J Indian Inst. Sci., Jan.-Feb 1998, **78**, 59-70.

A document processing system for detecting text on cover pages

R. MAHESH AND S. CHAUDHURI
Department of Electrical Engineering Indian Institute of Technology, Bombay, Powai, Mumbai - 400 076
Email: sc@ee.iitb.ernet.in

Abstract

A technique for detecting sparsely located texts on cover pages of books or magazines is presented in this paper. The task of such a textual segmentation is very difficult due to wide variations in the shape, size, fonts and colors in the documents. However, the solution to this problem finds a very useful application in building an automatic cataloging system

The proposed technique yields good results for a large class of cover page documents. Results of application of the technique on a few sample documents are presented here.

1. Introduction

A document image analysis deals with the transformation of images of printed and handwritten documents, including both graphics and text into computer revisable forms. The decreasing cost of hardware and the increased facility of networking makes it a good idea to have textual information on the electronic media in a computer-readable form. Several advantages of such of a document image processing system include creation of full-text on-line files which would assist in the remote access of libraries and information transfer, automatic sorting of letters, automatic preparation of hyper texts from printed documents, etc.

Analysis of scanned documents usually involves the following steps.

- Preprocessing: It includes steps like acquiring the image from a scanner and skew detection. Detection of the skew in the image of the scanned document and its rectification has been the focus of a large number of publications[1-3].
- Segmentation: It involves the separation of text from non-textual matter which may include graphics, figures, tables etc. Several techniques, such as smearing[4], profiling[5,6], have evolved to perform segmentation, especially for a binarized image of the document.
- Text Recognition: The textual part of the document has to be identified in terms of paragraphs, sentences, words and finally individual characters. The detected characters are then recognized using a number of methods[7,8]. One may also be interested in finding and recognizing special logos in the text. It may be mentioned here that, very often, the text recognition and the task of text segmentation supplement each other as some isolated non-characters may be erroneously segmented as text.

- Understanding: The internal data structure need not merely describe geometric properties and the meaning of words, it should possess knowledge about the topics being addressed in the document page. One may also be interested in understanding the lay-out used in the document[5,9,10]. Further, the relevance of graphics in the document needs to be understood.

In this paper, we deal with a very specific problem in document image processing: *Given a scanned image of the cover (front) page of any arbitrary book or magazine or report, we would like to detect its textual contents.* This has a great deal of applications in designing an automated library cataloging system.

The task of textual segmentation of images of internal pages of a book is quite simple compared to that of the cover page due to the following reasons:

- There is a good amount of tonal variations on the cover page and hence it must be digitized to 256 grey levels, unlike normal documents which are binary in nature.
- Cover pages invariably include a variety of art works which may often make the reading very difficult. Presentation of textual matters, although quite sparse in nature, varies in shape, size, style, fonts and shading. In the same cover, the size of the font may vary dramatically from being very huge to tiny.
- A cover designer prefers to embed the text within a graphics region. This poses a difficult problem for text detection.

A technique has been proposed by Zhong *et al.*[11] for the textual segmentation of color images for similar purposes. The technique hinges around the observation that the spatial variation, computed using a sufficiently long horizontal window, has higher values in textual region than in non-textual region. The spatial variance over a running window of the input image is generated and the Canny edge detector is applied; the resulting edges are suitably paired giving rise to bounding boxes indicating textual regions. Initially, the color image is used to obtain a color histogram in which neighboring peaks are merged together. The color of text within each box is determined and the text components extending beyond a bounding box is incorporated to complete the segmentation process.

However their approach fails to yield acceptable results if a word contains differently shaded (or colored) characters or if the variation in font size is quite large in the same document. The method that we propose here yields better results under these circumstances. Further, we note that the color of the font is redundant as far as the information content of the text is concerned, although it may be aesthetically more convincing. Hence we discard the color information to save computation, and scan the document in black and white.

The organization of the paper is as follows, section 2 discusses the proposed method and section 3 has a few typical results and a discussion of these results. Section 4 concludes the paper.

2. The proposed method

Observations on gray images of a large number of cover pages reveal that a distinctive feature of text on covers is that related pieces of text are generally of the same color and intensity implying that they will correspond to a particular gray value. Also there is usually a contrast in

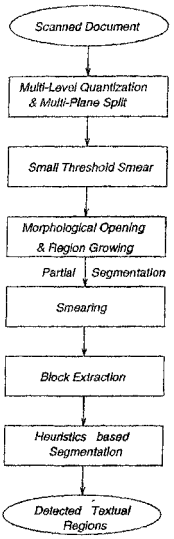


FIG. 1. A flow chart of the proposed method.

gray values of text and its immediate surroundings. One may, of course, cite some pathological examples.

The assumptions succinctly used in this study are:

- Related pieces of text are of the same color and intensity, *i.e.*, similar gray levels in the scanned image.
- There is a contrast in the gray value of text and its immediate surroundings.
- Text is nearly horizontal.
- The size of text portions exceeds a minimum preset value. This may result in isolated alphabets being discarded.

The various steps in the proposed method, depicted in the flow chart given in Fig. 1, are elucidated below.

2.1. *Multi-level quantization & Multi-Plane Split*

The scanned image contains several gray levels. Even within a region that appears uniform visually, the gray-values vary within some limits about the mean. In this step of multi-level

quantization a histogram of all pixel values with a strength above a minimum threshold (0.5\% of the image size) are chosen. All peaks which are close to each other are merged, the measure of closeness being given in terms of difference in gray levels. The neighboring peaks are merged if the difference is less than 20-25 gray levels. The process of merging starts by selecting two closest peaks in the histogram and these adjacent gray levels are combined to have a single gray level having a population strength equal to the sum of the two populations and the resultant gray value being a weighted mean of the two gray values. This procedure is repeated on the resultant histogram until no two peaks are closer than the minimum threshold. Typically the resultant number of quantizations (N) is between 4 to 8. The given gray image is now split into N such quantized planes, each being represented by the corresponding bit map.

2.2. *Small threshold smearing*

It was observed that images like those of a face and other gradually varying graphics, often appear as sets of granular points which are closely spaced, but not as continuous patches. The given area may be represented in more than one plane. This is due to some dithering technique used by the publisher while printing the cover page. Hence in order that the next processing step (opening) give good results, a smearing[4] with a small threshold of 2 pixels in both the x and y directions is done. This results in a smoothing of the granular portions of the non-text regions to a great extent.

2.3. *Morphological opening and region growing*

A distinct feature of text is that the stroke width is quite small and hence it cannot be opened with a circular structuring element of a moderate value of radius. After the previous step, a picture or any such graphical feature present on the cover page tends to contain local patches of uniform regions in each of the image planes. If each frame is independently opened with a circular structuring element (rC) of a large enough radius (r) what is left in that plane after this operation should ideally correspond to non-textual regions. The opened image in a plane thus contains a subset of seeds that may be grown to capture the non-textual regions.

A corollary of the assumption that the text substantially differs from its neighboring regions in terms of gray values is that the text and its immediate surroundings in any given plane are *completely disconnected*. Now morphological growth of the opened image in each plane is performed. This is done by repeating the following two operations until it cannot be grown any further. The two operations are

- Dilation: The image is dilated with a circle of unit radius (C). This results in an extension of the seed region in a plane by a unit distance in all directions.
- Logical Anding: The dilated image is logically anded at every pixel with the smeared image (I) as discussed in section 2.2.

The region growing (I_g) can thus be considered as a process of conditional dilation and may be mathematically represented as

$$I_g = \{\ldots \{\{\{\{I \circ rC\} \oplus C\} \cap I\} \oplus C\} \cap I\} \oplus \ldots \text{ } n \text{ times.}\} \tag{1}$$

where $\oplus$ represents the dilation operator and $\circ$ represents the opening.

This process would result in a growth of the opened segments (considered as seeds here) of a particular image plane and they will extend to cover most of the non-textual regions in the plane. Since the textual regions cannot be opened and they are disconnected from the non-textual regions, they do not appear in the planes obtained by such a growth. Thus large portions of the non-textual regions are identified.

During quantization, the transition pixels between any two distinct (as concerning gray values) and spatially adjacent regions may often result in thin lines of 'border' pixels that lie in a third plane. The aim of our current processing step is to remove, as far as possible, all portions of non-text. These 'border' pixels, by virtue of their lying in another plane, are not affected by the above processes of opening and region growth. Hence the obtained grown images are dilated by an extra pixel so that they extend by a pixel in all directions. This may, however, cause the text regions to shrink by a uniform pixel in all directions. If the strokes in the text are very thin, this may induce discontinuity in the textual fonts.

The regions belonging to non-text as identified in this step are removed from each of the planes.

2.4. *Run length smoothing (smearing)*

The run length smoothing algorithm is now used on each of these planes. This results in sets of neighboring pieces of text getting connected to be classified as a single connected component. If the plane contains isolated points, smoothing may result in deterioration of the proposed method. Hence before smearing, isolated points are removed from each one of the planes. There could be some stray non-textual segments in the smeared image, and they need to be removed at a later stage (The choice of the run length may be quite crucial for different types of cover pages).

2.5. *Block extraction*

Run length smoothening is followed by a 'block' capture. In block capture each connected object is enclosed by the smallest rectangle within which it may fit in. The algorithm works as follows

- Starting from the top left corner of the image, the image is raster scanned from left to right.
- On encountering the first pixel that belongs to any object, the connected segment is traversed in the clockwise direction along the boundary capturing the minimum and maximum coordinates in both X and Y directions. The bounding block is now determined from the coordinates.
- Once a bounding block has been determined for an object, the segment is removed and the raster scanning is continued till all connected components are captured.

2.6. *Heuristics based segmentation*

Each connected component in the image, specified by a delimiting rectangular block, must now be identified whether it is textual or not. Certain heuristics are used to arrive at a conclusion. Any block that may constitute text requires a certain amount of width and length. The minimum

thresholds for a text block are decided experimentally and any block not meeting these requirements is discarded. In our experiments, the minimum width was chosen to be 5 pixels and the minimum length to be 50 pixels.

For any block that passes the above test, the ratios of the numbers of object pixels within each block in a frame before and after smearing, to the size of the block are calculated. Let us denote these ratios by $\mathcal{B}$ and $\mathcal{A}$ respectively, where

$$\mathcal{B} = \text{number of object pixels before smearing/block size}$$
$$\mathcal{A} = \text{number of object pixels after smearing/block size}$$

For any block to qualify as a text block these parameters must satisfy some heuristic conditions, *i.e*,

$$Min < \mathcal{B} < Max \quad and$$
$$\mathcal{A} > \delta$$

Typical values of *Min* and *Max* are around 0.2 and 0.5, respectively. A choice of $\delta = 0.5$ was experimentally found to perform well for most documents.

Having segmented the planes individually, the various planes with the identified regions of text, are OR-ed together yielding the final result.

FIG. 2. Input image of a cassette cover.

3. Results and discussion

The proposed technique is applied on a large number of cover-page documents to test its performance. A few of the results are presented here for brevity.

3.1. *Experiments on cassette covers*

The technique was applied to cassette covers, scanned at 200dpi as a gray image on a HP scanner. The obtained results at various stages of the processing for a cover are displayed. The input gray image of the cassette cover is given in the figure 2. It is readily seen that it contains fonts of various styles, sizes and shades. It has logos of the producer and even has text embedded in graphics. Thus it is a very complex document for the task of textual segmentation.

Multi-level quantization results in five levels, they correspond to the gray values 53, 110, 135, 184 and 219. Three of the planes obtained after multi-level quantization are displayed for illustration in figure 3. Note that the dark regions in each plane corresponds to an object. The graphics region tends to be present in a number of planes, while the textual component primarily shows up in the two planes. For example the text 'ANAND' appears in the plane - c, but an outline of it does appear (as discussed in section 2.3) in plane - b.

The planes results after the morphological opening and region growing step as mentioned earlier contains only graphics. A sample result corresponding to the gray level 135 (figure 3b)

(a) (b)

(c)

FIG. 3. Illustration of multi-level quantization of the document given in figure 2 into a number of planes. Planes a, b, and c correspond to gray levels 110, 135 and 219 respectively.

FIG 4 Illustration of the process of capturing the non-textual regions using morphological opening and subsequently region growing. The result corresponds to the processing of figure 3b

is displayed in figure 4. The circular structuring element used for the purpose of opening, had a radius of 9.

In figure 5, the final result of textual segmentation is displayed. It is noticed that words like 'Abhimaan', 'Anand', 'Aandhi', 'S. D. Burman', 'Gulzar', etc have been captured very well inspite of their drastically differing sizes, thickness and font styles. Words like 'Rupam Chitra' have been captured but with less clarity, this is because after quantization the pixels corresponding to these pieces of text split into in more than one plane. The above problem during quantization results in the complete loss of some thinly typed words like 'SPHO'.

It may be noted that some amount of post processing is now needed as parts of non-text have also been classified as text. A solution to this problem would be to send the segmented image to a character recognizer which would discard these as graphics.

The various parameters used in the experiment are:

Smearing thresholds : In the horizontal direction = 20 pixels, In the vertical direction = 1 pixel;

$$Min = 0.15, Max = 0.60; \delta = 0.50.$$

FIG. 5. Segmented text in the cassette cover image.

3.2. *Experiments on a book cover*

The figure 6 displays the gray image scanned at 100 dpi which was taken up for segmentation. The only bit of text lost is the encircled 'R' to the top right of the word 'mouseware'. This is due to the fact that this piece of text is of extremely small dimensions. The various parameters like the radius of the structuring element used for opening, smearing thresholds, and the parameters used in heuristic segmentation are same as those used in the previous case.

3.3. *Bengali book cover*

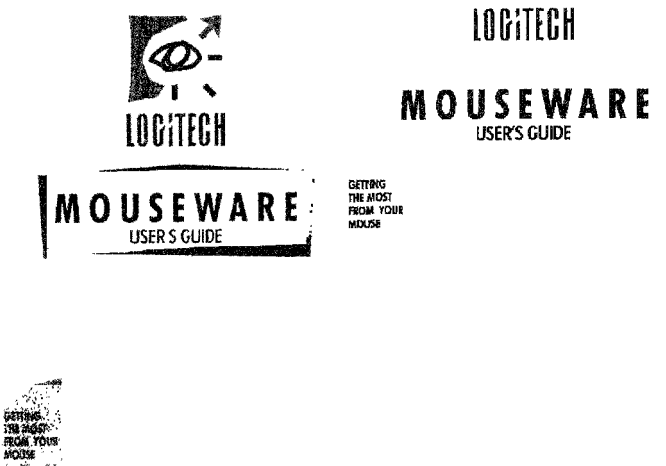
We now show that the proposed technique does not depend on the type of linguistic script used in the document. The cover page of a Bengali book scanned at 150dpi was chosen. In this case, in order to capture the name of the book certain parameters had to be changed. This is due to its extremely large font size. The parameters are:

Radius of structuring element used in opening: 14,

$\delta = 0.35$ (in place of 0.50 used in the other cases),

Smearing threshold in the horizontal direction : 30.

Also to remove the dotted pattern at the center, the value of *Max* was reduced to 0.40. The value of 0.60 used in the cassette cover is needed only in cases where the text fonts are very



(a) (b)

FIG. 6. (a) An example of a cover page of a book and (b) the segmented output.

thick with very little space in between characters as in the word 'AANDHI' of the cassette (see figure 2) otherwise the limit of 0.40 on *Max* is quite adequate.

The initial gray image is displayed in figure 7. The final segmented output is given in figure 8. The results of the segmentation shows clearly that our technique does not depend on the nature of the script. It may be further noted that the logo of the 'Ramakrishna Mission' in the bottom left corner has also been captured.

3.4. *Handwritten text*

We now illustrate the performance of the proposed scheme on the cover page of a hand written document. Figure 9 shows one such document in Bengali, scanned at 150dpi resolution. Since the text here is handwritten and is very thin, the values of *Min* and *Max* have been modified to give good results. In this case:

$Min = 0.1$; $Max = 0.3$.

The opening radius was chosen to be 8 (similar to the one chosen in the case of the cassette cover). The final segmented output is displayed in figure 10. It may be noted that the segmentation has left behind a portion of the graphics, that need to be eliminated through postprocessing. The textual regions have been captured very well. Thus the segmentation technique works well even in the case of hand-crafted documents.

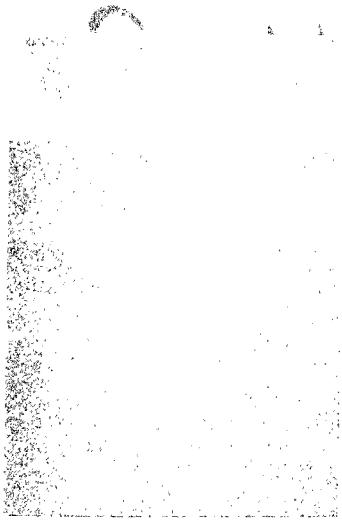


FIG. 7. Input gray image of a bengalie book cover.

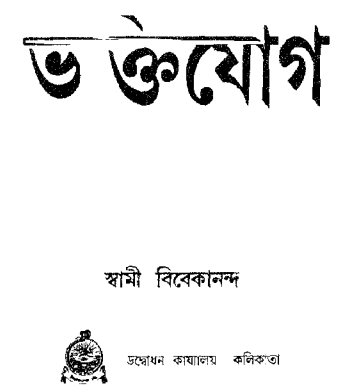


FIG. 8. The segmented text for the document in figure 7.

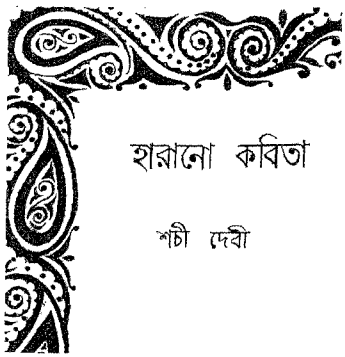


FIG. 9. Example of a handcrafted coverpage for a Bengali book of verses by an amateur poet.

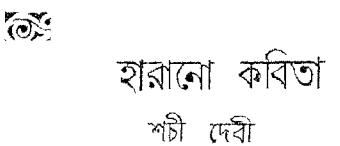


FIG. 10 Results of textual segmentation of the document in figure 9.

4. Conclusions

We have presented a method for detecting textual regions in scanned documents of cover pages. Such texts are quite sparse but it includes a wide variabilities in terms of shape, size, style and fonts. The proposed technique has been found to yield good results for a large class of documents, and it may prove to be a useful technique while building an automated cataloging system. However it does involve quite a bit of heuristics and it would be ideal if one can automatically determine the values of the various parameters used in this experiment from the scanned document itself. Current study continues in this direction.

References

1. LE, D. S., THOMA, G. R. AND H. WECHSLER, Automated page orientation and skew angle detection for binary document images, *Pattern Recognition*, 1994, pp. 1325–1344.
2. POST, W., Detection of linear oblique structures and skew scan in digitized documents, *Proc of 8th Int. Conf. on Pattern Recognition*, pp. 738–743, 1986.
3. CHAUDHURI, S. AND AVANINDRA, Robust detection of skew in document images, *IEEE Trans on Image Processing*, vol-6, 1997.
4. WONG, K. Y., CASEY, R. G. AND WAHL, Document analysis system, *IBM J. Res. Develop.*, Vol. 26, 1982, pp. 647–656.
5. DENGEL, A , ANASTASIL: A system for low-level and high-level geometric analysis of printed documents, *Structured document image analysis*, H. S. Baird, et al. (Eds), Springer-Verlag, 1992, pp. 70–99.
6. NAGY, G. AND SETH, S., Hierarchical representation of optically scanned documents, *Proc. 7th Int. Conf. on Pattern Recognition*, 1984, p. 347, Montreal.
7. KHOTANZAD, A. AND HONG, Y. H., Invariant image recognition by zernike moments, *IEEE Trans. PAMI*, vol. 12, 1990, pp. 489-495.
8. TEH, C. H. AND CHIN, R. T., On image analysis by the method of moments, *IEEE Trans. PAMI*, vol. 10, 1988, pp. 496–513.
9. NAGY, G , Towards a structured document utility, *Structured document image analysis*, H. S. Baird *et al.* (Eds), Springer-Verlag, 1992, pp. 54–69.
10. Viswanathan, M., Analysis of scanned document—A syntactic approach, *Structured Document Image Analysis*, H S. Baird, *et al.* (Eds), Springer-Verlag, 1992, pp. 115–136.
11. ZHONG, Y., KARU, K AND Jain, A. K., Locating text in complex color image, *Pattern Recognition*, Vol. 28, 1995, pp. 1523–1535.