Extraction of type style-based meta-information from imaged documents

B.B. Chaudhuri, U. Garain

Computer Vision & Pattern Recognition Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India; e-mail: {bbc,utpal}@isical.ac.in

Received July 12, 2000 / Revised October 1, 2000

Abstract. Extraction of some meta-information from printed documents without carrying out optical character recognition (OCR) is considered. It can be statistically verified that important terms in technical articles are mainly printed in italic, bold, and all-capital style. A quick approach to detecting them is proposed here. This approach is based on the global shape heuristics of these styles of any font. Important words in a document are sometimes printed in larger size as well. A smart approach for the determination of font size is also presented. Detection of type styles helps in improving OCR performance, especially for reading italicized text. Another advantage to identifying word type styles and font size has been discussed in the context of extracting: (i) different logical labels; and (ii) important terms from the document. Experimental results on the performance of the approach on a large number of good quality, as well as degraded, document images are presented.

Key words: OCR – Meta-information – Type style – Font size – Information retrieval

1 Introduction

A huge amount of document-based information in the form of books, journals, magazines, manuals, legal documents, newspapers, etc., is prepared every day throughout the world. The documents thus prepared can be classified into two broad categories, namely: (i) soft documents or hypertext-based documents; and (ii) hard documents or paper-based documents. Both categories have their advantages and disadvantages. Soft documents that are generated by computers and maintained in machine memory help in automated information processing within the electronic world. They are easy to process but less handy because a PC is needed to access them. However, with the emergence of the world-wide web (WWW) as a common platform for sharing information, the quantity of such electronic documents is now exploding. On the other hand, paper-based documents are rigid but easier to store, carry, and read. Because of a long-ingrained tradition, many people still rely on paper-based documents as their favorite medium and source of information.

For interchange and interaction of information, it is useful to convert one category of document into another. Because of the remarkable advances in computer-based printing technology, a soft document can be easily converted into a paper-based hard document. However, conversion of a paper-based document into a corresponding soft document is more difficult. One of the possibilities is the use of optical character recognition (OCR) systems. Given a document, an OCR system tries to recognize the text and converts it into the corresponding electronic format. However, for complex documents containing graphics, logos, and halftone pictures, and for documents with handwritten text, the OCR systems are still far from perfect.

Since the amount of both types of documents is increasing day by day, there is a pressing need for a fast and efficient system for information extraction from the documents. For electronic documents, many information retrieval (IR) techniques have been reported in the literature [7, 16, 21]. Commercial search engines like Yahoo, Alta-vista, etc., are in constant use.

The problem of information extraction from paper-based document is more complex. One possible approach is to read the document with an OCR system, convert it into an adequate electronic format, and then apply the conventional IR approaches for electronic documents. Taghva et al. [26] have undertaken an in-depth analysis of the interaction between OCR and IR. They point out that though OCR errors do not affect average retrieval effectiveness, there are other consequences that should be considered when OCR-generated text is applied for IR. Such an approach is bound to be slow and error prone. Thus, some alternative methods based on image manipulation have been tried. Doermann [9] presents an elaborate survey on indexing and retrieval of document images. Among others, Chen and Bloomberg [5] demonstrate that English-like textual document images can be

Correspondence to: B.B. Chaudhuri

summarized without OCR. In another study, Chen et al. [6] presented a segmentation and recognition-free approach for searching key words in document images.

In this paper, we have presented some OCR-free document processing techniques. Our main objective is to extract some meta-information from the imaged documents and apply them to make document processing smarter and more efficient. Meta-information can refer to information related to character font, size, style, document layout, etc. We have focussed our attention on detecting the character styles, namely, bold, italic, and all capitals for each word in an imaged document and show their application potential in: (i) spotting important terms; (ii) layout analysis; and (iii) OCR performance improvement.

The work is motivated by a survey [4] on the relative abundance of italic, bold, and all-capital words in printed documents. For conducting this survey, we covered more than 6,000 document pages of different technical journals, conference proceedings, technical books, etc. It was observed that, in most cases, the paper's title, section/sub-section headings, chapter headings, figure captions, table titles, mathematical text, and important terms (in the context of the document content) are written in a style other than the normal ones.

The initial concern of this paper is to present image-based approaches for the detection of character style and size for each word. The detection of type style for words helps in finding the paper's title, section headings, subsections or chapters without actual character recognition. In addition, it helps in spotting some terms that are important in the context of the document. The list of such important terms helps in identifying some of the keywords from the document image.

Identification of italic-styled words has another important application in the field of OCR. Baird and Nagy [2] demonstrate that a significant improvement in recognition accuracy of an OCR system could be achieved by utilizing the font information. Studies on font recognition by Khoubyari et al. [13] and Zramdini [28] also point in this direction. We have observed that the performance of many commercial OCRs deteriorates with style variations. The degradation is drastic with italic style, which increases with the increase in the slant angle of italicized characters. If our system for identification of word-type style is used to detect the italic words and the corresponding slant angle is computed, then words can be de-italicized by an inverse shearing transform corresponding to the slant angle. If we feed the result to the OCR system then a significant improvement in the recognition rate can be achieved. The detection of italic style is easier and faster than identification of exact character font.

This paper is organized as follows. Section 2 describes the detection procedure of italic, bold, and all-capital words. A smart approach for determining the character size is also outlined in this section. In Sect. 3 the usefulness of identifying type style and font size is discussed. The results of using the proposed methods on real data are presented in Sect. 4. Section 5 concludes the paper and discusses the scope of future work.

2 Identification of type style and font size for words

First, we consider the identification of words that are written in italic, bold or in all-capital style. The following three sub-sections are dedicated to this. Next, we consider the measurement of character size in Sect. 2.4.

2.1 Detection of italic words

If the words of a text are in italic style, then their characters are slanted. We define 'slant' as the angle in degrees clockwise from the vertical at which the characters are sheared. Hence, to recognize italic words, our main objective is to determine the character slant angle.

Earlier approaches to the detection of italic style can be divided into four groups: (i) morphological analysis of the image; (ii) statistical analysis of stroke patterns; (iii) component analysis; and (iv) a knowledge-based approach. Bloomberg [3] used an interesting multiresolution morphological analysis that detects italic words using hit-and-miss transforms (HTM). In this author's approach, a structuring element (SE) is constructed, assuming that a distinguishing feature for italic words is that the edges are inclined at about 12 degrees from the vertical. However, such an SE is a somewhat weak filter as the slant angle varies from one particular font to another. Furthermore, it is prone to image variation and noise.

Among the studies based on the statistical analysis of stroke patterns, Shi and Pavlidis [23, 24] used the histogram of stroke slopes for separating italic versus non-italic styles as well as serif versus sans serif fonts. In their study [24], they considered documents where the whole text is typed in italic or upright (normal) style. In other approaches, Tsirikolias et al. [27] used a moment-based approach to recognize slanted characters. Kim and Kwon [14] suggested a method based on sampling and quantization to recognize the skewed characters, while Sun and Si [25] used gradient direction for the detection of slanted characters from document images having few italic words. However, the detection of italic words was not the main concern of the last three studies.

Doermann et al. [8] identified italic words by constructing a minimum upright bounding parallelogram for each component, but did not provide any detail. Among the knowledge-based approaches, Baird and Nagy [2] detected the italic styles utilizing the font information. They used a 100-font classifier, which was automatically adapted to a specific font. Khoubyari and Hull [13] presented an interesting work on font detection by recognizing the frequent function words such as 'the', 'of', 'and', 'a', 'to', etc. Zramdini [28] detected italic style as a part of his study on font detection. He calculated the first derivative of the horizontal projection profile and used it as one of the features stored in the knowledge base. None of the above studies presented the success rate for italic word detection.

In our approach, we have used the simple shape information of English characters. It is observed that most

Fig. 1a,b. A character showing property 1: **a** normal style **b** italic style

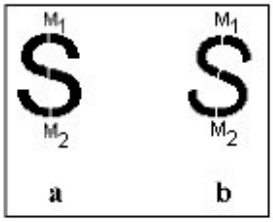

Fig. 2a,b. A character showing property 2: **a** normal style **b** italic style

English characters (irrespective of upper and lower case) show at least one of the two properties defined below:

Property 1. The character has at least one straight-line segment that is 80% or more of the total character height. If such a line exists then it is almost vertical for normal-styled characters and slanted for italic-styled characters (see Fig. 1).

Property 2. The character has a single black run if we scan horizontally at the top and bottom row. For such a character, let the midpoints of the black run at the top-most and bottom-most row be M_1 and M_2, respectively. If we draw an imaginary line through M_1 and M_2 then the line is vertical if the character is in normal style and slanted if the character is in italic style. Figures 2a and 2b show the midpoints M_1 and M_2 for the character 'S' when it is in normal and italic style, respectively.

We use the above two properties in a hierarchical fashion. Property 1 is tested for first. For the characters that do not exhibit property 1 we test for property 2. In both cases, we start horizontal scanning from the row which is some rows (say, two rows) below the actual top-most row and stop at the row which is some rows (say, two rows) above the actual bottom-most row. This provides immunity to serif and sans serif style variation and the noise at the contour.

To test for property 1 we draw all possible straight lines satisfying property 1 and for each such straight line we calculate the slant angle defined before. From these slant angles we get a measure of the overall slant of the character. If the calculated slant angle is θ, then for

- $\Phi_1 \leq \theta \leq \Phi_2$ we decide that the character is in italic style.
- $\theta \leq \Phi_1$ we decide that the character is in normal style.
- $\theta \geq \Phi_2$ we delay our decision till property 2 is tested for

where Φ_1 and Φ_2 are two predefined threshold values used to provide immunity to image variations and noise. It is interesting to note that property 1 is able to distinguish italic vs non-italic styles for 30 out of 52 English characters including both upper and lower cases. The characters for which this discrimination is achieved are listed in Fig. 3a. For some characters (like 'z' or 'Z') that contain a slanted part even in normal style, we get a very large slant angle (which is beyond our threshold) and invoke the test for property 2. On the other hand, there are eight characters (see Fig. 3b) that do not have any vertical line but contain both slanted left and right lines. They have vertical shape symmetry when they are in normal style. Hence, during the test for property 1 these characters show nearly zero slant angles when they are in normal style and give positive values when in italic. Thus, italic/non-italic style for 38 out of 52 characters is detected at this stage.

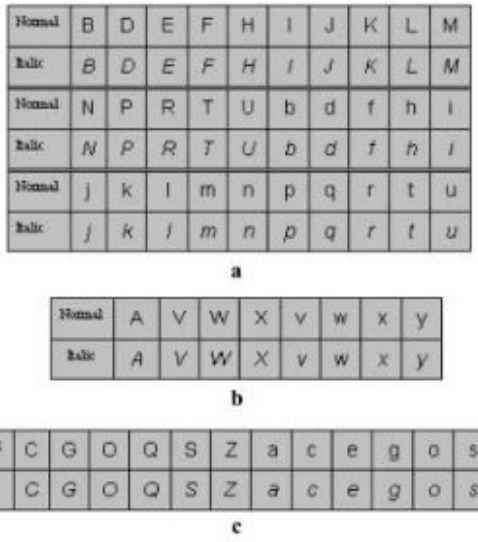

Fig. 3a–c. Detection of italic words: **a,b** italic characters detected through the test for property 1 **c** italic characters detected through the test for property 2

For the rest of the characters, where the test for property 1 does not succeed, property 2 is tested for. If M_1 and M_2 exist according to property 2 then the angle made by the line M_1M_2 with the vertical is considered as θ. Then the decision is made using the first two conditions given above. If θ does not satisfy any of these two conditions (we have not come across any such situation) then consider the character as a normally styled one. From typographical knowledge and testing on a large set of italic texts, we set $\Phi_1 = 8°$ and $\Phi_2 = 20°$. The test for property 2 distinguishes italic/non-italic styles for 13 characters (see Fig. 3c) for which style (italic/non-italic) detection could not be done through the test for property 1.

Thus, italic style for 38 + 13 = 51 characters can be correctly identified. No decision can be made for the character ('Y'). But this hardly affects our final goal of detecting italic words because, after all the characters within a word are classified, the majority voting approach is used. A word-level decision (whether the word is italic) is taken by counting the majority of the styles (normal or italic) in which the characters within the word are printed.

Fig. 4a,b. Bold characters: **a** uniformly bold **b** non-uniformly bold

2.2 Detection of bold words

Extraction of words in bold-type style is a more difficult problem. A bold character is mostly identified by the thickness of character strokes. However, stroke thickness in a scanned image may differ significantly from that in the original document. This is mainly due to two reasons: (i) image degradation because of paper quality, print quality, photocopying, etc.; (ii) choice of inappropriate threshold values for gray tone to two-tone conversion, etc.

Very few researchers have studied the problem of detecting boldface words. Among earlier studies, Bloomberg [3] adopted a multiresolution morphological analysis for this purpose. He used two structural elements (SE) for thinning vertical lines from left and right. In his approach, the image is progressively thinned in the horizontal direction. At each thinning iteration, the number of black pixels is counted and a ratio of previous/current counts on successive iterations is constructed. If there is a small amount of bold text intermixed with normal text, at some point the majority of the normal text will disappear. No success rate for the detection of bold words is discussed. Doermann et al. [8] also identified the boldface words using a morphological approach. They applied an erosion transform but did not provide any details.

We observed that the thickness of some or all stroke segments of bold characters is more than those of normal characters. Based on this relative thickness, bold characters can be of two types (see, Fig. 4). In one type, all stroke segments of the character are more or less uniformly bold (thick) as in Fig. 4a. In another type, some strokes are bold while others are normal (or even thinner than normal) as in Fig. 4b. Our idea is to detect the bold strokes by measuring their thickness.

To do so, we compute the run-length of black pixels along several directions at some boundary points. Consider the starting point for each new run of black pixels for each horizontal row scan and measure the run-lengths in three directions namely, horizontal, $+45°$ (i.e., upward), and $-45°$ (i.e., downward) directions. For a point P (see Fig. 5), let the run-lengths in these three directions be $w_h(P)$, $w_u(P)$, and $w_d(P)$, respectively. We take the minimum of these three values as the thickness $w(P)$ at point P i.e., $w(P) = min[w_h(P), w_u(P), w_d(P)]$. At a particular point P, we consider its thickness $w(P)$ provided the thickness in the neighborhood of P along the boundary of the stroke is nearly equal to $w(P)$. In this way, thickness at inconsistent points, such as corner and serif, are avoided. Let C_w denote the number of stroke border points at which the thickness w is encountered. Consider the histogram C_w against w. For the characters in Fig. 4a, the histogram will be unimodal, while for the characters in Fig. 4b the histogram will be bimodal. This is shown in Fig. 6.

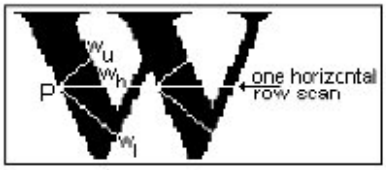

Fig. 5. Detection of bold characters: thickness at a stroke boundary point

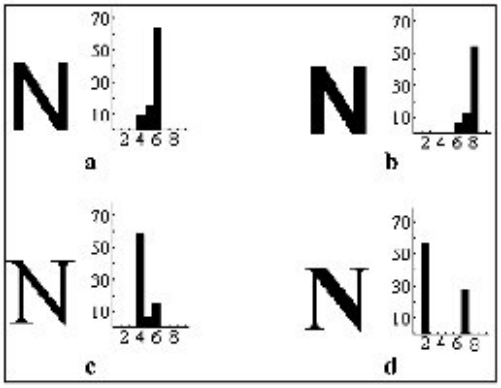

Fig. 6a–d. Detection of bold characters: W vs C_w histogram for **a** a normal (uniform) character **b** a bold (uniform) character **c** a normal (non-uniform) character **d** a bold (non-uniform) character

Next, we partition the histogram halfway between the maximum thickness w_{max} and minimum thickness, w_{min}. Let, $w_o = (w_{max} + w_{min})/2$. Now the average thickness, say w_t, of the bold (thick) strokes is estimated as

$$w_t = \frac{\sum_{w=w_o}^{w=w_{max}} wC_w}{\sum_{w=w_o}^{w=w_{max}} C_w} \quad (1)$$

The thickness of character strokes depends on the character size (height). We classify a character as bold for which w_t is more than α times ($\alpha < 1$) the character height. From typographical knowledge and testing on a large set of bold texts, we choose $\alpha = 0.20$. The above process is repeated on each character of the word.

2.3 Detection of all-capital words

The detection of words in all-capitals style is relatively easy. An English line (or word) can be partitioned into three zones as shown in Fig. 7a. In our approach, zone detection is carried out on a text line. First, the number of black runs (or crossing counts) is calculated for each horizontal row scan. Next, a histogram is constructed for crossing count (C_h) vs row-number (i). Figure 7b shows the histogram for the text line shown in Fig. 7a. The histogram has four peaks estimating the four virtual lines that actually define the three zones as shown in Fig. 7a.

- $H_{top} = \max\{i \text{ such that } C_h[i] > 0\}$.
- $H_{bottom} = \min\{i \text{ such that } C_h[i] > 0\}$.

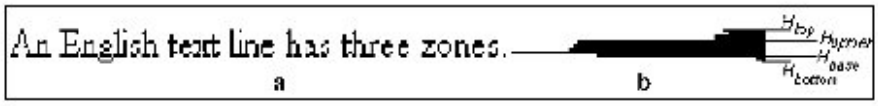

Fig. 7a,b. Detection of three zones of an English text line: **a** three zones of an English text line **b** zone detection through histogram of horizontal crossing counts

- $H_{base} = i$ such that $C_h[i]$ - $C_h[i\text{-}1]$ is maximal.
- $H_{upper} = i$ such that $C_h[i]$ - $C_h[i\text{-}1]$ is minimal.

Among these three zones, we are interested in upper and middle zones since any capital character covers only these two zones. In our experiment, we do not encounter any text line (a text line may consist of 10–30 words) having all characters lying in the middle zone. On the other hand, if a text line shows only one zone then we decide that all words in that line are capital ones. However, all the words in a text line may not be in capital letters. To detect occasional all-capital words in a text line, we detect the three zones and group the characters into the following subsets:

- Subset 1. Characters having their parts only in the middle and in the upper zone.
- Subset 2. Characters which do not belong to subset 1.

Since all the capital characters belong to subset 1, we are interested in that subset. However, as shown in Fig. 8 there are some small (lowercase) characters that also belong to subset 1 and we want to distinguish them without doing any character recognition. The following two stages are used hierarchically for this purpose. Figure 9 shows the characters for which the capital style is detected at each stage.

Stage 1. At first, we check the number of black runs (or cross count) in the horizontal scan for some rows just above the middle zone. If we get at least one black run then the processing of stage 2 is invoked for that character. A zero black run implies that the character is not a capital one. If the number of black runs is greater than one, we declare that the character is capital. In fact, the processing of stage 1 requires very little computational effort as the cross counts for the rows are calculated before while detecting the three zones of a line.

It is interesting to note that 19 out of 26 uppercase characters are detected using this simple feature at stage 1 (see Fig. 9). For some serif fonts, though rarely used, the lowercase character 'f' is identified as having a capital style at stage 1. Another confusion arises for the pattern 'ff' (where two 'f's touch each other). Such problems hardly affect our final decision since we take a majority decision for word classification.

Stage 2. When stage 1 fails to identify case information stage 2 is invoked. Let l be the run-length of black pixels of the row for which we get a single black run in stage 1 and let L be the maximum run-length of black pixels obtained among rows in the upper zone. Then the following decisions are taken:

- if $l < L/2$ then the character is capital,
- otherwise the character is small.

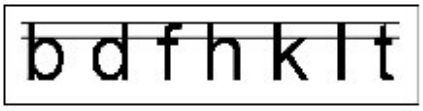

Fig. 8. Small characters that belong to subset 1

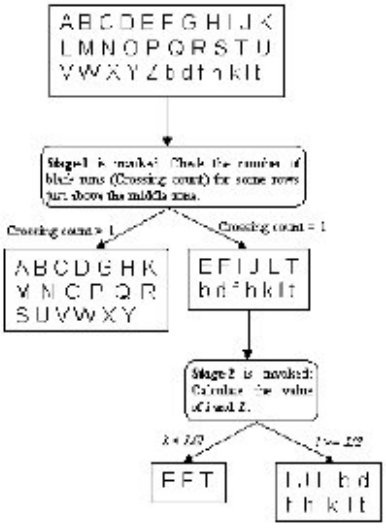

Fig. 9. Detection of all-capital characters

Figure 10a shows that the l of the lowercase character (having character parts in the upper zone) almost equals L whereas, in Fig. 10b l is much less than $L/2$ for the capital character. In Figs. 10a and 10b, rows (just above the middle zone), for which a single black run is obtained in stage 1, are shown by the white stripes on the character images.

Confusion arises for three capital characters 'I', 'J', and 'L'. Among them, the identification of 'I' is easier in English text, as it often occurs as a separate word whose width (or aspect ratio) is minimum. The capital style of character 'L' is detected by identifying the horizontal line segment at the bottom position. In our approach, explicit capital-style detection for the character 'J' is not done for two reasons: (i) the occurrence of 'J' is relatively rare; and (ii) the final decision regarding a word style (capital or small) is taken by calculating the majority of the style (capital/small) in which the individual characters are printed. The success rate for the detection of capital style at the word level is presented in Sect. 5.

2.4 Determination of character size

Character size is determined by calculating the character height in terms of the number of pixels. The algorithm for the determination of character size picks one line at a time and calculates the height of the middle zone, which

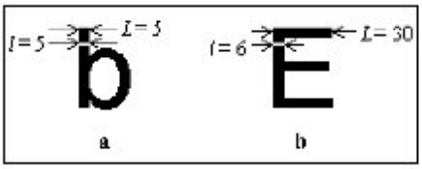

Fig. 10a,b. Some of the characters that need stage 2 processing

is called the character x-height. The three zonal divisions of English text have been discussed above. Let H be the x-height of the text line, say T, being processed and let H_p be the x-height for known character at point-size, say P. The character size for the characters of T is calculated as:

$$P' = P + \text{int}\left[\frac{H - H_p}{h} + \text{sign}(H - H_p) \times 0.5\right] \quad (2)$$

where, h is the increment in x-height when the character size changes from point size P to point size $P+1$.

P, H_p, and h of (2) are known for a particular scanning resolution and H is calculated for a character of unknown size. The x-height of a character may vary from one font to another at a fixed size. Thus, instead of exact character size we get an appropriate value that is very near to the actual one. If any confusion arises during determination of character size for a word, contextual information, such as the character size of the surrounding words, is used to resolve it.

In our experiment, we use $P = 10$. H_p and h are set to 18 and 2, respectively, where documents are scanned at 300 dpi. But the lines containing words with all capital characters do not exhibit the three zones of a text line. However, we detect all-capital words prior to the determination of character size using the approach stated in Sect. 2.3. The character size for all-capital words is determined using the same expression as in (2) but using different values for H_p and h. For example, we use H_p = 28, h = 3, for P = 10 where the scanning resolution is 300 dpi. The details of the experimental results are discussed in Sect. 4.

3 Usefulness of identifying type style and font size

Identification of different type styles and font size has various applications in the field of document image analysis (DIA). One of these is to improve the recognition accuracy of an OCR system. It is observed that the performance of existing text recognition systems degrades as the type styles deviate from the normal. In this section, we propose a simple but efficient technique to take advantage of homogeneous type style to improve accuracy in character recognition.

Another useful application of identifying type styles and font size is to provide some OCR-free document processing for technical articles. Most of these documents are organized under title, section, sub-section or chapter headings. In this section, we outline how these headings are extracted using type style and font size information.

Normal style text may be easier to read

Normal style text may be easier to read

Fig. 11. Examples of text printed in normal and italic style in the same font

Italic, bold, and all-capital words typically have special significance in a document. Once identified, they can be used as keywords for automatic indexing into a database of scanned document images. This has been quantitatively demonstrated in the next section.

3.1 Improvement of OCR performance

It is observed that the OCR systems produce good results when a document is printed in normal style and in commonly used fonts. However, whenever there is any deviation from the normal, in font as well as in style, recognition accuracy degrades. Hence, information about the font and style of the text may be used to improve the performance of an OCR system [2, 13, 23, 28].

Here we do not attempt to identify the character fonts but propose a practical approach to improve character recognition accuracy. Compared to normal characters, the italic characters differ significantly in shape. Figure 11 exhibits two text lines, one in normal, and the other in italic style in the same font. The performance of many commercial OCR systems deteriorates with style variations and the degradation is drastic with italic style. We noted that the recognition score continually decreases with increases in the slant angle of the italicized characters. A typical example is shown in Fig. 12.

To improve the performance, we run our italic detection algorithm on the document. Once a word is detected as italic, the slant angle is computed and the word is de-italicized by an inverse operation. Thus, all words in the document become homogeneous regarding style. Next, the resultant document is fed into the existing OCR systems. We observe that better recognition accuracy could be achieved through this approach. Figure 12c shows the text in Fig. 12a after de-italicizing the italic words. Figure 12d exhibits the OCR output when the document in Fig. 12c is processed. Here the misrecognition rate has been reduced from 6.85% to 0.33%. Further results are presented in the next section.

3.2 Extraction of headings and other meta-information

To extract headings, at first, we take the vertical projection profile of the document and extract the regions surrounded by larger white spaces. In each extracted region the type style for each word is checked. If all words (or the majority of them) are written in italic, bold or all-capital style, or the size is larger than the normal, then we may decide that the line or lines in that region is a kind of heading. Headings are categorized by using

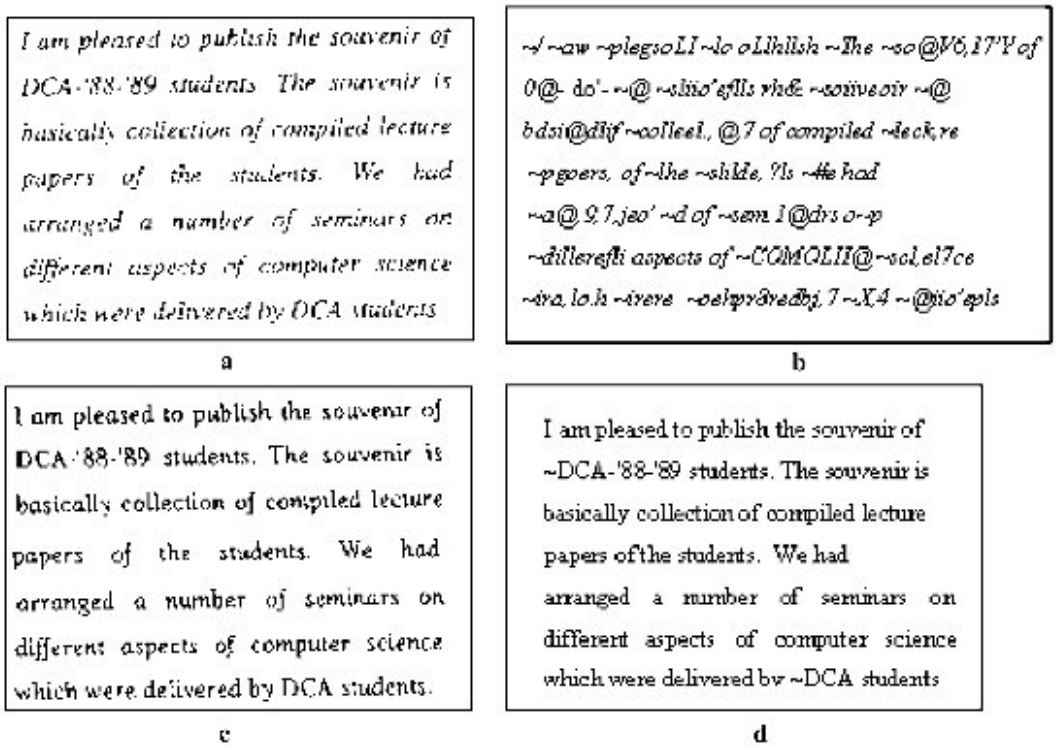

Fig. 12a–d. Improvement of the OCR system performance: **a** italic styled text **b** OCR output of the text in part a **c** de-italicized text of part a **d** OCR output of the text in part c

the conventions for preparing paper documents. A heading is identified as the main title of the article by its font size (i.e., if the font size is the largest one). Physical position (i.e., layout information) also helps us to reach such a conclusion. Other headings, such as section or sub-section titles, are also identified and categorized by examining their size and position in the document image. For a multi-column document we assume that the column segmentation is already done. The segmentation method by Jain and Yu [11] or by Pavlidis and Zhou [18], for example, may be used for this purpose.

In some cases, there exist wide white spaces above and/or below the author names, figure captions, table titles, etc., and these texts are generally written in italic or in bold style. Regions containing such text are also extracted. The figure caption is identified by finding a non-textual region [18] just above the extracted region. Similarly, the region containing a text for table titles can also be detected by finding a table structure [10] just above or below the region. Regions extracted by examining the white space around them may also contain other types of text, such as mathematical expressions. These are useful meta-information that can be extracted and stored separately for an imaged document.

Sometimes, instead of a heading, a whole paragraph may be extracted by using the white space heuristics. Such situations can be avoided by looking at the word type styles, because words in a paragraph are written in normal style only. In Figs. 13a and 13c the images of two successive pages of a technical paper taken from the proceedings of the 14th ICPR are shown. Figures 13b and 13d show the extracted regions from Figs. 13a and 13c, respectively. These regions have wide white spaces around them and contain words in italic, bold or all-capital styles only. Since the abstract in Fig. 13a is written in italic style, the region that contains the abstract is also extracted along with the headings as shown in Fig. 13b. Similarly, the mathematical equation of Fig. 13c is also extracted and shown in Fig. 13d.

4 Test results

The proposed methods are implemented on a 166 MHz PC with 32 MB RAM using 'C' language. The algorithms are tested on a database of two hundred pages. On the average, each page contains about fifty lines and each line contains about ten words of printed English texts in different font style. Details of the database are given in Table 1. Apart from real-life documents, 20 pages are prepared using software packages like the LaTeX formatting system [15] and MS-WORD 97 [17]. These pages are more or less clean and noise-free documents and contain a relatively large number of italic, bold, and all-capital words. The algorithms are tested on another 25 pages that are degraded versions of the noise-free documents. There are a few models [1, 12] for generating degraded documents from noise-free versions. We follow the model proposed by Kanungo et al. [12] and generate the degraded documents synthetically.

The success rate in identifying italic, bold, and all-capital words are shown in Table 2. Since we take the word-level decision by counting the majority of the styles in which the individual characters within the word are

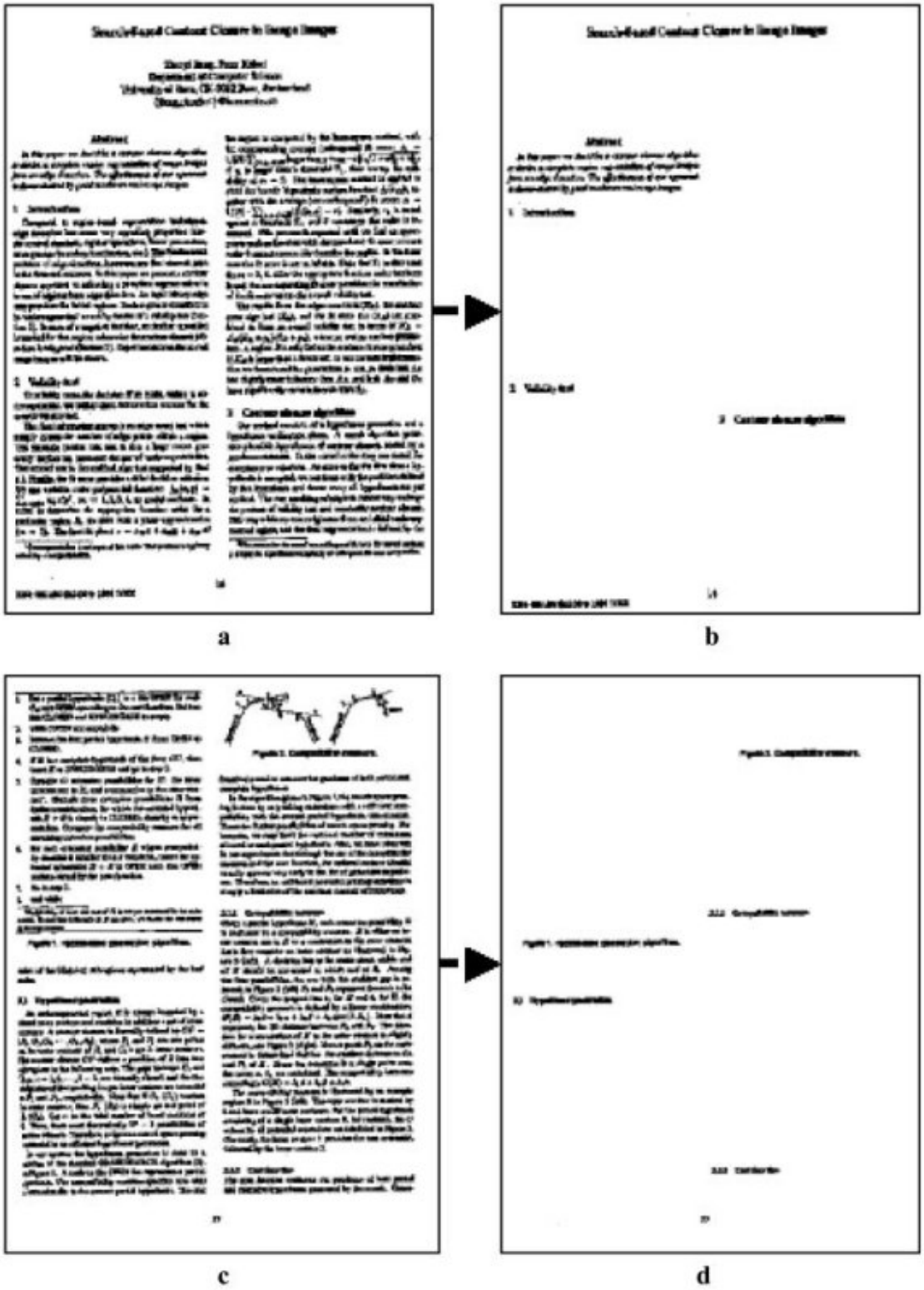

Fig. 13a–d. Extraction of headings and other meta-information: **a** image of a title page of a technical paper **b** extracted portions from the image in part a **c** image of another page of a technical paper **d** extracted portions from the image in part c

Table 1. Database used for testing algorithms

Source	Number of pages
Proceedings of 5th ICDAR	60
Proceedings of 4th ICPR	60
Univ. of Washington English Document Image Database I	35
Noise-free documents	20
Degraded documents	25

Table 2. Identification of word type styles

Font	Recognition rate (%) for				
	Italic		Bold		All-capital
	Char	Word	Char	Word	Word
Times	96.43	99.82	95.70	99.51	99.92
Arial	97.35	99.91	97.55	99.78	100
Courier	95.65	99.57	95.15	99.45	99.93
Gothic	97.43	99.87	97.76	99.81	99.98
Roman	96.69	99.62	96.95	99.69	99.98

Table 3. Identification results for degraded documents

SNR	Recognition Rate (%) for words in		
(in DB)	Italic	Bold	All-capital
20	99.34	99.63	99.83
15	98.67	98.02	98.34
10	96.12	95.52	97.91
5	94.76	93.22	95.05

printed, we note that the recognition rate at the word level is far better than that at the character level. Experiments show that the methods are quite robust on a variety of fonts and font sizes. The operational speed for identifying italic and bold words is about 1.9 Mpixel/s and 1.2 Mpixel/s, respectively. The detection of all-capital words is much faster at a rate of 3.8 Mpixel/s. As a comparison, Bloomberg [3] reports the speed of 1.3 Mpixel/s and 0.5 Mpixel/s, for detection of italic and bold words, respectively.

As mentioned, a model due to Kanungo et al. [12] was used to generate degraded documents. The model accounts for: (i) pixel inversion (from foreground to background and vice versa) that occurs independently at each pixel; (ii) blurring; and (iii) perspective distortions. Pixel inversion occurs due to light intensity fluctuations, sensitivity of the sensors, and the thresholding level, etc., whereas blurring occurs due to the point-spread function of the scanner. Perspective distortions occur when a thick book is scanned under the flatbed scanner. Figure 14 shows a degraded image generated from a portion of a noise-free image. Table 3 shows the recognition rate for italic, bold and all-capital words in the degraded documents. It shows that our algorithms perform with accuracy higher than 95% even for degraded documents having an SNR equal to 10 dB.

Table 4. Determination of character size

Character size (in Points)	% of correct estimation	Errors (Erroneous Pt sizes are in brackets)
6	98.45	0.35% (5 Pt) 1.2% (8 Pt)
8	98.67	0.26% (6 Pt) 1.07% (9 Pt)
9	98.83	0.19% (8 Pt) 0.98% (10 Pt)
10	99.12	0.23% (9 Pt) 0.65% (11 Pt)
11	98.76	0.42% (10 Pt) 0.82% (12 Pt)
12	99.17	0.17% (11 Pt) 0.72% (14 Pt)
14	98.93	0.47% (12 Pt) 0.60% (16 Pt)
16	99.07	0.28% (14 Pt) 0.65% (18 Pt)
18	99.05	0.52% (18 Pt) 0.43% (20 Pt)
20	99.17	0.24% (18 Pt) 0.69% (22 Pt)
24	99.02	0.37% (22 Pt) 0.61% (26 Pt)
36	98.97	0.25% (34 Pt) 0.78% (38 Pt)
48	99.11	0.21% (46 Pt) 0.68% (50 Pt)
72	99.23	0.21% (68 Pt) 0.56% (70 Pt)

Our method of determining character size shows significant accuracy. The results are presented in Table 4 where the percentage values are measured based on the number of words. Here, the errors occur because (2) does not always give the exact character size for all the words. This is due to the following: (i) the x-height of a character may vary from font to font for the same point-size (for sans serif fonts the x-height of a character is slightly larger than the corresponding x-height of the character in serif fonts at the same point size); (ii) the method determines the character size not on a word-basis but on a line-basis. Thus, if there is an isolated word printed in a size larger than the other words in the sentence then it is difficult to detect this larger-sized word. However, in our experiment we have not encountered any word printed in this fashion and it is indeed a rare possibility in a technical document. We have tested our method using both serif and sans serif font families.

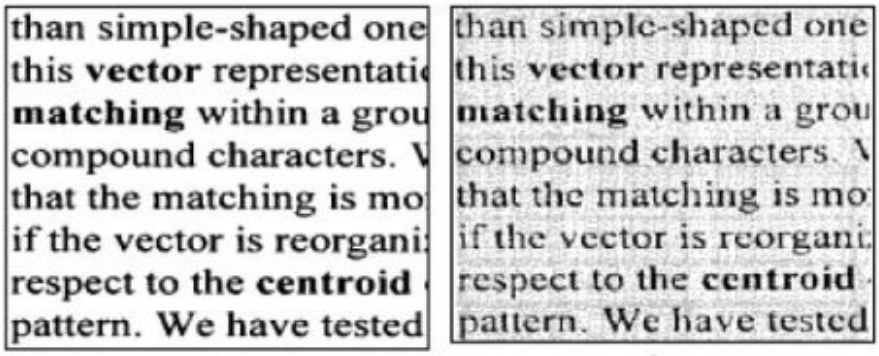

Fig. 14a,b. Example of degraded documents: **a** original noise-free document **b** degraded version of the document in part a

Table 5. OCR accuracy for italic words

Font	Recognition Rates (%)	
	Original	De-italicized
Times italic	99.15	99.75
Arial italic	98.60	99.82
Courier italic	98.55	99.45
Gothic italic	98.70	99.80
Roman Italic	70.65	99.55

Table 6. Extraction of logical labels and other information (results are computed on 40 technical papers)

Type of information	#Items in test data	#Properly extracted items
Paper Title	40	40
Paper Abstract	40	32
Section/sub-section Heading	268	264
Figure Caption	152	121
Table Title	43	29

To test the efficiency of our approach to improve the performance of OCR systems, we took pages that contain a reasonable number of italic words. At first, a document is processed by an existing OCR system and the recognition output for the italic words is checked. For this purpose, we used the popular OCR system 'OmniPage Version 4.0'. Next, our method for identifying italic words is applied to the document. Once identified, the italic words are de-italicized by the estimated slant angle, fed to the OCR system, and the output is compared with the earlier one. The details of this comparison are shown in Table 5.

Our method for extracting logical labels from the technical articles is tested on 40 technical papers taken from the proceedings of the 14th ICPR, the 5th ICDAR, and from some other sources listed in Table 1. Here, emphasis was put on extracting five types of information, namely: (i) paper title; (ii) paper abstract; (iii) section/sub-section headings; (iv) figure captions; and (v) table titles. The results of extracting these labels are shown in Table 6. It is noted that all the paper titles are extracted from each of the 40 papers with 100 percent accuracy. In 32 papers, the paper abstracts are typed in italic style and they are extracted properly. However, for the remaining eight papers, abstracts are not extracted because they are printed in normal style and size.

Out of a total of 268 section and sub-section headings, 264 headings are extracted accurately. This shows 98.5% accuracy in extracting different label-headings from an imaged document. Our algorithm fails to extract four section-headings properly. Of these, in two cases, the heading is partly extracted, but in the two other cases complete failure is noticed. On the other hand, for three cases, normal text lines are misidentified as sub-section headings. It is observed that each of these three lines contains a few acronyms typed in all-capital characters and, hence, the error in identification has occurred. Similarly, 121 out of 152 figure captions and 29 out of 43 table titles are properly extracted. They are extracted based on only the white space around them and word type style information. Use of other layout information will increase the accuracy in extracting figure captions and table titles.

Earlier we stated our observation that italic, bold, and all-capital words typically have special significance in a document. To verify this, we considered 40 technical articles where each article is treated as a single document. For these documents, our methods for identifying italic, bold, and all-capital words are first applied. Once identified, these words are extracted and tagged with the respective document name to which they belong. Next, the documents are processed by an OCR system and are converted into computer-readable text. For each of these document texts, the term frequency for each word is calculated. A list of words (L_w) is built for each document and is sorted according to the descending order of term frequencies. L_w excludes the stop words that are maintained in a predefined list of such words. To test the importance of the previously extracted italic, bold, and all-capital words, the ranks of these words are manually checked in L_w for each of the documents under consideration. We have observed that on an average 3.1 italic, bold or all-capital words are included in the list of the ten most frequent words in the documents. A summary of the results of this experiment on 40 documents is given in Table 7. Table 8 presents more detailed results for ten documents.

Table 7. Summary of results on the importance of italic, bold, and all-capital words

Doc. ID	Total number of words	# Italic, bold and all-capital words	# Italic, bold, and all-capital words in the list of top N mostly frequent words in the document		
			N = 10	N = 20	N = 50
40	322,680	5,240	3.1	4.7	10.6

Table 8. Results of the experiment on the importance of italic, bold, and all-capital words

Doc. ID	Total number of words	# Italic, bold and all-capital words	# Italic, bold, and all-capital words in the list of top N mostly frequent words in the document		
			N = 10	N = 20	N = 50
Doc-1	8,567	239	5	6	17
Doc-2	7,579	193	4	7	15
Doc-3	5,609	151	4	6	12
Doc-4	6,258	105	3	5	12
Doc-5	8,105	212	3	5	15
Doc-6	8,073	121	3	5	11
Doc-7	5,934	113	3	4	9
Doc-8	7,003	131	3	5	10
Doc-9	4,701	93	3	4	9
Doc-10	3,145	55	2	4	9
Total	64,974	1,413	33	51	119

5 Discussion

Some simple and efficient algorithms for the automatic detection of italic, bold, and all-capital words have been developed. All approaches are applicable for text in the English (Roman) alphabet only. In this paper, we do not consider numerals and other symbols. The algorithms are largely font and size independent. Actual character recognition is not required in detecting the words. The application potentials have been discussed in the context of spotting important terms and for extracting useful meta-information from documents. Our experiment for spotting important terms based on type styles shows that the approach can be used to detect keywords for automatic indexing into a database of scanned document image. On the other hand, de-italicization of italic words shows significant improvement in OCR accuracy.

IR systems [16, 21] are mostly used to extract information from computer readable text. Attempts have been made recently for IR from paper-based documents. One possible approach is to integrate the OCR device with IR systems. Studies by Taghva et al. [26] are directed towards the integration of OCR and IR technologies. Their studies indicate that the traditional IR approaches are also applicable for OCR-generated text. However, the typographical aspects (or format information) are lost once the documents are OCRed. If identified, these aspects can be used to improve the IR efficiency from the paper documents. For example, type style and the font size of a word typically indicate its special significance in a document.

At present, we are trying to outline an efficient approach for integrating type style and font information in the IR from paper-based documents. Our approach is centered on the vector space model [20] where a document is represented as a list of terms or keywords with associated weights. The words or phrases are considered as terms and the weight corresponding to a term is a measure of its importance in representing the information in the given document [19, 22].

In our approach, the weight of a term is incremented by some value based on its style and size information. The determination of the amount of increment is subject to rigorous experiments. Our initial approach for modifying the term weights has been tested on a document database consisting of 200 document files and 20 queries. All these documents are OCR-generated from their respective paper documents. We compare our method of assigning term weights with that of a vector space model. It is observed that the term weights to the terms having a type style different from normal and having a size larger than the predominant size, are assigned in a more judicious way than they were just after applying the vector space model. These modified weights change a document's rank in a positive way and improve the retrieval efficiency against a given query. However, to reach a final conclusion, our approach has to be tested against a database much larger than the one we have used. We plan to report our final results elsewhere.

Acknowledgements. The authors would like to thank Dr. M. Mitra for some useful discussions on this work.

References

1. Baird HS.: Document image defect models. In: Proc. IAPR workshop on Syntactic and Structural Pattern Recognition. Murray Hill, N.J., USA, pp. 38–46, 1990
2. Baird HS, Nagy G.: A Self-Correction 100-Font Classifier. In: Proc. SPIE Conf. on Document Recognition, pp. 106–115, 1994
3. Bloomberg DS.: Multiresolution Morphology Analysis of Document Images. In: Proc. SPIE Conf. on Visual Communications and Image Processing, 1818:648–662, 1992
4. Chaudhuri BB, Garain U.: Detection of Italic, Bold and All-Capital Words in Document Images. In: Proc. 14th Int. Conf. on Pattern Recognition (ICPR), 1:610–612, 1998
5. Chen FR, Bloomberg DS.: Summarization of Imaged Documents without OCR. In: Comput. Vision Image Understanding, 70(3):307–320, 1998
6. Chen FR, Wilcox LD, Bloomberg DS.: Detecting and locating partially specified keywords in scanned images using hidden Markov Models. In: Proc. Int. Conf. on

Document Analysis and Recognition (ICDAR), pp. 133–138, 1993
7. Craven.: Learning to extract symbolic knowledge from the World Wide Web. In: Internal report, School of Computer Science, CMU, 1997
8. Doermann D, Rivlin E, Rosenfeld A.: The function of documents. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), Germany, 2:1077–1081, 1997
9. Doermann D.: The Indexing and Retrieval of Document Images: a Survey. In: Comput. Vision Image Understanding, 70(3):287–298, 1998
10. Green E, Krishnamoorthy M.: Recognition of Tables Using Table Grammars. In: Proc. Graphics and Table Recognition Workshop, Pennstate, pp. 261–277, 1995
11. Jain AK, Yu B.: Document Representation and Its Application to Page Decomposition. In: IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 20(3):294–308, 1998
12. Kanungo T, Haralick RM, Phillips I.: Non-linear local and global document degradation models. In: Int. J. Imaging Syst. Technol., 5(4):220–230, 1994
13. Khoubyari S, Hull JJ.: Font and Function Word Identification in Document Recognition. In: Comput. Vision Image Understanding, 63(1):66–74, 1996
14. Kim M, Kwon Y.: Multi-font and Multi-size character recognition based on sampling and quantization of an unwrapped contour. In: 13th Int. Conf. on Pattern Recognition, ICPR, vol. 3, Track C, pp. 170–174, 1996
15. Lamport, L.: LATEX: A document Presentation System. Addison-Wesley, Reading, Mass., USA, 1986
16. Lancaster FW.: Information Retrieval Systems: characteristics, Testing and Evaluation. Wiley, New York, 1968
17. Microsoft Word 97: Copyright©1983-1996, Microsoft Corporation, USA
18. Pavlidis T, Zhou J.: Page Segmentation and Classification. In: CVGIP: Graphical Models Image Process., 54:484–496, 1992
19. Salton G, Yang CS, Yu CT.: A theory of term importance in automatic text analysis. In: J. Am. Soc. Inf. Sci., 26(1):33–44, 1975a
20. Salton G, Wong A, Yang CS .: A vector space model for information retrieval. In: J. Am. Soc. Inf. Sci., 18(11):613–620, 1975b
21. Salton G, McGill MJ.: Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983
22. Salton G, Buckley C.: Term-weighting approaches in automatic text retrieval. In: Inf. Process. Manage., 24(5), pp. 513–523, 1988
23. Shi H, Pavlidis T.: A System For Text Recognition Based On Graph Embedding Matching. In: Proc. of Int. Assoc. for Pattern Recognition Workshop on Document Analysis System (DAS), pp. 413–427, 1996
24. Shi H, Pavlidis T.: Font Recognition And Contextual Processing For More Accurate Text Recognition. In: Proc. 4th Int. Conf. on Document Analysis and Recognition (ICDAR), 1:39–44, 1997
25. Sun C, Si D.: Skew And Slant Correction For Document Images Using Gradient Direction. In: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), 1:142–146, 1997
26. Taghva K, Borsack J, Condit A.: Information Retrieval and OCR. Handbook of Character Recognition and Document Image Analysis. H. Bunke, P.S.P. Wang (eds), World Scientific, Singapore, pp. 755–777, 1997
27. Tsirikolias K, Mertzios BG.: Statistical Pattern Recognition Using Efficient two-dimensional Moment With Application To Character Recognition. Pattern Recognition, 26:877–882, 1993
28. Zramdini A.: Study of Optical Font Recognition Based on Global Typographical Features. Ph.D. Thesis, Institute of Informatics of the University of Fribourg, Switzerland, 1995

B.B. Chaudhuri is currently the head of Computer Vision and Pattern Recognition Unit of Indian Statistical Institute, India. His research interests include Pattern Recognition, Image Processing, Computer Vision, Natural Language Processing and Digital Document Processing including OCR. He has published about 200 research papers in reputed International Journals and has authored the books entitled *Two Tone Image Processing and Recognition* (Wiley Eastern, 1993) and *Object Oriented Programming: Fundamentals and applications* (Prentice Hall, 1998). Professor Chaudhuri received many awards and prizes including Sir J.C. Bose Memorial Award (1986), M.N. Saha Memorial Award (twice: 1989, 1991), Homi Bhabha Fellowship award (1992), Dr. Vikram Sasabhai Research Award (1995), and C. Achuta Menon Prize (1996) for his contribution in the field of Engineering sciences, Indian language processing, computer applications, etc. Professor Chaudhuri is a Senior member of IEEE, member secretary of IAS (Indian Section) and Fellows of IAPR, Institution of Electronics and Telecommunication Engineering (India), National Academy of Sciences, and National Academy of Engineering (India). He is serving as associate editor of Pattern Recognition, Pattern Recognition Letters (Elsevier Sciences) and VIVEK as well as guest editor of a special issue of Journal IETE on Fuzzy Systems.

U. Garain received both of his B.E. and M.E. in Computer Science and Engineering from Jadav pur University, Calcutta in 1994 and 1997, respectively. He worked in two multinational software firms for one year and a half. Later on he joined as a research personnel in Indian Statistical Institute, Calcutta, where he is a full-time faculty now. He is one of the key scientists involved in the development of a bilingual (Devnagari & Bangla) OCR system. For last two years Mr. Garain has published several technical papers in reputed international journals and conferences. His areas of interest include digital document processing, OCR system development for Indian language scripts, document data compression, etc.