

Automated Detection and Segmentation of Table of Contents Page from Document Images

S. Mandal, S. P. Chowdhury, A. K. Das
CST Department B. E. College (D.U.)
Sibpur, Howrah 7111 103
sekhar, shyama, amit@becs.ac.in

Bhabatosh Chanda
ECS Unit, Indian Statistical Unit
Calcutta 700 035
chanda@isical.ac.in

Abstract

With an aim to extract the structural information from the table of contents (TOC) to help develop digital document library the requirement of identifying/segmenting the TOC page is obvious. The objective to create digital document library is to provide a non-labour intensive, cheap and flexible way of storing, representing and managing the paper document in electronic form to facilitate indexing, viewing, printing and extracting the intended portions. Information from the TOC pages be extracted to use in document database for effective retrieval of the required pages. In this paper we present fully automatic identification and segmentation of table of contents (TOC) page from scanned document.

Keywords: Document image segmentation, Table of contents detection, Digital document library.

1. Introduction

Table of contents (TOC) detection from scanned document pages is important for a user of the digital document library as an index for the contents of the books, journals, and reports etc. It is also necessary for the document database in the library to keep structural information, like chapters, sections and subsections for easy retrieval of the intended portions as demanded by the user. As a result the identification/segmentation of the TOC from the scanned pages has attracted researchers [16, 2] to put forward a couple of schemes to do the same.

It has been observed from the existing literatures that most of the works are directed toward higher level understanding of the TOC page so as to extract the structural information and representing the whole to a suitable meta-structure like HTML or XML etc. [16]. In doing so they assume that either the TOC is already segmented or some sort of character/symbol recognition technique is applied

to identify the TOC pages and their underlying structures. Though, symbol recognition is a part of OCR activity when it is applied to the non-segmented mixed material (text with math-zone and others) computation will be expensive and success far from satisfactory.

We on the other hand contend that a better approach is to identify the TOCs from the mixed material thereby helping the subsequent image processing and OCR activities to focus its processing only on the respective zones. In this paper we propose a fully automated technique for identification of TOC page or the portion of TOC in the text page exploiting *a priori* knowledge of the underlying structure of possible types of TOCs in books, journals etc. It may be noted that we did not use any type of symbol recognition techniques for identification/segmentation of TOCs.

Our goal is to identify whether a scanned page or its part is TOC or not using a top down approach that starts with an expectation of encountering a couple of structures available in common TOC forms. And after identification we segment the TOC to its constituent parts into number, title and corresponding page number of each section and subsections to facilitate OCR and help search and browse the document database. We assume that the input is the text portion which have been already segmented from mixed objects in a document page like, text, graphics and half-tones [4].

1.1. Past works

There are a number of schemes for page layout analysis and segmentation [10, 9, 4, 1, 3, 8]. Most of the works are directed towards segmentation of text, graphics and half-tones. [12, 10, 9] did not go further to extract tables and other structures from the document. In the system *CyberMagazine* Takasu et al. [15] proposed segmentation of blocks and syntactic analysis of their contents. Article recognition is done using a decision tree classifier and a matrix grammar based syntactic analysis. In [11, 14] O’Gorman and Story proposed method for TOC structure extraction in their Right Pages Electronic Library Systems

utilising OCR techniques. By 'Docstrum' analysis blocks are first extracted and then the articles are indexed with the help of a *a priori* model, fed manually, of the TOC of different journals. The relationship information is used to help human user to quickly go to the relevant sections (say from article title to the exact pages) on mouse clicks. Part of speech tagging, a labelling approach, for automatic recognition of TOCs is done by Belaid et al. [2] utilising an *a priori* model of the regularities present in the document structures and contents. POS tagging technique is directly applied to the text files produced by the OCR without any preprocessing.

In a recent work Tsuruoka et al. [16] proposes a method of image-based structure analysis and conversion of TOC of books into XML document. As observed in the TOCs, structural elements such as chapter, section etc. are featured by the extent of indentation and the size of characters. These are used to classify the text lines into different groups. A structure tree, prepared in advance, is related with the groups for the XML conversion. However this technique is limited only to the TOCs of books with discernible difference in the indentation and fonts used in the TOCs.

This paper is organised as follows. Section 2 describes the proposed work, section 3 has dealt with the experimental result while concluding remarks is given in section 4.

2. Proposed Work

The work carried out for automatic segmentation of TOC page or the portion of the TOC in a page provides a solution by developing morphology based algorithms to i) convert the text lines to word halos using morphological operations, ii) detect the rightmost word from every text line and counting the number of characters in it, iii) identify TOCs and iv) segment TOCs. This work is the continuation of earlier work on segmentation where the document image containing text, graphics, half-tones, tables and headings have been identified and segmented [4]. The work started with the gray scale image of the page with text, half-tones and graphics. The half-tone is removed first [4]. The image is then binarised [13] and skew corrected [7]. Utilising the power of morphological operators to extract shape based features we isolated tables and headings from the document image [5]. This facilitates segmentation of regular text and graphics from residue image. Moreover, extraction of tables and headings from the text improves the possible OCR operation and interpretation of the segmented text. Finally text and graphics are segmented [6]. Here we start with the segmented text which may contain TOC.

2.1. Observation

The TOC is nothing but text lines with a structured format. Hence to formulate TOC finding rules we scanned 75 pages containing TOCs' from books, journals and reports. We summarize the following characteristics of TOCs.

1. TOCs' may be available as a full page; this is the case for books and reports. It may be a part of a page; some journals have their TOCs in the front page along with LOGO, headings and other texts.
2. Rightmost columns usually contain the page numbers. This kind of right-alignment, however, in many TOCs are not maintained and the page number appears just after the name of the sections and subsection headings or the name of the articles. The page numbers in this case is printed with enough gap and in certain cases the same is printed with gaps slightly more than the normal word gaps.
3. Leftmost columns usually have the section and subsection numbers and the majority of the text lines (subsection headings) begin from a particular column and the gap between the title and section/sub-section number is much more than the normal word gaps.
4. Many dotted lines (or thin lines) may appear at regular intervals. This is used to show the correlations between the end of the text line indicating name of the articles etc. and the corresponding page numbers at the rightmost columns.
5. TOC page may have multiline title. Many TOCs are available with horizontal separator lines in the beginning and sometimes at the end.
6. TOCs have some characteristics similar to the tabular structure.
7. Page numbers are short (usually 2 to 4 characters, mostly 3 for average books) and the number of characters used to denote page numbers in a particular TOC page has very little variation for obvious reason.

The format information of TOCs has led to the formulation of the rules for finding out TOCs from the text. Though TOCs have many possible structures they can be classified into two main types, namely, TOC-I and TOC-II. In TOC-I page numbers are printed in right aligned form and in TOC-II they appear, without any alignment, after the name of the section/subsection or the author's name as shown in fig. 1. It may be noted that certain observations are deliberately ignored and some has no bearing with the technique used to identify the TOCs. For example, in TOCs, page numbers follow an ascending sequence may not be utilised unless we

use OCR technique. Similarly detection of a series of dotted lines, a feature unique to some of the TOCs, is ignored as the page numbers are invariably right aligned. Moreover the section and subsection numbers, which do appear, in the left hand side may not be used due to possible conflict of pages containing exercises.

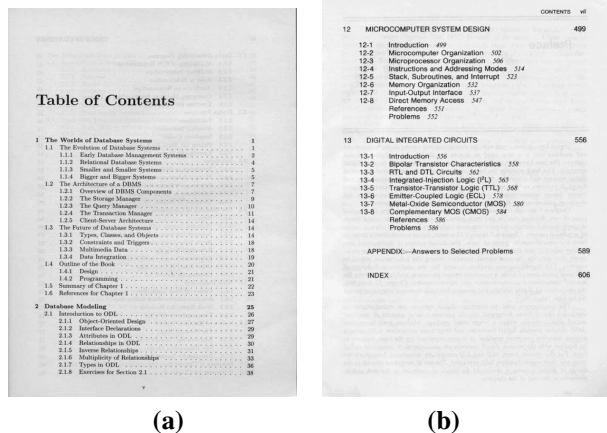


Figure 1. Two types of content pages; (a) TOC-I, Right aligned page numbering; (b) TOC-II; Distributed page numbering.

2.2. Identification steps

Identification of TOCs' are based on finding page numbers associated with the name of sections, sub-sections or articles/author. Note that the page number, considered as word as a whole, will be available¹ as the rightmost word of a text line. However, as an added complexity the rightmost word of all the lines may not be the page number as the text associated with a sub-section or article may spread over a couple of lines. We elaborate below the steps to find out the rightmost word and counting the number of characters in those words.

1. **Find word halos:** This is done by successive closing with structuring elements (SEs) of increasing area and component counting of the closed image. As inter-character gaps are much less than inter-word gaps the component count after couple of iteration falls sharply indicating a coalescing of characters (in individual words) forming word halos. Since this step is terminated based on the characteristics of the input image variable gap between the words does not pose any problem. Figure 2 shows line of word halos of a part of the TOC page (see fig. 1(b)). Each text line are then

separated by taking horizontal projection of the line of word halos.



Figure 2. Line of word halos.

2. **Find the rightmost word:** For an isolated text line with word halos the rightmost word is located by vertical scan. We consider the rightmost word as a possible candidate for page number if it is preceded by at least another word and count the characters by component labelling.
3. **Detection of TOCs:** After the common steps described earlier we go for detecting TOCs (both type I and II) with the help of a decision tree as shown in fig. 3.

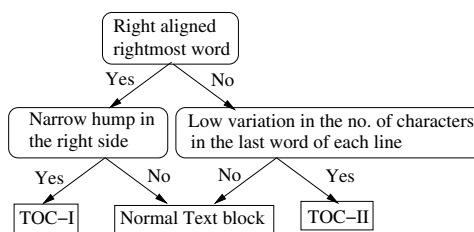


Figure 3. Decision tree to identify two types of content pages; TOC-I and TOC-II.

We elaborate below the logic used in the decision tree in order to identify TOC-I and TOC-II.

TOC-I: In this case the rightmost word is right aligned. There may be a lot of gap between the page number and the previous word. Sometimes dotted or thin lines are used as markers between the section/author etc. and the page numbers. However for TOC-I the right aligned page numbers will have their reflection in the vertical projection in the form of an isolated narrow hump at the rightmost section (see fig. 4(a)). To facilitate hump detection we compute the median of the pixel count and using the median value as the threshold we remove regions (see fig. 4(b)) which came due to the dotted lines or thin lines commonly used in TOCs as marker between the names and corresponding pages.

TOC-II: This type is characterised by the presence of page numbers at the end of almost all the lines but they are

¹This is by and large true except a small number of cases

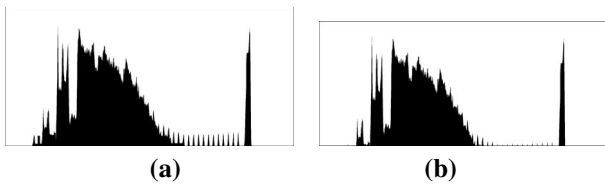


Figure 4. Result of vertical projection of fig 1(a); (a) Before threshold; (b) After threshold.

not printed in right aligned form. As a result there is no tall protruded part (see fig. 5(a)) in the right hand side. Detection is done by checking the variation in the number of characters in the rightmost word of each line (see fig. 5(b)).

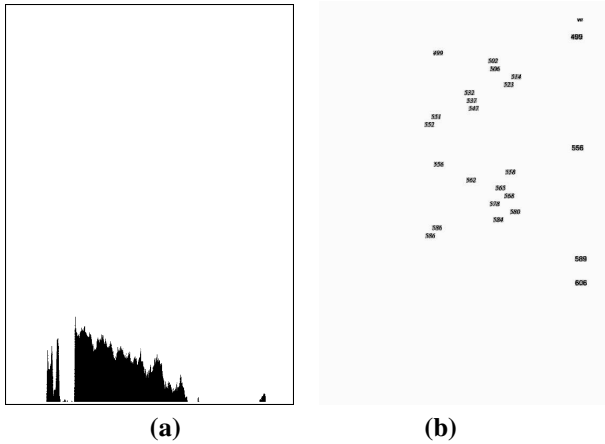


Figure 5. Features of TOC-II; (a) Vertical projection of the binary version of the image shown in fig. 1(b); (b) Spatial distribution of the rightmost words and the number of characters in each word of the image shown in fig. 1(b).

2.3. Segmentation of TOC

After detection of the TOC we segment each line into three parts; namely number, title and page number and we put the output information in a table as shown below. The table has three pointer fields pointing to the bounding box containing the number, title and page number that may be present in each text line in the TOC. This is done by taking the output of the step 1 for identification where each line is available as a line of word halos as input. Next the words are counted and coalesced by closing the band with SEs of increasing size. As the section (sub-section/chapter) number

Number	Title	Page number
L_{10}	L_{11}	L_{12}
L_{20}	L_{21}	L_{22}
...
L_{n0}	L_{n1}	L_{n2}

and page number are printed on either sides of the section names with more than the usual word gap the closing operation will combine the name portion to a single component leaving the number and page number portions unaltered. So in most of the cases we get three components; two short components at both ends and one wide in the middle. However it may also be noted that due to less gap between the title and page number portion for TOC-II we may get two components instead of usual three. We use a decision tree (see fig. 6) to put them in the output table.

A sample result of the TOC segmentation is also shown in the figure 7.

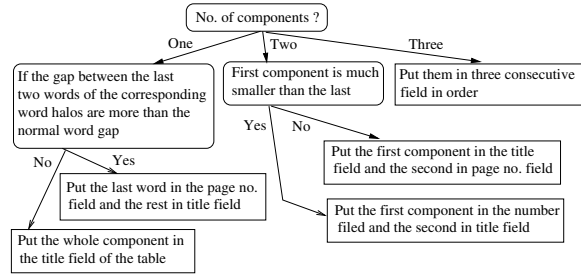


Figure 6. Decision tree for TOC segmentation

L_{10}	→ 12
L_{11}	→ MICROCOMPUTER SYSTEM DESIGN
L_{12}	→ 499
L_{20}	→ 12-1
L_{21}	→ Introduction
L_{22}	→ 499
	12-2
	Microcomputer Organization
	502
	12-3
	Microprocessor Organization
	506
	12-4
	Instructions and Addressing Modes
	514
	12-5
	Stack, Subroutines, and Interrupt
	523
	12-6
	Memory Organization
	532
	12-7
	Input-Output Interface
	537
L_{90}	→ 12-8
L_{91}	→ Direct Memory Access
L_{92}	→ 547
L_{101}	→ References
L_{102}	→ 551
	Problems
	552

Figure 7. Segmentation result showing segmentation of a part of the TOC of fig. 1(b).

It may be noted that each line may or may not be partitioned into three fields. For example we see in fig. 7 that line number 10 has two parts; title and page number.

3. Experimental results

Here we present results to identify and segment the TOCs from the dataset. Note that all experiments were carried out in a COMPAQ DS 20E workstation and all the programs are written in C. The performance of the system is evaluated based on the manual ground-truthing as well as manual checking. We have used document pages from University of Washington's document image database (UW-I, II, III) and our own collection. The dataset contains 143 TOCs of different kinds. Average page is approximately of 'letter size' and the pages are scanned at 300 dpi. We could identify 137 TOCs and no non-TOC pages were detected as a TOC. Average computation time is about 1.8 second per page. Though failure rate is small we would like to highlight the exceptional cases where our algorithm fails.

- Content pages with page numbers appearing on the left hand side of the document.
- Tabular structure without any separator lines and spreaded over whole width as well as presence of a narrow rightmost column.

It may also be noted that we did not try to combine several lines of a multi-line title (section/sub-section name etc.) to a single string and put it in the name field of the table. As a result we may have table entries with null pointers in the number and page number field for a particular line. We have assumed that the OCR or the search/browse engine will provide the logic to sort out the problem. As declared earlier that the TOCs have a fair degree of similarity with tabular structure and chance of mis-identification is high. We avoided this phenomena, by and large, as our table detection logic has built in protection against it [5].

It may be noted that our treatment is based only on the spatial distribution of the connected components and low resolution image may not greatly affect the performance of the segmentation. We have tested this by a 4 fold reduction of the resolution (300 dpi to 75 dpi) of the input images and we got a 7 fold timing improvement with a minimum degradation in segmentation performance.

4. Conclusion

We conclude this paper by pointing out the uniqueness of our work which includes i) identification and segmentation of the table of content (TOC) without any character recognition steps, ii) low computation cost, and iii) detection of TOCs without any explicit training or backtracking

if the page contains fair amount lines with associated page numbers.

References

- [1] O. T. Akindele and A. Belaid. Page segmentation by segment tracing. In *ICDAR93*, pages 341–344, Tsukuba, Japan, 1993.
- [2] A. Belaid, L. Pierron, and N. Valverde. Part-of-speech tagging for table of contents recognition. In *ICPR 15th International Conference*, pages 4451–4454, Barcelona, Espagne, 2000.
- [3] S. Chen. *Document Layout Analysis Using Recursive Morphological Transforms*. PhD thesis, University of Washington, 1995.
- [4] A. K. Das. *Document Image Segmentation: A morphological approach*. PhD thesis, Bengal Engineering College (Deemed University), Sibpur, India, 1998.
- [5] A. K. Das and B. Chanda. Detection of tables and headings from document image: A morphological approach. In *International Conf. on Computational linguistics, Speech and Document Processing (ICCLSDP'98); Feb. 18–20, Calcutta, India*, pages A57–A64, 1998.
- [6] A. K. Das and B. Chanda. Segmentation of text and graphics from document image: A morphological approach. In *International Conf. on Computational linguistics, Speech and Document Processing (ICCLSDP'98); Feb. 18–20, Calcutta, India*, pages A50–A56, 1998.
- [7] A. K. Das and B. Chanda. A fast algorithm for skew detection of document images using morphology. *Intl. J. of Document Analysis and Recognition*, 4:109–114, 2001.
- [8] K. C. Fan, C. H. Liu, and Y. K. Wang. Segmentation and classification of mixed text/graphics/image documents. *Pattern Recognition Letters*, Vol. 15:1201–1209, 1994.
- [9] Y. Ishitani. Document layout analysis based on emergent computation. In *ICDAR97*, pages 45–50, 1993.
- [10] A. K. Jain and B. Yu. Page segmentation using document model. In *Proc. ICDAR'97, Ulm, Germany*, pages 34–38, August, 1997.
- [11] L. O'Gorman. Image and document processing techniques for the right pages electronic library system. *ICPR*, 2:260–263, 1992.
- [12] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, No. 11:1162–1173, 1993.
- [13] N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. SMC*, 9, No. 1:62–66, 1979.
- [14] G. A. Story, L. O'Gorman, D. Fox, L. L. Schaper, and H. V. Jagadish. The right pages image-based electronic library for alerting and browsing. *Computer*, 25, No. 9:17–26, September 1992.
- [15] A. Takasu, S. Satoh, and E. Katsura. A rule learning method for academic document image processing. *ICDAR*, 1:239–242, 1995.
- [16] S. Tsuruoka, C. Hirano, T. Yoshikawa, and T. Shinogi. Image-based structure analysis for a table of contents and conversion to a xml documents. In *Workshop on document layout interpretation and its application (DLIA 2001)*, Sep 9, 2001, Seattle, Washington, USA, 2001.