March-April 2020
ISSN: 0193-4120 Page No. 2404 - 2410

Image Recognition and Segregation using Image Processing Techniques

Parita Jain[1], Amit Singhal[2], Diksha Chawla[3], Vineet Shrivastava[4]
[1]KIET Group of Institutions, Ghaziabad, India
[2,4]RKGIT, Ghaziabad, India
[4]PDM College of Engineering, Haryana, India
1paritajain23@gmail.com, [2]amitsfcs@rkgit.edu, [3]dikshu_chawla@yahoo.co.in, [4]vineetfcs@rkgit.edu.in

Article Info
Volume 83
Page Number: 2404 - 2410
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 19 March 2020*

Abstract:
Document image structure analysis is to achieve computerized document management and storage and retrieval. The main objective is to identify different pages of books, journals and reports and segregate it from other document images. This process helps to create a database of digital documents. It provides an immediate solution to enable the archived valuable materials so that it can be searchable and usable by users in order to achieve its objective. It also provides way of representing and managing the paper document in electronic form. For, the purpose of viewing, indexing and extracting the intended portions. The proposed technique work on every aspect of image including its size, shape and orientation of image.

Keywords: *Threshold, ROI, Histogram, Centeriod, Projection.*

I. INTRODUCTION

All books and documents contains table of contents usually abbreviated as TOC. It represents metadata for the whole content of book. The contents usually divide in to three level formats such as, chapter's main title, section titles within the chapters and subsection titles. Table of contents are part of books, reports and journals, etc. There are various kinds of documents exists specifically table of contents which differs from each other language wise and quality wise These are all used in various documents, books, magazines and journals of different types and formats. All types of printed table of contents indicate page numbers where each part starts. In order to recognize table of content pages and differentiate it from other document images without using any character recognition technique to provide digital document library effective way to store and manage database. This is cheap, flexible and effective way to store and manage the huge collection of database. Different types of table of contents are collected from books, journals, reports, magazines and from open internet. Tables are important part of any document technical or non-technical for representing data in structural format for this purpose correct detection of structural feature are important.

II. PAST WORKS

Previous work done by Megan Elmore, Margaret [21] on image preprocessing. They work on text detection, autorotation and noise cleaning which prepare document for further analysis. Liangcai automatic TOC analysis method has been proposed by GAO and Zhi Tang [3] proposed a method of clustering for detection of table of content pages. According to Young-Bin kwon and jaehw [4] the article titles, author names and page numbers from scanned image of a journal's table of content scan

be extracted with the help of automated segmentation method. A knowledge based method to recognize table in scanned documents. This approach works on predefined information about which table types may appear. A non-labor intensive, cheap and flexible way of storing, representing and managing the document in electronic form to facilitate indexing and extracting the intended portions proposed by S. Mandal, A.K Das and S.P Chowdhury [1]. Tsuruoka et.al [2] used the indentation and font size to extract the structural elements such as chapters and sections in a book. To deal with layout variance some researchers applied functional knowledge to the TOC recognition [3]. Wolfgang Tersteengen [5] proposed a method that is very accurately able to cope up with distorted tables providing little layout information for example number of lines, bad alignment or few rows.

III. PROPOSED SOLUTION

In proposed approach, the work has been carried out for identification of Tabular structure page from document images. The overall workflow is divided into two parts (i) TOC detection and (ii) TOC recognition by applying (a) Horizontal projection to separate individual line row wise by identifying spaces(b) Vertical projection [20] which is used to identify gaps in txt lines column wise. This technique is based on identifying geometrical concepts of image and used those

concepts to segregate TOC page from other document pages of books, journals and reports.

A. Observations

Toc's from different online and offline sources have few characteristics in common these can be summarizing as:

- Table of content page may have multiline title.
- Table of content have characteristics similar to tabular structure.
- Right most columns usually contain page numbers and left most columns usually have section and subsection numbers.
- Many dotted lines or gaps may appear at regular intervals.
- Table of content pages are usually divided into three column format: Right column usually have page numbers, left columns usually have section and subsection number and middle column include title of chapters.

Based on the above observations our approach mainly classified into two table of content pages: TOC-I and TOC-II.

B. Identification steps

This paper presents a technique for TOC identification by taking projection profile and measuring space in between page.

TOC Detection - Detection of Table of content Page is done by applying binarization and taking ROI.

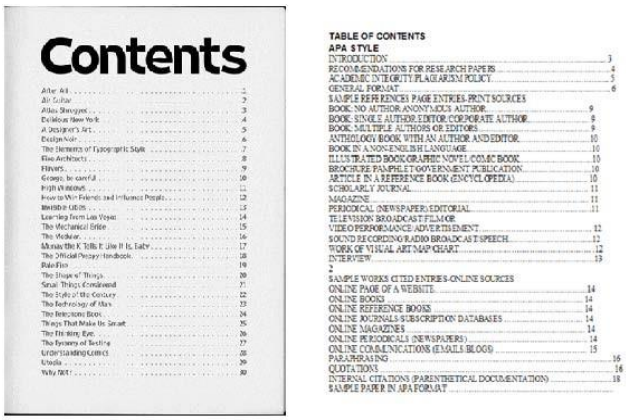


Fig. 1. Two types of table of content pages ;(a) TOC-I; Right aligned page numbering ;(b) TOC-II; Distributed page numbering.

March-April 2020
ISSN: 0193-4120 Page No. 2404 - 2410

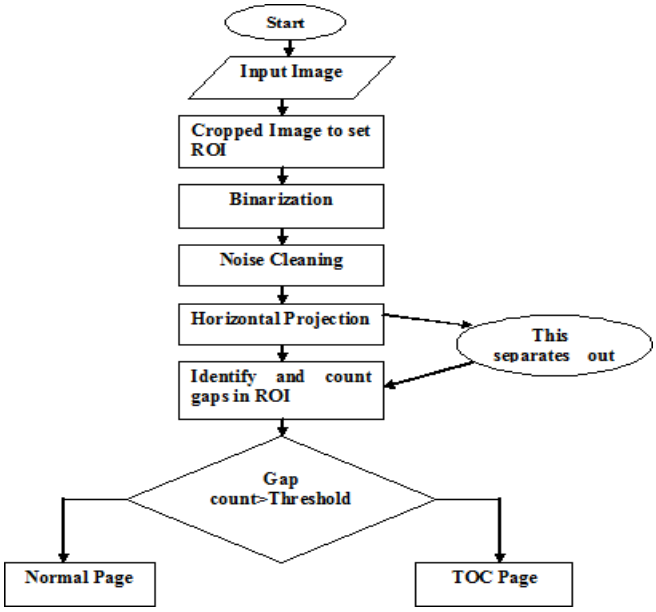


Fig. 2. Flowchart of TOC Detection

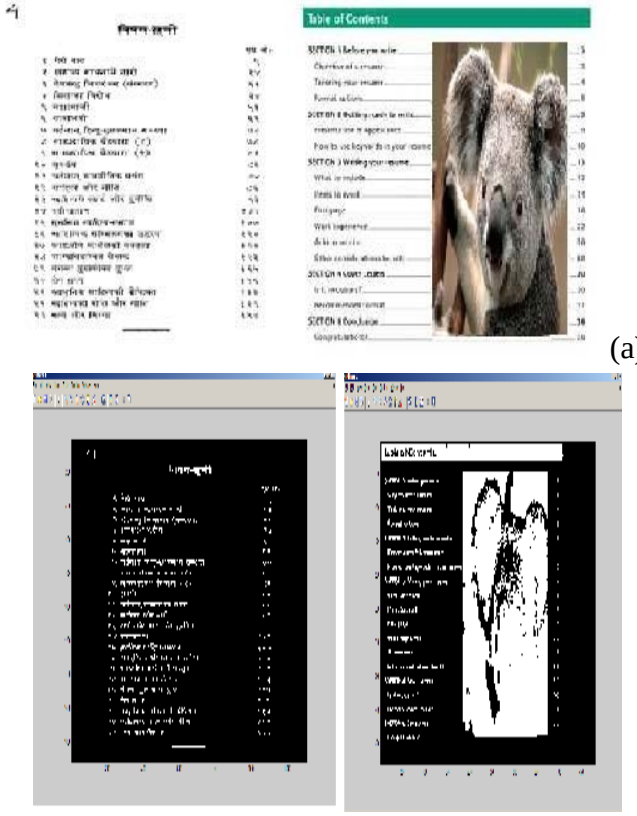
Procedure for identification of TOC page can be described

TOC-I in this the page numbers are on right side and aligned. In this dot between title and page number are used to separate different columns. Gap count between three columns: section number, title and page number and be identified using vertical projection. We can differentiate TOC's from normal pages by considering total space between columns.

TOC –II in this type of Table of content page the main challenge is to identify equal gaps between columns. As page numbers are not aligned in this case gap count of each line is different.

Preprocessing operation refers to processing an image for improving quality of an image. Binarization of input image and noise removal are the two main preprocessing operations. There are two methods of Binarization (i) Global threshold (ii) Otsu (local) thresholding method [9]. In this paper, Otsu thresholding method [6] is applied. Global thresholding method [8], does not take in to account localization effects. For example, if a pixel is say of intensity 150 in gray scale [15] and if global threshold calculated is 200 then it will classify that pixel as a black. However, it is possible that the neighboring 8 pixel of the above one are all of intensity <50.In that case, although that pixel qualifies as a foreground pixel, we wrongly classify it based on our global threshold [10].

Otsu method finds a normalized threshold which basically caters to all these localization effects. Hence, in above case the pixel is classified as a foreground pixel than background.



(a)

Fig. 3. (a) Document pages with images embedded in between

(b) Binarization Results of images

Noise Cleaning -Images taken from different sources like cameras and scanners will pick up noise. Images are collected from different sources they are often different in color or intensity. For noise cleaning we use two operations:(i) To expand or reduce size of object Also the binary image contains connected set of 1's which can be identified using dilation is used.

March-April 2020
ISSN: 0193-4120 Page No. 2404 - 2410

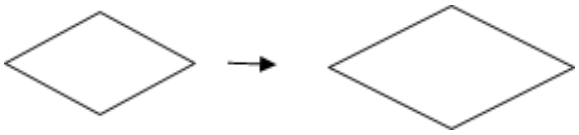


Fig. 4. Expanding shapes

Fig. 5. Filling, holes and gaps

To increase the size of objects in image dilation method is used which is very useful for object recognition. To increase the brightness of objects we can also use dilation method. This method works by taking the maximum value of neighborhood pixel (ii) Another Preprocessing operation "opening" is used in this operation will smooth various edges. It will also remove small protrusions from an image.

Horizontal Projection- It gives us information about the total number of strips in an image, number of rows in a strip, and strip byte count. By utilizing this information apply horizontal projection which gives total number of black or white pixels in a row. To extract each horizontal line horizontal projection [19] is applied.

Vertical Projection - Vertical projection is applied to separate connected characters. The troughs in vertical projection mostly correspond to the junction of two characters. After applying projections, we can identify and count total gaps in required region of interest (ROI).

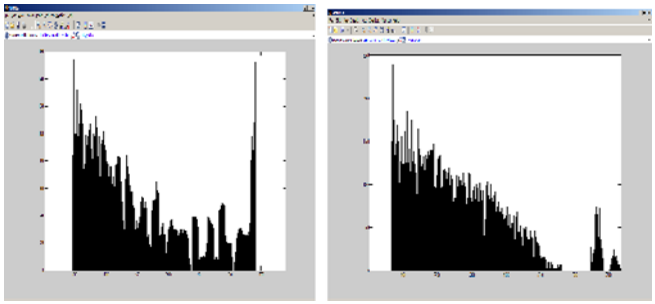


Fig. 6. Histogram of TOC-I &II

We can the calculate sum of pixels in each column of an image. This is used for the purpose of identification of number of black and white pixel in each column. This is called vertical projection. Vertical projection to separate connected characters [13] an application of OCR.

By applying both of these projections we can identify gap between the section/subsection and title and between title and page numbers. Once recognize the given page as a TOC page then taken for its main features recognition so that it can be further classified into aligned and not aligned pages. Steps to be followed for recognition and segregation of table of content page from other document pages include:

TOC Recognition – Recognition of TOC is done by finding out last connected component and calculating centeriod of each individual character.

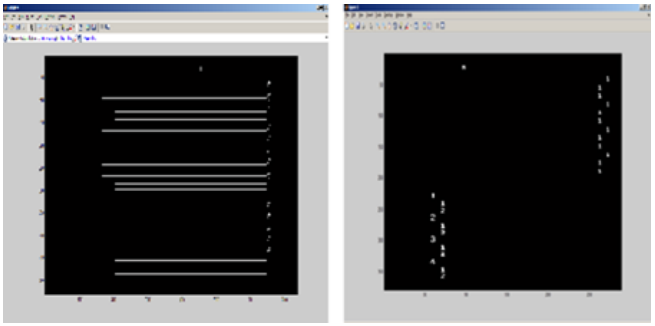


Fig. 8. Last connected components

IV. EXPERIMENTAL RESULTS

There is no fixed pattern of table of content pages. Each author or writer can make TOC for his book, magazine or report in his own way. Books and reports usually follow simple column wise division format (section and subsection headings, title and page number) while table of content in magazines and articles are in totally different format.

March-April 2020
ISSN: 0193-4120 Page No. 2404 - 2410

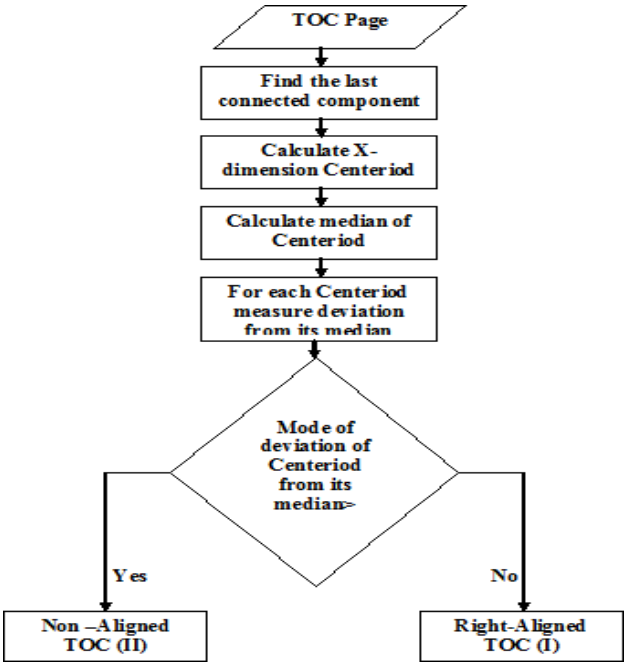


Fig. 7. Flowchart of TOC Recognition

1) Centeriod - Once last connected component is extracted then we have to calculate Centeriod of the last connected component of each line. For each centeriod we have to measure its deviation from its median. Then based on the mode of deviation compare the threshold for the required region of interest.

2) Centeriod Median- Once Centeriod of last connected component is calculated then we will calculate median of all centeriod and find out maximum and minimum value of centeriod value.

3) Distance Calculation - Centeriod of each line is now compared with maximum and minimum of centeriod. We will finalize one threshold value if distance between Centeriod is grater then that threshold value then we will consider it as a non-aligned page else we will consider it as a aligned page i.e. for each Centeriod measure the deviation from median of Centeriod. Then mode of deviation of Centeriod is analyzed and based on that we will consider it as an aligned or not aligned table of content page. This process segregate different types of table of content pages.

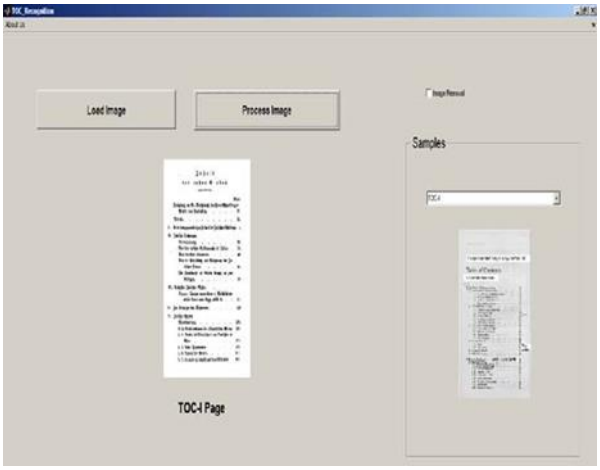


Fig. 9. Recognition of TOC-I(Aligned Pages)

Our technique for TOC identification is based on finding space on a page which is more than a normal page. More than 100 tables of content pages (TOC) and few normal pages of books, Journals and magazines present in a data set to be identified.

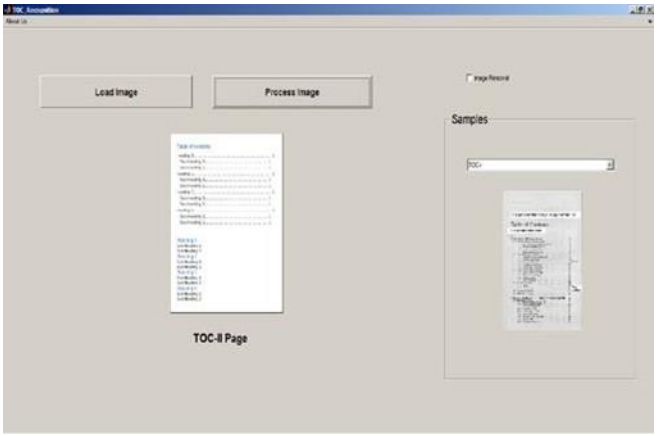


Fig. 10. Recognition of TOC-II (Nonaligned pages)

There are few limitations of proposed algorithm:

- This technique will not work for those pages in which we do not find sufficient spaces which let them differentiated from other normal pages of books.
- This technique will also not work for those pages where we do not find section or subsection numbers and page numbers only title is available this approach will consider it as a normal page.

Fig. 11. Table of content without space (b) With images in between

V. CONCLUSION

Identification and Segregation of different types of table of content pages from other document pages of books, journals and report is the main objective of this paper. It specifically works on documents which contains images in between. It works on differently oriented images and shapes as well. The main advantage of our algorithm is its simplicity. Predefined template set is not required images can be collected from any source. It does not follow any character identification technique [12]. It gives results based upon structure analysis. Due to simplicity and robustness it has also low computational cost.

REFERENCES

[1] S. Mandal, S. P. Chowdhury, A.K Das, "Automatic Detection and segmentation of table of contents and index pages from document images", In Proceedings of Seventh International Conference on Document Analysis and Recognition, 2003.

[2] S. Tsuruoka, C. Hirano, T. Yoshikawa and T. Shinogi "Image based structural analysis for a table of contents and conversion to XML documents", In Proceedings of the DLIA '01 Seattle, 2001.

[3] L. Gao, Z. Tang, X. Lin, and X. Tao, and Y. Chu "Analysis of Book Documents' Table of Content Based on Clustering", In Proceedings of 10th International Conference on Document Analysis and Recognition, 2009.

[4] Y. B. Kwon and J. Park, "Implementation of Content Analysis System for Recognition of Journals_Table of Contents", In Proceedings of Seventh International Conference on Document Analysis and Recognition, 2007.

[5] W. Tersteengen, Table Recognition by Reference Tables, 2002.

[6] M. Huang, W. Yu, and D. Zhu, "An Improved image segmentation algorithm based on Otsu method", In Proceedings of 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2012.

[7] X. Lin and Y. Xiong, "Detection and Analysis of Table of Contents Based on Content Association", International Journal of Document Analysis and Recognition (IJDAR), Vol. 8, pp. 132-143, 2006.

[8] H. M. Som, J. M. Zain, and A. J. Ghazali, "Application of thresholding Techniques for readability Improvement of Jawai Historical Manuscript Images", Advanced computing: An International Journal (ACIJ), Vol.2, 2011.

[9] H. J. Vala, and A. Baxi, "A review on Otsu Image Segmentation Algorithm", International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 2, pp. 387-389, 2013.

[10] M. Sezgin and B. Sankur, "Survey over image Thresholding Techniques and quantitative performance evaluation", Journal of Electronic Imaging, Vol. 13, 2004.

[11] O. I. Singh, T. Sinam, O. James, and T. R. Singh, "Low contrast and Mean Based Thresholding Technique in Image Binarization", International Journal Computer Applications, Vol. 51, pp.5-7, 2012.

[12] D. Deodhare, N. N. R. R. Suri, and R. Amit, "Preprocessing and Image Enhancement Algorithms for a Form Based Intelligent Character Recognition System", International Journal of Computer Science and Applications, Vol. 2, pp. 131-144, 2005.

[13] W. Bieniecki, S. Grabowski, and W. Rozenberg, "Image Processing for Improving OCR Accuracy", In Proceedings of International Conference on Perspective Technologies and Methods in MEMS Design, 2007.

[14] M. Elmore and M. Martonosi, "A Morphological Image Preprocessing Suite for OCR on Natural Scene Images", Georgia Institute of Technology, 2008.

[15] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 3rd ed., Pearson Education, 2009.

[16] L. Krasula, M. Klima, E. Rogard, and E. Jeanblanc, "MATLAB-Based Applications for Image Processing and Image Quality

Assessment", Radio Engineering, Vol. 21, pp. 154-161, 2012.

[17] M. Soni, A. Khare, and S. Jain, "A Survey of Digital Image Processing and Its Problems", International Journal of Scientific and Research Publications, Vol. 4, pp.1-6, 2014.

[18] N. Mahajan and K. Jaidka, "Various skew detection and correction Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, 2015.

[19] B. Jain and M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical Projection Profile Analysis", International Journal of Scientific and Research Publications, Vol. 4, pp. 1-4, 2014.

[20] S. Sonavane, A. khade, and V. B. Gaikwad, "Localization and Segmentation of Indian Car Number Plate System: A Projection Based Multistage", International Journal of Scientific and Engineering Research, Vol. 4, 2013.