International Conference on "Recent Trends in Technology and its Impact on Economy of India"
Guru Nanak College for Girls, Sri Mukstar Sahib, Punjab (India) (ICRTTIEI-17)
24th October 2017, www.conferenceworld.in ISBN: 978-93-86171-74-0

# A Survey on Newspaper Image Segmentation Techniques

**Rupinder Pal Kaur[1], Manish Kumar Jindal[2]**

[1]*Assistant Professor,Guru Nanak College for Girls, Sri Muktsar Sahib, Punjab, (India)*

[2]*Professor, Panjab University Regional Centre, Sri Muktsar Sahib, Punjab, (India)*

***ABSTRACT***

*Newspaper digitization is emerging field in OCR. Recognition of newspaper page is done by scanning the newspaper page and then segmenting scanned image into various regions. This paper presents a review on techniques of segmenting newspaper images into various blocks. We had studied various research papers written by various authors and keenly analyzed pros and cons of each technique. Section 1 covers brief introduction about need of newspaper digitization and entities in newspaper image. In section 2 various techniques proposed by different authors for segmenting newspaper images are discussed. In section 3, table to represent overall view of techniques is presented and discussion on techniques is done. Section 4 gives scope of future work. In last we conclude the paper*

.

## I. INTRODUCTION

Newspapers are a great source of information but the main obstacle is accessing particular newspaper and required information from piles of old newspaper. So there is dire need of newspaper digitization so that information can be retrieved from the desk of user. That is why newspaper digitization is major area of interest these days with many running projects of digitization worldwide. The National Digital Newspaper Program (NDNP) run by the Library of Congress and the National Endowment for the Humanities in United States is funding statewide newspaper digitization projects [50]. NDNP program make digitized availability of newspapers of United States from 1836-1922 [1]. To ensure that all Americans should get old newspapers from 1982, NEH run projects in various states with the help of NDNP [2]. Various links to projects held in states can be found from website of NEH. Digitization is storing scanned image of newspaper image on computer or on microfilms so that it can be accessible through internet. Digital image is only scanned image of newspaper article; no further text processing is possible on this scanned image. So, for text processing digitized image of newspaper text need to be converted to machine readable or we can say computer processable form. To convert newspaper article into machine readable form one option is to type article one by one and we can assume time consumption for this task, so the best option for conversion is OCR. Optical character recognition is conversion of scanned image of printed, handwritten or typewritten text into machine readable (encoded) text, so this converted text can be used for further machine processes. Through OCR manual typing errors can be removed and time can be saved.

To convert digital image of newspaper article into machine readable form two main steps are to be followed. First step is article segmentation into various regions or blocks like headline, sub headline, paragraphs, captions,

framed paragraphs etc. before feeding to OCR. Segmentation is the basic step for Character recognition. Segmented blocks shall be further segmented into recognizable unit. Second step is segmented block text recognition through OCR. Many hurdles are present in digitization of old newspapers like very complex layout of newspaper article , script mix with roman digits and most likely major problem is of poor paper and printing quality of newspapers. When old newspaper articles are scanned it causes many types of degradation in scanned image like distorted border of characters due to aging of paper, paper and printing quality, marks on paper due to time factor, folding of paper at spine of paper etc.

### 1.1 Newspaper Entities

A newspaper page image can contain various entities which need to be segmented for better recognition. In images various entities are shown which are very common in each newspaper and generally are candidate of segmentation.

1. Text region: text region (fig. a) is main body of any news that describe about the event occurred. The font of text varies newspaper to newspaper but generally body text font is less than title lines. Style and font remains same for whole body text.
2. Title: Title of newspaper article gives introduction about news. Font of title is large than body text. Many lines can be in title region. Sub titles can also be present in article. Titles run through columns, do not break in columns like body text. (fig. 2).
3. Horizontal and vertical lines: horizontal and vertical continuous or broken lines that separate articles in newspaper page. (fig.3)
4. Images: Digital photo present in article related to event described in article.(fig 4)
5. Drawings: graphics present in article like any map etc.


LALIT MODI EPISODE FALLOUT
Govt. draws up three
Idea is to close fault lines in BJP and c

**Fig. 1 Text region (highlighted with color blue) of article**


LALIT MODI EPISODE FALLOUT
Govt. draws up three
Idea is to close fault lines in BJP and c

**Fig.2 Title and sub title (highlighted in blue) of news article**


ay be used
nces
ambulances, he added.
Plea to amend rules
Meanwhile, the Transport Commissioner, also the State Road Safety Commissioner, has written to the Transport Department to amend rules to ensure that speed restrictions are not applicable to ambulances.
As per the Central Motor Vehicles Act, ambulances are exempted from speed restrictions when they carry the critically injured. The Transport Commissioner is seeking amendment to the Kerala
tickets likely for tourist spots
Shankar Bennur
MYSURU: Common entry tickets for prime tourist destinations here may be introduced, either before or after Dasara, which will end long queues outside the ticket counters.
Instead of buying tickets at every spot, visitors can purchase one ticket that will allow them entry to multiple destinations, such as Amba Vilas Palace, Mysuru zoo, Karanji Lake Nature Park, and Chamundi Hills. There is a plan to include Brindavan
Tribes den
collect fore
K.A. Shaji
PALAKKAD: The Forest Department's refusal to renew the annual licence of 27 cooperative societies under the Kerala State Development Corporation for SC/ST Communities (KSDCSSC) to collect minor forest produce (MFP) has put to difficulty a large number of tribal beneficiaries in areas such as Attappady, Idukki, and Wayanad.
The annual licence of the

**Fig. 3 horizontal lines separating news articles**


Four killed in clash at IIT Mandi
The victims were private security guards who allegedly provoked agitating workers
Staff Correspondent
SHIMLA: At least four people were killed and nine others were injured in a violent clash between the agitating workers and private security guards of a contractor in Kamand in Mandi district, where a new Indian Institute of Technology (IIT) campus is under construction. The workers on the site were agitating from past few days against the non-payment of wages and EPF by the construction company.
The contractor who had brought some other workers and armed security guards from outside was not heeding to the demands of old workers. In a face-off between the two the private security guards, who were well built armed bouncers from Punjab, opened up the fire and injured many workers including their leaders.
The police on the IIT construction site in Mandi where agitating workers and private security guards employed by a construction company clashed on Saturday. The workers were agitating for past few days in protest against non-payment of wages and EPF by the company. – PHOTO: PTI

**Fig. 4 Images in newspaper article**

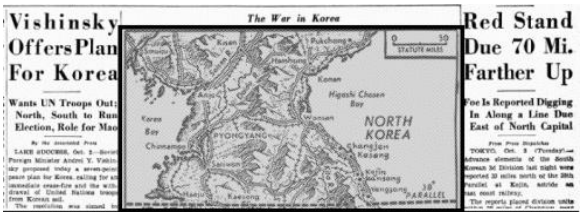


**Fig.5 graphics in article**

## II. SEGMENTATION METHODS PROPOSED BY VARIOUS AUTHORS

Segmentation of article into blocks or regions is necessary for better recognition through OCR. For segmentation of newspaper many techniques are proposed by various authors. Two main approaches used are bottom up and top down approach. Bottom up approach starts with segmentation of low level components and merge components into a region. For example start with line and merge to form a paragraph. Second is top down approach which starts with segmentation of higher level component like segment into paragraph and then into lines. Further techniques of segmentation of newspaper are proposed under these two basic techniques.

Lam et al.[3] Proposed technique for segmentation of newspaper article and recognition of text as well. This paper is among very first papers on research on newspaper segmentation. Lam et al. first segmented article image into various blocks using bottom up approach, segmented at character level. Segmented characters are merged using connected component analysis based on size and threshold of character component. Textual block and non textual block are filtered to get text region only. Textural analysis is then performed to classify

remaining text into different blocks .Segmented blocks were then feed to OCR for recognition. Segmentation technique failed to recognize text in framed paragraph, distorted characters, touching characters.

Page decomposition based on smearing and labeling of region is proposed by Gatos et al [4] on Greek newspaper "To Vima" to extract image components such as lines, images, drawings, text and title blocks. To extract and remove lines Hough transformation and morphological transformations are used. To extract images authors used technique of surrounding box height. But this technique also segment framed paragraphs as image which is incorrect segmentation. For further segmentation of text and title block, letter height and RLSA (run length smearing algorithm) with connected component analysis techniques are used with adaptive parameter. Using this technique text, images and lines are extracted with results up to 98% or above but title segmentation accuracy achieved only 89.10%.

Gatos et al. [5] experimented on two technique proposed in literature that were smearing and labeling, and image profiling but found inefficient in segmentation of newspaper. Gatos et al. proposed new technique based on horizontal and vertical image projection which provides image segmentation and identification of regions. Proposed technique consist of three parts that are first calculating horizontal and vertical image profiling using mask at each point depending upon letter size to get rectangular regions, second find out local minima of both horizontal and vertical profile to group together the segmented regions depending upon existing foreground pixels and last identify text regions by analyzing dominant frequencies through FFT (Fast Fourier Transform) of horizontal segmented regions.

Arabic newspaper decomposition is done by Hadjar et al. [6] by adapting technique proposed by Gatos et al.[4] with little modifications. Authors used connected component analysis and RLSA. Using connected component authors extracted thread (threads are separators between columns or between different entities), frames and images. For extraction of text line RLSA technique is used to extract lines horizontally and vertically, then applied connected components to extract blocks by threshold chosen manually according to text size. Using this technique segmentation results are sufficient when applied on Annahar and Alayat newspapers but low recognition rate are reported in title with diacritics (special symbols in Arabic) and threads and images with texture.

Beside segmentation of English newspapers work is performed on Chinese newspaper image. Jie Xi et al. [7] used bottom up approach to segment Chinese newspaper. According to author proposed techniques in literature do not segment Chinese image well because of two main reasons. First is Chinese character are not single component so RLSA technique is used twice to connect components of Chinese characters. Secondly articles are horizontally as well as vertically aligned in a newspaper image so to segment horizontal and vertical text connected components are generated. To generate connected component Jie et al. used inter block distance, between line distance, within line distance and a threshold that lie between-line distance and within-line distance to smoothen horizontal and vertical lines. This technique is best for Chinese newspaper although this can be applied on English newspaper too. This technique fail to segment headline if headline font size is little larger than body text.

ISBN: 978-93-86171-74-0

Techniques for segmenting textual and non textual regions are reported by Anderson et al.[8] .Authors first segment regions using X Y cut algorithm ( X Y cut algorithm work on horizontal and vertical projection profile to detect peaks. Peaks are thick black or white gaps at which cuts are placed [11]) then neural network is used to classify and merging of regions. In neural network multilayer perceptron is trained with back propagation. This paper proposed techniques for segmenting textual and non textual (with low accuracy rate of non textual region as compared to textual region) further segmentation of textual region is not proposed.

Only headline extraction is reported by [9] based on run length smearing algorithm.

A new method to segment newspaper article based on connected line was proposed by Mitchell et al. [10]. Bottom up approach is used for segmentation of page. First authors extracted patterns from image. These patterns are extracted using rectangular regions of 9 pixels (called rect in paper) containing at least one black pixel. The process of locating rects continues until a square of white pixels is found. Then connected rects are merged to form one of the nine listed regions of same characteristics. Region classification is based on some rules as defined by authors. These rules are based on pattern height, width, area, rect area etc. If horizontal and vertical lines which separate article are connected to text, then segmentation of component become necessary to separate text of two articles , separate algorithm is proposed by [10] to solve this problem.

Bottom up approach is basic approach used by Mitchell et al. [12] to segment document into components. This work is modified work of their previous work. Authors reported previous work using same technique of rects in [10] in which experiments was carried out on black and white images but in this approach no such conditions are implemented. Moreover limitation of [10] that text paragraphs containing single line were not segmented properly and columns with fewer gaps were merged, are tried to solve in this paper. Patterns are extracted to locate rectangular regions known as rect of loosely connected black pixels, once all patterns are extracted, patterns need to classify according to regions as text, photograph or something else. A set of rules are formed to identify each pattern based on size, shape, black pixel member, run length characteristics. This technique reports 76% of image segmentation. But disadvantage of this technique is complexity of technique. Secondly this technique does not work well with poor quality image and segmentation of title near to text.

Yuan et al. [13] proposed method for segmentation of newspaper article image based on edge detection and merging method. Yuan et al. used gray scale image for the purpose of segmentation. The method is based on continuous shape of characters and a fixed distance between lines. Canny edge detection method is used to locate edges in horizontal direction of lines. The small horizontal edges are merged to form a longer text line. While merging into lines extra noise lines are removed by using thin line coding (TLC). A little skew while merging to form a line is managed by Eigenvectors. After detection of text lines, region merging is done to form a block by pairing straight lines from upper and lower edge. To eliminate some noise, connected component analysis was performed. The proposed method gives accuracy of results if lines are straight. This method fails if text lines are not aligned and method fails if text block is not well separated from background.

Boirgiu et al. [14] proposed method for classification of text based on geometrical characteristics. This technique would also be applied on newspapers text. Authors experimented on roman script newspapers and gave idea of classifying text consisting homogeneous structural properties or measurements like texture, font size, font boldness, italic text and line spacing etc. Authors first extract entities by using run length connected

pixels and then applied filters on extracted entities to remove insignificant entities. Then applied algorithm to extract text measurements and based on text measurements and texture analysis homogeneous text is clustered which will separate individual blocks of newspaper article. Limitation of this algorithm is if a block contains text with different text features like caption which could contain capital as well as lower case letters; it will separate it into different blocks.

Fixed point model is introduced by Bansal et al. [15] to automatically segment English newspaper article by labeling different regions of article. To segment article author used leptonica software designed by Bloomberg to segment into text and graphics. Labeling is used to identify blocks as headline, sub headline, text blocks and caption. Labeling of each node (block) is based on features of node like appearance and contextual features. Appearance features try to associate each block to label using characteristics of that block and contextual feature try to label a block using information from neighboring blocks. Results of block segmentation are compared using SVM and KLR, around 96% and 97% accuracy is found in segmentation of heading and text blocks, but only 82.4 % accuracy is found in segmentation of sub heading because of mixing with heading and 83.74% accuracy is reported in segmentation of caption because of long caption confused with text blocks.

## III. DISCUSSION ON TECHNIQUES

Various techniques are proposed by researchers. In this section we are presenting a table which will provide overall view of techniques. Which techniques are used in paper and what are limitations resulted after implementing these techniques, are discussed in Table.(table 1.)

**Table 1: Representing overall view of segmenting techniques**

| Reference No. | Author | Input Data Type | Techniques used | Limitations if any |
|---|---|---|---|---|
| [3] | Lam et al. | English Newspaper | Connected Component Analysis and Texture Analysis | Do not segment framed paragraphs |
| [4] | Gatos et al. | "To Vima" Greek Newspaper | Surrounding box height and RLSA | Title segmentation accuracy low |
| [12] | Mitchell et al. | English newspaper | Connected line(formed rects of 9 pixels) | Text columns are segmented well if gap is low in columns |
| [7] | Jie Xi et al. | Chinese Newspaper | RLSA and Inter block distance | Fail to segment headline if font is little larger than body text size. |

| | | | | |
|---|---|---|---|---|
| [6] | Hadjar et al. | Arabic Newspaper | Connected line component and RLSA | Fail to segment title with special symbol |
| [5] | Gatos. Et al. | Greek Newspaper | Image projection profiles and FFT | |
| [8] | Anderson et al. | English Newspaper | X Y cut algorithm | Proposed only textual and non textual segmentation |
| [13] | Yaun et al. | English Newspaper | Edge Detection and merging | Text blocks are not segmented if lines are not straight |
| [14] | Boiangui et al. | Roman newspaper | Used geometrical features for segmentation | Divide single block into various if block contains variable text size. |
| [10] | Mitchell et al. | English newspaper | Connected Lines | Do not work well with poor quality image and do not segment title near to text. |
| [15] | Bansal et al | Indian English newspaper | Fixed point model | Sub headings are not segmented well. |

### 3.1 Discussion

Segmentation is necessary before feeding any newspaper page image to OCR for proper recognition of text. To merge homogenous regions connected component analysis technique is used in almost papers. RLSA technique is used by many authors to segment image into homogeneous regions . Reviewing all these papers common resulted limitations are that sub titles are not extracted accurately. Mitchll et al.[10] used techniques of dividing image into rectangular rects of 9 pixels to extract features and merge rects to form homogenous regions. These techniques do not work well on degraded documents. Gatos had done a lot of work to segment article and finally proposed image projection technique to segment images. X Y cut algorithm works well to segment textual and non textual regions. Canny edge detection algorithm for segmentation of newspaper images can be used but it could not handle text blocks if text is not well separated from background. Methods are also proposed based on geometrical features but this technique failed to segment a block if it contains font of two different sizes for example text in caption.

ISBN: 978-93-86171-74-0

## IV. SCOPE OF FUTURE WORK:

After reviewing the papers we found that most of the techniques failed to segment sub headlines from body text if headlines are very close to text. Second limitation is segmentation of caption which is present below image in article. Captions are segmented as of body text.. New Algorithms are required to correctly segment framed paragraph present in articles as framed paragraphs are incorrectly segmented as images.

## V. CONCLUSION

This paper presented overall view of segmenting newspaper image techniques used by different researchers .We conclude that in most of papers combination of techniques are used for segmentation which give better results. Still many improvements in techniques are required to overcome above discussed limitations. Most of these techniques are experimented on different English, Greek, Chinese and Arabic newspapers. Indian language newspapers segmentation is on very initial stage with lot of scope.

## REFERENCES

[1.] http://www.loc.gov/ndnp/

[2.] http://www.neh.gov/us-newspaper-program

[3.] Lam, Stephen W., Dacheng Wang, and Sargur N. Srihari. "Reading newspaper text." Pattern Recognition, Proceedings of 10th International Conference on document analysis and recognition, Vol. 1. IEEE, pp. 703-705, 1990.

[4.] Gatos, B ."Integrated Algorithms for Newspaper Page Decomposition and Article Tracking." Proceedings of the Fifth International Conference on Document Analysis and Recognition. IEEE Computer Society, pp. 559-562, 1999.

[5.] Gatos, Basilios,. "A new method for segmenting newspaper articles." Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, pp. 695-696, 1998.

[6.] Hadjar, Karim, and Rolf Ingold. "Arabic newspaper page segmentation." 12th International Conference on Document Analysis and Recognition. IEEE Computer Society, Vol. 2, pp. 1186-1189, 2003.

[7.] Xi, Jie, Jianming Hu, and Lide Wu. "Page segmentation of Chinese newspapers." Pattern recognition, Vol. 35 (12), pp. 2695-2704, 2002.

[8.] Andersen, Tim, and Wei Zhang. "Features for neural net based region identification of newspaper documents." Proceedings. Seventh International Conference on Document Analysis and Recognition, IEEE, pp.403-407, 2003.

[9.] Liu, Qing Hong, and Chew Lim Tan. "Newspaper headlines extraction from microfilm images." International journal on Document Analysis and recognition, Vol. 6, pp. 201-210, 2004.

[10.] Mitchell, Phillip E., and Hong Yan. "Newspaper document analysis featuring connected line segmentation." Proceedings of the Pan-Sydney area workshop on Visual information processing, Australian Computer Society, Vol. 11, pp. 1181-1185, 2001.

ISBN: 978-93-86171-74-0

[11.] Niyogi, Debashish, and Sargur N. Srihari. "An integrated approach to document decomposition and structural analysis." International Journal of Imaging Systems and Technology, Vol. 7(4), pp. 330-342, 1996.

[12.] Mitchell, Phillip E., and Hong Yan. "Newspaper layout analysis incorporating connected component separation." Image and Vision Computing, Vol. 22 (4), pp. 307-317, 2004.

[13.] Yuan, Qing, and Chew Lim Tan. "Page segmentation and text extraction from gray-scale images in microfilm format." Photonics West 2001-Electronic Imaging. International Society for Optics and Photonics, vol. 4(2), pp. 323-332, 2000.

[14.] Boiangiu, Costin-Anton, et al. "Automatic text clustering and classification based on font geometrical characteristics." Proceedings of 9th WSEAS International Conference on Automation and Information, pp. 468-473, 2008.

[15.] Bansal, Anukriti, "Newspaper article extraction using hierarchical fixed point model." Document Analysis Systems (DAS), 11th IAPR International Workshop on IEEE, pp. 257-261, 2014.