**“Newspaper Article Segmentation”**

**Contents**

1. **Introduction**
2. **Literature Survey**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Authors | Paper Title | Methodology | Datasets | Results | Limitations/ Research Gap |
| Anukriti Bansal, Santanu Chaudhury, Sumantra Dutta Roy and J.B. Srivastava | Newspaper Article Extraction Using Hierarchical Fixed-Point Model | Machine Learning (SVM and KLR) | Customised Dataset |  |  |
| B. Gatos, S.L Mantzaris, K.V Chandrinos, A. Tsigris, S.J Perantonis | Integrated Algorithms for Newspaper Page Decomposition and Article Tracking | Rule-Based Approach | Dataset containing 100 images of “TO VIMA” newspapers |  |  |
| Minghao Li1, Yiheng Xu2, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, Ming Zhou | DocBank: A Benchmark Dataset for Document Layout Analysis | Rule-Based Approach | DocBank Dataset containing 500k Images |  |  |
|  |  |  |  |  |  |

1. **Objectives of the work**

To segment and identify the blocks such as paragraphs, titles, etc within the newspaper which helps in extracting features for the article segmentation. Block identification is achieved using certain methods such as thresholding, morphological dilations, RLSA, etc. Each block within an article contains main attributes such as title, contents, and images and these can be segmented using the identified blocks as one of the features to the machine learning model.

To create ground truth dataset for the newspapers by mapping and annotating the original image for articles, advertisements, headers, footers, images, etc and generating the coordinate points of each annotated groups. Along with these coordinate points other attribute information such as image resolution, annotation rules, etc are stored in to XML (Extensible Markup Language) files for each image found within the dataset.

1. **Methodology**

For block identification various methodologies are implemented where initially binary thresholding of images is carried out and then run-length smoothing algorithms are applied in order to extract those blocks as separate elements.

* 1. **Binary Thresholding**

Thresholding of an image is said to be segmentation of image pixels into two

different regions, one as foreground and other as background. These regions are segmented based on the threshold value set to the intensity of the pixels. Above the threshold is said to foreground and rest is said to be background which helps further methods to identify the region of interests to be analysed/processed. Thresholding can be easily applied only if the image is a grayscale. In our process most of the newspapers have white background and hence the threshold value will be selected close to the maximum pixel intensity as 255 for any 8-bit image.

* 1. **Morphological Dilation**

Generally, morphological operations are the operations carried out in the pixel intensities which changes the structure of the areas of interest within an image by altering the boundary pixels. After thresholding certain regions within the image will not be clear as no of pixels will be found missed. Those boundaries of regions and missing pixels can be expanded or distorted by applying a structural element of a particular square matrix which is also a binary image said to be kernels. For dilation this structural element is superimposed in the original image and adds the pixels to both the inner and outer boundaries of regions. Hence certain pixel regions found around the texts in the newspapers are expanded such that whole paragraphs can be segmented as blocks.

* 1. **Run Length Smoothing Algorithm**

The run length smoothing algorithm is a procedure to identify the regions/blocks within an image consisting of texts, graphics, etc. These blocks can be segmented as words, paragraphs, titles, images, etc by using a sequence-based transformation along the n number of row and column pixels. This algorithm is applied in an image with a threshold value representing no of pixels in a sequence considered within the horizontal and vertical pixel arrangements. Input sequence is row-wise pixels found in the image and no of background pixel values (0’s) along the sequence are calculated. Only if the no of 0’s in the sequence consecutively is less than the threshold value, then those pixels are altered to foreground pixel values of (1). When condition fails, the pixels remain the same.

Both horizontally and vertically, the sequences are considered and applied with the algorithm to obtain as new binary images. From these two images, a logical OR operation is carried out followed by a logical NOT to eliminate the white spaces and retain them as blocks.

1. Work carried out so far
2. Results and Discussions
3. Further work to be carried out
4. Conclusion
5. References