paper

by Plagiarism Checker

Submission date: 14-Nov-2021 09:37AM (UTC-0500)
Submission ID: 1702023874
File name: SHILPA_PAPER-ID_29_without_references.doc (1.27M)
Word count: 3117
Character count: 17346

A Novel Approach for Newspaper Block Segmentation using Run-Length Smoothing Algorithm

Shridevi SOMA[a,1] and Shilpa[b]
[a]*Computer Science and Engineering Department, PDACE,Klb,Krnataka,India*
[b]*Computer Science and Engineering Department, SUK, Klb,Karnataka,India*
[a]*shridevisoma@gmail.com* , [b]*s123shilpa@gmail.com*

Abstract. Region findings and analysis plays an important role in document understanding. Due to the existence of complex layouts, document understanding is a very challenging task for researchers. Layouts in newspapers are derived for the articles where article consists of multiple blocks. These various blocks must be segmented and identified within the whole newspaper which helps further in article segmentation. This paper proposes a novel method to identify and segment blocks found within the newspaper irrespective of its layout using simple image processing operations such as morphological dilation, run-length smoothing algorithm and rule based algorithm. These methods have been tested on dataset consisting of digital enewspaper images of recent years from the TOI and Financial Express Newspapers with different layouts and complexities. The experimental results exhibits our method proposed outperformed region findings with the precision 0.83, recall 0.72 and F1 score 0.76. Effectively these blocks will be used as features and evaluation measure for various document analysis.

Keywords. Complex Layouts ,Document understanding, Dilation ,Layouts,Region findings ,Segmentation

1. Introduction

Document Digitisation is the process of taking a physical document or document representation and transferring it into a faithfull digital representation. The ideal goal is to preserve all the information contained within the physical document as physical documents deteriorate overtime [1,2,3]. It also achieves the goal of improved access, as digital documents can be quickly retrieved from archives and returned to a user, without them having to be in the same physical location as the document. Digitisation is necessary to allow sophisticated computer processing of documents [4,5].

Document layout analysis plays a vital role in processing the digitized documents in an efficient way and acts as a pre-processing step. The task of analysis of layout involves pointing out and designating regions of interest on a given document[6,7,8,9,10]. Due to the complex layouts found in document image, document understanding and analysis becomes difficult and in past many works has been initiated on this, but due to complexity of document it still remains as a challenging task and need to be addressed. This paper proposes a simple method to identify and segment blocks found within the newspaper irrespective of its layout using simple image

processing operations. These methods have been experimented on digital enewspaper images of recent years with different layouts and complexities.

Contribution of the work [11,12] are as fallows considering the complex documents with irregular layouts, detecting the major whitespaces, an attempt to segregate where an image is surrounded by text and vice versa, dealing with varying fonts, headers, captions, finally segmentation findings blocks using run length smoothing algorithm and rule based algorithm the proposed model. As a result, our algorithm segments various blocks including texts, images and graphics in the newspaper and extract them as separate elements. Effectively these blocks will be used as features and evaluation measure for article segmentation algorithm.

The rest part of effort is as structured in the next section related work on block segmentation is presented. Section III explains proposed methodology and gives details about each module of proposed approach. Section IV describes experimental results & discussions respectively. Section V provides the inferences of work and ensuing orientations

2. Related Work

Looking back to complexities involved in understanding the structure of a newspaper document Hui-Yin Wu et al.(2019) [3] proposed an approach towards understanding the structure and design of newspaper image. The approach fallows computer vision techniques for segmentation via curvilinear structure detection and CNN techniques for classification via fastai library. Nasid Habib Barna et al. (2018) [6] presents a language independent segmentation system based on morphological operations which takes the input as heterogeneous document and produces output as homogenous components. The system segments homogenous components as title, image, and tables. Authors proposed two more methods RIFR and TETC and have achieved accuracy of 93%. As a replacement to conventional CNN Sai Chandra Kosaraju et al. (2019) [7] introduces a new model “texture-based CNN” for layout findings and analysis. The model uses dilated CN layers which takes a tile image and generates a class label thereby outperforming the available conventional CNN techniques in classifying documents. Hybrid methods based on morphological operations, contour classification, connected components have been proposed by Nikos Vasilopoulos, S.W. Alarcon Arenas et al. ((2017)(2018)) [[11],[12]] their findings include titles, text and non-text regions respectively. Focusing towards developing document layout evaluation system and page segmentation methods Thomas Strecker et al. (2009) [13] developed MyNews ground truth dataset generation system for newspaper layout analysis. The system takes as input news corpus and generates digital newspaper and XML file as output. A key contribution by Anukriti Bansal et al. (2014) [14] in automatically learning the rules to identify the logical units of a given document image outperforms the available systems that rely on learning the rules to detect logical units. The proposed model captures the conceptual information for labelling of block predictions present in news articles and therefore the model attained promising results. Tuan Anh Tran et al. (2018) [15] proposed hybrid method. Connected component analysis and white space analysis are used for classifying textual and non-textual regions. Textual regions during segmentation are retrieved by white space analysis and mathematical morphology. Non-text regions like graphs, tables, charts and many more are retrieved by ML process. Empirical computations on UWIII dataset found to be

promising and faster. Processing time reduced by 1/3 when compared to multilevel algorithms..

. Dario Augusto Borges Oliveira et al. (2017) [16] proposed reduction in dimensionality based classification approach. The proposed architecture takes as an input two 1D arrays of horizontal and vertical projections of image tile and classifies them into text, images, tables then final class labelling assignment done. Empirical results shows that the proposed architecture outperforms state-of-art techniques by low data usage and increasing the processing time. Annus Zulfiqar et al. (2019) [17] proposed a deep learning technique for logical layout analysis. The proposed method uses two layered RNN where text is passed as an input to the model where it tries to find and label zones. Empirical computations shows that the model achieves accuracy of 95.38% with random split and 96.21% for unseen layouts from this it is clear that the model works well for even the unseen layouts when compared to random split. A. Almutairi et al. (2019) [18] proposed a deep learning technique to generate a language-agnostic model to semantically segment newspaper images .S. Ramesh et al. (2021) [22] in their work have proposed model to segment the given newspaper document image into different categories by fallowing the concept of Mask R-CNN based on image augmentation and transfer learning method that generates separate masks and their associated labels. S. Biswas et al. (2021) [23] presented a model depicting the techniques that got evolved in field of segmentation particularly in they have highlighted about importance of instance segmentation inspired by MaskR-CNN techniques. S. Pletschacher et al. (2010) [25] describes XML based PAGE framework in evaluating datasets in any infrastructure. Due to its promising and efficient evaluation results the framework is now extensively used in many competitions series ICDAR, PRImA.

From the above literature survey, it is found that the research work on document layout analysis has been initiated long ago due to the complexities of document it still remains as a challenging task and need to be addressed. It is noted that the concept of rule based method, non rule based method, hybrid methods have been used for document understanding and analysis. Recent survey says that works based on deep learning and neural network have been initiated and neural networks seems to be advantageous[22,23,24,25], more competitive and needs to address.

3. Proposed Methodology

The proposed methodology is implemented in two phases pre-processing and blocks identifications. The aim of pre-processing is improve the likeness of image to analyse it in a better way. By pre-processing we can reduce undesired disturbances and enhance some features which are required for the applications we are working for.

3.1. Need for Preprocessing

(i) Resizing of collected dataset images must be done before applying segmentation (ii) All images should be made feature extractable with same kind of thresholding (iii) Detecting major whitespaces in the newspaper to understand the unique layouts (iv) Finding the grid structure (rows and columns) used within the article layouts (v) De-noising the images for better feature extraction especially in old newspapers (vi) Segregating images along with respective articles within multiple grid sharing (vii)

Extracted elements and components are chosen according to their close influences the feature extraction.

3.2. Block Identifications and Dataset

The second phase of the proposed work is carried out using binary thresholding ,RLSA and Rule based algorithm. The Dataset used for implementing the proposed approach is as fallows- Digital newspaper images of recent years from The TOI and Financial Express enewspapers from https://archive.org/details/TOIDELFEB18, https://archive.org/details/TOIDELJAN18, https://dailyepaper.in/finanmcial-express-newspaper with different layouts complexities, images of size128*128 and in png format.
Figure. 1. depicts the flow of the methodology proposed. As an initial step the data are read in form of Numpy array where RGB values for images are initialize

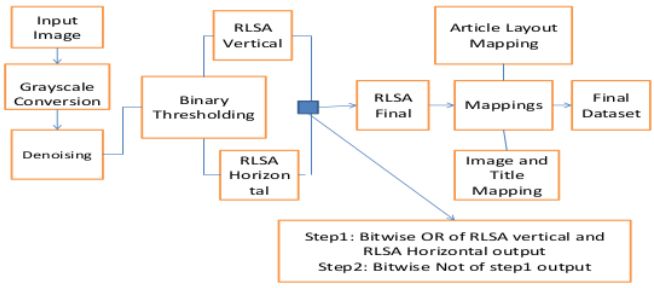

Figure 1. Flow of the methodology proposed

A. Denoising

Denoising a method used to remove noise fastNIMeansDenoising is considered for smoothing. It looks for a particular change in the grayscale image with a parameter deciding filter strength, a templatewindow size to calculate weights and a search window size to calculate weighted average for given picture element

B. *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 entry value set to intensity of the pixels. Above the entry value is said as 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 the 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.

C. 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 number 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.(Figure 2)

1	1	1
1	1	1
1	1	1

Set of coordinate points =

{ (-1, -1), (0, -1), (1, -1),
(-1, 0), (0, 0), (1, 0),
(-1, 1), (0, 1), (1, 1) }

Figure 2. A 3x3 kernel (structural element) used in dilation

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.Next phase of our work focuses on Block findings. The most important and required phase in document understanding/classification. Methods used for finding each block found in the whole image are a) OCR b) RLSA- run length smoothing algorithm and c) Rule based algorithm.

A.OCR

For performing OCR (converts image to string, image to image) a commonly available pyTesseractocr engine a framework is used. It is majorly useful in text analysis alone but in proposed case for layout extraction it is noticed that it will identify text and make bounding boxes without using any contours method, thresholding method, the overall detected bounding boxes can be merged to say it as particular paragraphs for region detection. But here image is also considered as a text with an error that is observed and failed to detect properly (Figure. 3a & 3b).

(a) (b)

Figure 3. Applying OCR method results (a) original document image (b) OCR_result

A. *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 number of pixels in a sequence considered within the horizontal and vertical pixel arrangements. Input sequence is row-wise pixels found in the image and number of background pixel values (0's) along the sequence are calculated. Only if then 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 (Figure. 4e & 4f) 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 (Figure. 4g) and retain them as blocks (Figure. 4h). The major and important contribution of the RLSA algorithm chosen helps us to identify blocks in a better way when compared to normal dilation and thresholding methods.

B. *Rule Based Algorithm*

The Rule Based Algorithm has been derived to retrieve intelligence properly out of the RLSA output. Hence certain rules involving the above-mentioned operations have been involved and are as listed below.

Algorithm: Rule based Algorithm

Input: RLSA output- Blocks
Output: Filtered contours representing blocks
Start

Step 1: Contour and their bounding boxes extraction
Step 2: Compute the average width and height of all bounding boxes
Step 3: Filter contours with height >1.6 times the average height
Step 4: Again filter those heights<2.1 times the average height
Step 5: Create masks
Step 6: Apply morphological dilation with 2*5 kernel matrix to Step 5
Step 7: Remove title regions from Step 6
Step 8: Apply morphological dilation with 1*2 kernel matrix
Step 9: Perform contour extraction to the output of Step 8
Step 10: Generate masks for blocks identified
Step 11: To the Bounding boxes from Step 6 and Step 10 filter the masks that are of negligible width and height where criteria is either 0.1 times below the height or 0.1 times below the weight
Step 12: Extract the filtered contours
Stop

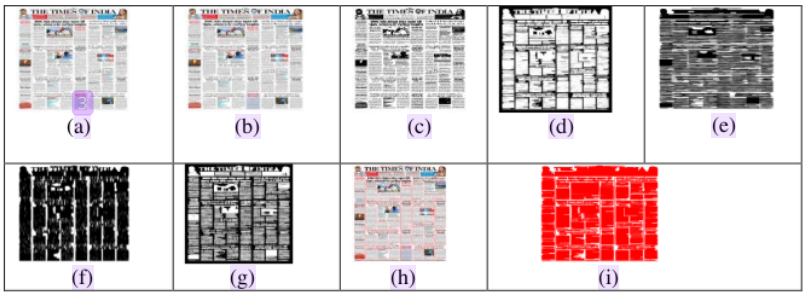

Figure 4. Results of the proposed methodology (a) original image (b) Denoised image (c) Threshold image (d) Dilated image (e) RLSA_h (f) RLSA_v (g) RLSA_f (h) Final result (i) Final result1

4. Experimental Results & Discussions

The proposed method for finding the blocks has been implemented in OpenCV language on a laptop with windows OS and other configurations. Experimented on a dataset consisting of digital enewspaper images of recent years from The TOI and Financial express enewspapers. It is observed that the proposed algorithm segments various blocks as shown in Figure. 4. Later these blocks will be effectively used as features and evaluation measures for article segmentation algorithm. In comparision with existing works as shown in Table 2 it is noted that author S.Naik has achieved an overall performance of 0.61 in segmenting the English newspaper images by fallowing otsu method, thickness of white pixels varied with block of black pixels, black runs with maximum widths methodologies. Author N.Vasilopous by following the hybrid technique for complex layout analysis of high resolution scanned newspaper images has achieved score of 0.76. Furthermore, from the results acquired, to determine how accurate the algorithm have segmented and identifies the blocks within the original data, certain calculations have been performed from blocks identified. Precision and Recall calculations are the major factors to determine accuracy scores of our algorithm and F1 score is calculated from Precision and Recall calculations. In the Eq. (1) and (2) TP represents True Positives, FP represents False Positives and FN represents False Negatives.

$$Precision = \frac{TP}{TP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FP} \quad (2)$$

$$F1\,Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (3)$$

For testing images, each calculation is performed and scores has been determined and for overall results, the averaged values of those calculations are used. Accuracy of our

algorithm can be determined from the F1 score achieved between the range 0 to range 1. It is observed that when compared to performance of S.Naik in segmenting the unstructured newspaper document N.Vasilopolous with the hybrid approach attained the better performance rate but the proposed approach which combines both the methods have achieved the good performance rate on a large number of English newspaper images .

Table 1. Accuracy results for segmenting the blocks

Block Segmentation	Precision	Recall	F1 Score
	0.83	0.72	0.76

Figure 5. Shows the results obtained by run length smoothing algorithm and Generated Ground Truth and Table 1clearly depicts the results of block findins with F1 score 0.76. By comparing both the results (Figure. 5a & 5b) it is found that our proposed algorithm has performed a decent score to detect the regions as blocks. Figure. 6. Shows the graph of performance measure for block segmentations attained by our proposed work where in the Figure. 6. Images indicate the newspaper images of 128*128 size drawn from TOI enewspapers and Financial express enewspapers.

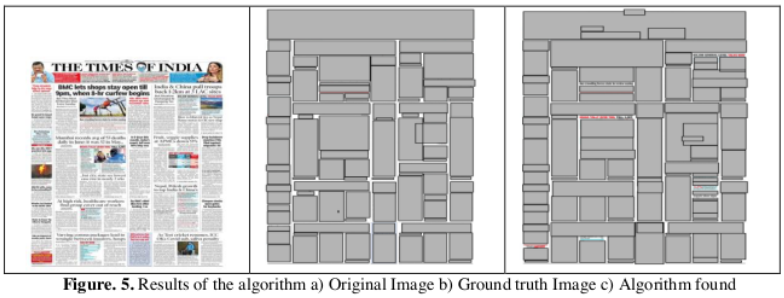

Figure. 5. Results of the algorithm a) Original Image b) Ground truth Image c) Algorithm found

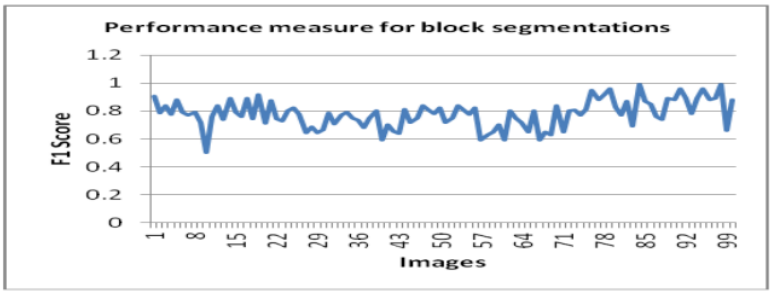

Figure. 6. Performance Measure for Block Segmentations. Images on the X-axis indicate images retrieved from Times Of India enewspapers as mentioned in Dataset Section. The ImageID's are TOI_0001,TOI_0004,TOI_0006-TOI_0016,TOI_0018-TOI_0046,TOI_0051-TOI_00074. F1 Score on the Y-axis indicate F1 Score of all images considered.

Table 2. Comparisions of the proposed method with the available methods

Authors	Precision	Recall	F1 Score
Proposed Method	**0.83**	**0.72**	**0.76**
N.Vasilopous et al. [11]	0.75	0.81	0.76
S.Naik et al. [26]	0.63	0.74	0.61

5. Conclusion

Block findings and analysis forms a major and first step towards building an efficient and complete model for document classification and analysis systems. Block findings is attained using certain methods such as thresholding, morphological dilations, OCR, RLSA, Rule based approach . 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. Using RLSA we found outer region and individual regions as block findings. Experimental results shows that compared to existing methods RLSA helps us to identify blocks in a very better way with precision of 0.83, recall of 0.72 and F1 score of 0.76. As limitations it is observed that the present work fails to detect too complex blocks, could not detect small font sized captions Future scope of the block segmentation is to label those blocks within each articles to study the layouts of the newspaper, to extend the dataset size and improvise the performance rate.

paper

ORIGINALITY REPORT

5%	4%	2%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ses.library.usyd.edu.au Internet Source	2%
2	dokumen.pub Internet Source	1%
3	eprints.lancs.ac.uk Internet Source	<1%
4	Anukriti Bansal, Santanu Chaudhury, Sumantra Dutta Roy, J.B. Srivastava. "Newspaper Article Extraction Using Hierarchical Fixed Point Model", 2014 11th IAPR International Workshop on Document Analysis Systems, 2014 Publication	<1%
5	Muneeswaran, K.. "Texture image segmentation using combined features from spatial and spectral distribution", Pattern Recognition Letters, 200605 Publication	<1%
6	www.ncbi.nlm.nih.gov Internet Source	<1%

7	tel.archives-ouvertes.fr Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		