IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, Mar 2019 E-ISSN: 2347-2693

Classification of Text and Images from PDF Using Graph Based Technique

D. Selvanayagi

Dept. of Computer Science and Commerce (CA), Vellalar College for Women, Erode -, Tamil Nadu

Corresponding author: selvasubhika@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.11411146 | Available online at: www.ijcseonline.org

Accepted: 13/Mar/2019, Published: 31/Mar/2019

***Abstract*—**Today's e-book plays an important role in all fields to learn new things through personal computer, laptop or mobile phones. There are various formats available for an e-book. The extensively used format is PDF because it retains the original format of the document. Segmentation is for reusing the content but in existing system the documents are segmented as the text content only. It doesn't consider the non-text elements such as graphs, tables, and images. In this research layout analysis is performed by extracting both text objects and non-text objects from the PDF document and segmenting the objects separately using Support Vector Machine (SVM) classifiers. Finally we get the output as text objects and non-text objects separately. This method utilizes both bottom up approach for text line extraction and top down approach to divide graph tree created by Kruskal's algorithm into sub graph which use Euclidean distance between adjacent vertices. Both text and non-text objects are classified using SVM technique. For each segmented text and non-text different dimensional features are extracted for labeling purpose. Several E-book PDF documents are tested and some sample input and output PDF documents are shown in the experimental result.

***Keywords* —** E-book, PDF, Kruskal's algorithm, Euclidean distance, SVM.

I. INTRODUCTION

An e-book is an electronic model where we get a traditional print book from either a personal computer or by using an e-book reader. The e-book is available in various formats like MOBI, AZW, AZW1, AZW4, EPUB, and PDF. The most widely used e-book format is PDF because while transferring PDF documents it maintains the original formatting and security; no one can change the content of the document. The PDF document may contain text objects and image objects. Text objects contain only the text data. Image objects include graphs, tables, lists, and images. Document segmentation plays an important role in e-book which is used to reuse the content of the document. It is a method of sub dividing the document regions as text regions and image regions and it leads to layout analysis.

The document can be segmented as text segmentation and image segmentation. In the existing work only the text contents of a PDF document can be segmented. But it is more important of image segmentation along with the text segmentation. Text segmentation is a precursor to text retrieval, automatic summarization, information retrieval, language modeling and natural language processing. In written texts, text segmentation is the process of identifying the boundaries between words, phrases, or some other linguistic meaningful units, such as sentences and topics. The term separated from such processing is useful to help humans reading texts, and are mainly used to assist computers to do some artificial processes as fundamental units. Text line extraction is a preprocessing step for handwriting recognition and document structure extraction, and image segmentation is mid-level a processing technique. The main reason of the segmentation process is to get more information in the region of interest in an image.

Most of the PDF documents contain both text and image objects so in this research we consider both the text and image objects, for segmentation in the PDF documents. In the proposed research work the segmentation of both text and image components in e-book PDF format is considered. This overcomes the restriction of segmentation in tables, images, graphs, etc., in the PDF document. In this system both text layer and image layer are taken into consideration for segmentation. Each layer segments its data independently. Finally the results of both text and image layers are merged together for final segmentation. Text segmentation and image segmentation are used for reusable purpose.

II. REVIEW OF LITERATURE

Neha Gupta et al introduced a text extraction concept which is based on Image Segmentation. The text involved in these images includes critical and useful information. Text

extraction in images has been used in a large variety of applications such as vehicle license plate detection, document retrieving, mobile robot navigation, and object identification [1]. In this system, we retrieve text information from complex input images by using Discrete Wavelet Transform (DWT). But a preprocessing step is required for color to extract text edges in the color image. The edge map is formed using resultant edges. Morphological operations are applied to improve the performance on the processed edge map and then thresholding is applied in the image.

Chandranath Adak et al introduced a new method for Unsupervised Text Extraction from G-Maps. Text extraction [3] is a method to extract the textual section from a non-textual background. Due to an unsupervised approach there is no necessity of any preceding knowledge or training set about the textual and non-textual parts. For image segmentation and to detect the edge Fuzzy C Means Clustering Technique and Prewitt Method [5] are used respectively. The constraint of this system is that it is not fully automatic because of thresholding and selection of better result is dependent on human eye.

Q. Yuan et al introduced a new text extraction technique which is based on Edge Information. The designed scheme presents a well-designed approach [6] that uses area statistics to take out textual blocks from grey scale record pictures. The main objective of this scheme is to find out textual regions on heavy noise- infected newspaper photos and split them from graphical regions. The algorithm traces the function points in unique entities and then groups the ones with facet points of textual areas. From this method we can obtain accurate web page decomposition with green computation and reduced reminiscence size by copying with line segments.

Thai V. Hoang et al introduced a new text extraction method which is based on Sparse Representation. Input document image includes both text and graphics which is processed to produce two output images, one returns with text and the other returns with graphics [7]. Graphical file pictures containing textual content and graphic additives are taken into consideration as two-dimensional indicators. The proposed set of rules fully depends upon a sparse representation [8] framework with the content as it should be chosen discriminative over complete dictionaries. Every one offers sparse illustration above one type of signal and non-sparse illustration above the other. Separation of text and image additives is obtained via promoting sparse graphic of input images in these dictionaries. The proposed approach overcomes the problem of handling among text and images.

S.Ranjini et al introduced a technique is from Digital English Comic Image Using Median Filter. In the present work, Blob extraction functions are used and Japanese text is taken out vertically from Manga Comic Image [9]. Along with this, simultaneously text is taken out from multiple limitations using optical character recognition (OCR) [10] and convert Japanese language of Manga into a few additional languages in the traditional way to proportion the satisfaction of studying Manga through the net.

III. DRAWBACKS OF EXISTING METHODOLOGY

In the existing work the text documentation is to group text into visually homogeneous blocks. From PDF document we separate the text components from the image components such as images, tables and graphs. Here, line segmentation is considered over a horizontal reading order. This method involves three modules which are text information retrieval, the merging of words into text lines and the grouping of text lines into text blocks.

- In the existing system considered only text part of PDF document
- Text information retrieval technique retrieves only text context from PDF documents
- Bounding boxes considered for single column
- In merging of word into text lines, the quads were sorted either in ascending order or descending order
- The merging of words used either top down or bottom up technique
- The grouping of text lines into text blocks failed to group the text in image of PDF documents
- List and tables are not considered in text segmentation
- Text belonging to map regions often has various orientation and excess character space.

These are the most challenging cases for text segmentation. Even though lot of issues are text and image segmentation so we, need a new proposed model to overcome those issues.

IV. PROPOSED METHODOLOGY

In the proposed work the segmentation of both text and image components in e-book PDF format is considered. This overcomes the restriction of segmentation in tables, images and graphs in PDF document. In this system both text layer and image layer are taken into consideration for segmentation. Each layer segments its data independently. Finally the results of both text and image layers are merged together for final segmentation. Text segmentation and image segmentation are used for reusable purpose. This work is composed of text segmentation and image segmentation. In text segmentation text content in the PDF documents of e-book is segmented and in image segmentation image objects are segmented. Hence considering text and image objects in PDF document, the accuracy, precision, recall and F-measure for segmented documents will be increased.

A. Graph based Text Segmentation

In PDF document the words and quads are accessed through Word Finder and visual attributes are retrieved. Then get the bounding boxes of quads. The bounding boxes of each word or quad may vary from one to another word and it also varies from one line to another. Additionally the vertical center lines are computed from the bounding boxes. Further the words or quads are merged into text lines by selecting up a quad that has no longer been assigned a line identity to begin a new line segment. Then the line is extended by adding qualified quads on both left and right to the line. When no qualified quad can be added to the line, a new line is started until all quads are assigned a new line identity.

Input : PDF document
Output : Text content in text pad
Step 1 *: Access words and quads in PDF document*
Step 2 ***:*** *Check the document in horizontal reading direction*
Step 3 ***:*** *Calculate geometric center point and form block*
Step4 *: For boundary detection assign*
$lineseg\,ment_{boundary} = -1$
$boundary_{id} = 0$
Step5 ***:*** *Define boundary for each line and increase the by 1*
$boundary_{id} = boundary_{id} + 1$
Step 6 ***:*** *Merge the lines using queue lines. =*
Step 7 ***:*** *Using Kruskal define the edge weight. Sort the edge weight in descending order and calculate mean and variance value*

$$Mean = \frac{1}{vertices-1}\sum_{n=1}^{vertices-1} w;$$

Variance= $\frac{1}{vertices-1}\sum_{n=1}^{vertices-1}[w(edges_n$

Step 8 ***:*** *Set the threshold value Θ=n*variance*
Step 9 ***:*** *Remove the edges w( )-Mean>Θ*
Step 10 ***:*** ***Calculate*** *angle distribution for segmentation Angle distribution=*

// *=|* $vertex_j$|,

=| $vertex_j$|

Step 11: *Calculate line spacing and word spacing word spacing>interline spacing Merge according to width of block.*

If horizontal distance between two words is smaller than threshold value those words are merged horizontally and we cannot consider the vertical distance between the words. Here we use font size, vertical center and width of the quad which are assigned as attributes to form line segment. After getting the text line segment we build homogeneous text blocks which avoid the pitfall by decoupling line space and font size. Relative difference between two line spaces is defined in Kruskal's algorithm which is distance between vertical center lines and finds the block boundary is found by comparing relative line space difference with a threshold value. Kruskal's algorithm is a minimum spanning tree algorithm that helps to find the block boundary which determines an edge of the least possible weight between the vertical centers.

B. Graph based Image Segmentation

In image segmentation process a digital image is partitioned into a number of segments. The image may contain tables, lists, and graphs. By this segmentation process those contents are partitioned separately and saved in required location. It is a hybrid method. In this system both text and image layers are taken into consideration for segmentation. Each layer segments its data independently. Finally the results of both text and image layers are merged together for final segmentation.

For every layout analysis the image objects are not considered in the segmentation. This is the main goal of this research work considering both textual and image objects. It is considered that image objects are spatially far away from text blocks. Then cluster properties of Delaunay tessellation neighborhood system are used to reject non-textual objects. For layout segmentation only the clusters in the text region are considered. Hence it plays an important role in reflow the able reconstruction of PDF document structure. There are two systems available to segment identification for PDF document pages. One is from the PDF path which is directly used to extract geometric features of bounding boxes and to group the elements into desired physical segments by image streams.

Thus the bounding box ensures to include the elements for graphics but the smallest bounding box that encloses white background which is invisible to users. Such issues will return inaccuracy for graphic segmentation. Additionally, when the path and image elements for making a holistic graphic composite are vast in numbers, the computational speed will be reduced for the grouping process. One more option is to utilize the well- researched image based segmentation methods. In this research work image objects are processed as a separate layer using traditional image analysis method. From the visual perspective component analysis is obtained. Local text features describe the spatial closeness of graphic objects. Merging process is required to detect graphic composite holistically. Thresholds are set for connected component grouping based on inter text line spacing. As for graphics embedded or surrounded by text elements added, post processing of integration is handled.

Graph based Image Segmentation Algorithm

Input: PDF document
Output: Image objects
Step 1*: Components are analyzed from visual perspective*
Step 2: *Geometric features get from component analysis*
Step 3: *Define interline spacing and Set threshold*
Interline spacing < threshold
Merge component objects

C. SVM Classification

In this module the objects are classified. SVM classifier uses test data and train data to classify the data. The output of layout analysis is bounding boxes of text line composite objects for text layer and graphic composite objects for image layer. Then from the analysis result of text or image layer a feature vector is extracted for each composite object. The source of feature extraction comes from different layers for classification which are indicated by character features of graphic components with zero.

It is the main difference between the text and image features. For both textual content and image content segments, all the segmented sub images are saved, and image features describing texture spectrum are extracted. In this SVM is used for classification of text and image objects. In this work, a larger sort of class labels is considered. The previous analysis within the document is taken into consideration for segment extraction where document is segmented into physical class labels such as footer text, body text, page number text, graphic text, and header. Multi-class SVM classifier is used in labeling task to discover the dissimilarity capacity of the presented features.

V. EXPERIMENTAL RESULTS

In our research work we used 50 e-book PDF documents to evaluate the performance of graph based approach in terms of accuracy, precision, recall, time measure and F-measure. In bottom up growing region approach it segments the text content in PDF document while in the graph based approach the image objects are segmented by using a hybrid method.

Sample input and output of graph based approach

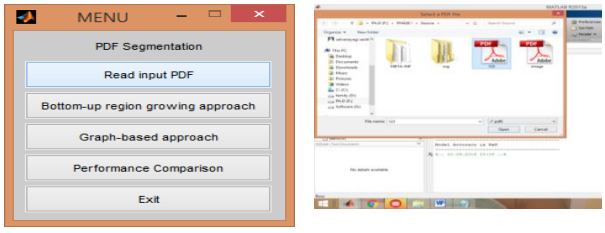


Fig. 1. *Input PDF Document. (figure caption)*

In the above Fig 1, PDF document is taken as input to evaluate text and image segmentation.

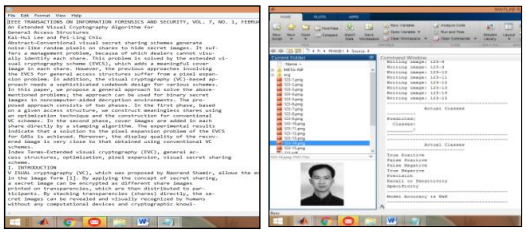


Fig. 2. *a) Text Segmentation b) Image Segmentation* Accuracy on Graph based approach and existing for text segmentation and image segmentation

Thus the image objects in the input PDF document such as graph, images, pictures and tables are segmented by using graph based approach which is shown in Fig 2(b). The text contents are segmented by using Kruskal's algorithm based approach as shown in Fig 2(a).The above fig 2, shows the first page text segmentation. Like this the text is segmented for whole document.

For performance evaluation, the proposed graph based text segmentation is compared with OCD document processing; XY cut segmentation and bottom up growing approach. Then for image segmentation the proposed graph based approach is compared with segmentation using connected components, Eigen vector and bottom up nearest neighbor application.

A. Accuracy

Social Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Accuracy is evaluated as

$$Accuracy = \frac{(Truepositive + Truenegative)}{Truepositive+Trunegative+Falsepositive+Falsenegative}$$

Fig.3 shows graph based segmentation for text which shows higher accuracy than the existing approaches.

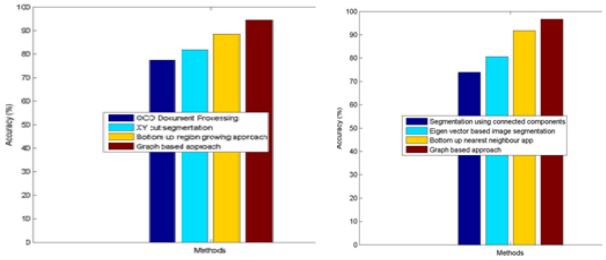


Fig. 3. *Accuracy on Graph based approach and existing for text Segmentation and image segmentation*

B. Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive

$$\frac{(Truepositive)}{}$$

$$Precision = \frac{}{(Truepositive + Falsepositive)}$$

Fig.4 shows graph based segmentation which shows higher precision than the existing approaches.

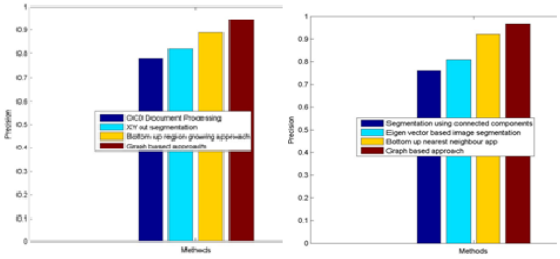


Fig. 4. *Precision on Graph based approach and existing for text segmentation and image segmentation*

C. Recall

The Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$Precision = \frac{(Truepositive)}{(Truepositive + Falsepositive)}$$

Fig 5 shows graph based segmentation which shows higher recall than the existing approaches.

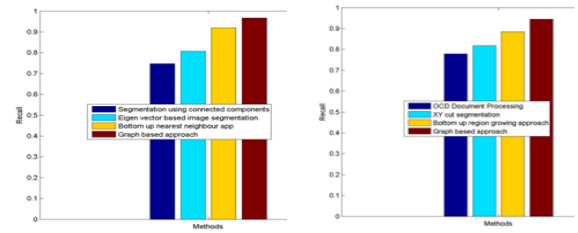


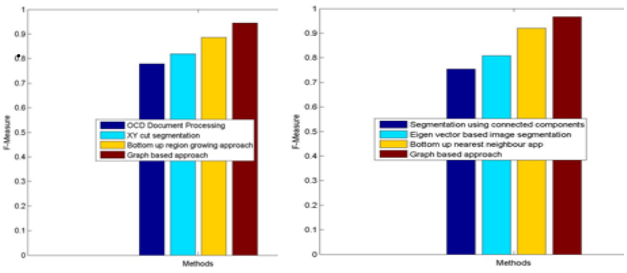
Fig. 5. *Recall as on Graph based approach and existing for text segmentation and image segmentation*

D. F-measure

F-measure is calculated from the precision and recall value. It is calculated as

$$f\text{-}measure = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

Fig 6 shows graph based segmentation which shows higher f-measure than the existing approaches.



The values of precision, recall, accuracy, time measure and f-measure are tabulated in the following table 1

TABLE I. COMPARISON TABLE

	Accuracy	*Precision*	*Recall*	*F-Measure*
Text Segmentation				
OCD Document Processing	77.5	0.8	0.8	0.8
XY cut segmentation	81.8	0.8	0.8	0.8
Bottom up region growing approach	88.5	0.9	0.9	0.9
Graph based approach	94.5	0.9	0.9	0.9
Image Segmentation				
Segmentation using connected components	73.9	0.8	0.7	0.8
Eigen vector based image segmentation	80.6	0.8	0.8	0.8
Bottom up nearest neighbor app	91.8	0.9	0.9	0.9
Graph based approach	96.6	1.0	1.0	1.0

VI. CONCLUSION

In this work, e-book PDF format is segmented considering both text objects and image objects. This process involves text layer and image layers processed separately and finally the objects are classified by SVM classifier. Then experimental results are conducted in various e-book PDF documents to prove that the proposed graph-based approach is better than the existing bottom up region approach in terms of accuracy, precision, recall, f-measure, time measure.

REFERENCES

[1] Gupta, N., &Banga, V. K. (2012, April). Image segmentation for text extraction. In 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE'2012) (pp. 182-185).

[2] Pasha, S., & Padma, M. C. (2015, December). Handwritten Kannada character recognition using wavelet transform and structural features. InEmerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on (pp. 346-351). IEEE.

[3] Adak, C. (2013, August). Unsupervised text extraction from G-maps. InHuman Computer Interactions (ICHCI), 2013 International Conference on (pp. 1-4). IEEE.

[4] Liu, J., Fan, X. Z., & Chen, K. (2007, October). Research on method of extracting Chinese domain terms based on rough and fuzzy clustering. InSemantics, Knowledge and Grid, Third International Conference on (pp. 366-369). IEEE.

[5] Chaple, G. N., Daruwala, R. D., & Gofane, M. S. (2015, February). Comparisons of Robert, Prewitt, Sobel operator based edge detection methods for real time uses on FPGA. In Technologies for Sustainable Development (ICTSD), 2015 International Conference on (pp. 1-4). IEEE.

[6] Gautam, A. (2013). Segmentation of Text From Image Document. International Journal of Computer Science and Information Technologies,4(3), 538-540.

[7] Tounsi,M., Mo Moalla, I., Alimi, A. M., & Lebouregois, F. (2015, August). Arabic characters recognition in natural scenes using sparse coding for feature representations. In Document Analysis

and Recognition (ICDAR), 2015 13th International Conference on (pp. 1036-1040). IEEE.
[8] O'Gorman, L. (1993). The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(11), 1162-1173.
[9] Nathiya, N., & Pradeepa, K. (2013, December). Optical Character Recognition for scene text detection, mining and recognition. In Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on (pp. 1-4). IEEE.
[10] Yuan, Q., & Tan, C. L. (2001). Text extraction from gray scale document images using edge information. In Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on (pp. 302-306). IEEE.
[11] Kumari, S., & Vijay, R. (2012). Effect of symlet filter order on denoising of still images. Advanced Computing, 3(1), 137.
[12] Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in images and videos. IEEE Transactions on circuits and systems for video technology, 12(4), 256-268.
[13] Wu, L., Shivakumara, P., Lu, T., & Tan, C. L. (2015). A New Technique for Multi-Oriented Scene Text Line Detection and Tracking in Video. IEEE Transactions on Multimedia, 17(8), 1137-1152.
[14] Ranjini, S., &Sundaresan, M. (2013). Extraction and Recognition of Text From Digital English Comic Image Using Median Filter. International Journal on Computer Science and Engineering, 5(4), 238.
[15] Mehta, A., Parihar, A. S., & Mehta, N. (2015, September). Supervised classification of dermoscopic images using optimized fuzzy clustering based Multi-Layer Feed-forward Neural Network. In Computer, Communication and Control (IC4), 2015 International Conference on (pp. 1-6). IEEE.
[16] Tehsin, S., Masood, A., &Kausar, S. (2014). Survey of Region-Based Text Extraction Techniques for Efficient Indexing of Image/Video Retrieval. International Journal of Image, Graphics and Signal Processing, 6(12), 53.
[17] Green, R., & Oliver, C. (2013, November). Layout analysis of book pages. In2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013) (pp. 118-123). IEEE.
[18] Hoang, T. V., &Tabbone, S. (2010, June). Text extraction from graphical document images using sparse representation. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (pp. 143-150). ACM.
[19] Moniz, N., & Rodrigues, F. (2012). Extracting Structure, Text and Entities from PDF Documents of the Portuguese Legislation. In KDIR (pp. 123-131).

