Segmentation of Textual and Non-Textual Regions from a Document

Rajath A N[1], Parashiva Murthy B M[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] GSSS Institute of Engineering and Technology for Women College, Mysuru

Abstract- *The proposed objective of this paper is to separate text and non-text regions from a document through segmentation based on histogram generation. Segmentation is a process of splitting an image in to its constituent's regions or objects. The level to which the sub-division is carried depends on the problem being solved. Image Segmentation algorithms generally are based on one or two basic properties of intensity values are Discontinuity and Similarity. In the first category, the approach is to partition an image based on abrupt changes in intensity such as edges in an image. The principle approaches in the second category are based on partitioning an image in to regions that are similar according to a set of pre-defined criteria, thresholding, region growing, and region splitting and merging are examples of methods in the second category. The input to the proposed model can be of two types. First, an image created using MS Paint. Secondly, a scanned newspaper image which is subsequent to pre-processing activities for removal of noise. Pre-processing is a method of enhancing the image for better feature extraction. After pre-processing, obtain a conditioned image input of the document, which implies that the document is noise free. Target of this project is to translate human identifiable document to machine identifiable codes. This process carried out using MATLAB R2013a software.*

Keywords- Horizontal Projection, Vertical Projection, Splitting Based on the Threshold Value.

I. INTRODUCTION

With the widespread use of computers in business and government sectors, organizations are converting their paper documents into electronic documents that can be processed further. Identifying the text and non-text regions from a document is written is one of the real-life applications that have motivated research issues in the context of pattern recognition and image processing based automatic document analysis and recognition. If a document has multiple segments, then both analysis and recognition become more severely challenging, as it requires the identification of the text and non-text regions before the analysis of the content could be made. The objective of segmentation of text and non-text regions is to translate human identifiable document to machine identifiable codes. This project is based on Image processing. Image Processing is enhancing an image or extracting information or features from an image. Document analysis and recognition is a challenging issue in the area of image processing. In document analysis and recognition, the major work is focused on the process of extracting the contents of the document. Image processing involves manual and digital techniques used to improve image geometry and appearance to identify factors and to extract selected information.

The field of digital image processing refers to processing digital images by means of a digital computer. A digital image is composed of a finite number of elements, each of which has a particular location and value. These elements are referred to as picture elements, image elements, and pixels. Pixel is the term most widely used to denote the elements of a digital image.

The rest of this paper is organized as follows. The next section composes a review of similar researches that have been implemented and tested for segmentation of text and non-textual regions from a document. In Section III, the proposed method is described. In Section IV, the horizontal projection, vertical projection and splitting based on the threshold value are discussed in detail. In Section V, experimental results are reported. Finally, some conclusions are given and future work is proposed in Section VI.

II. REVIEW OF OTHER METHODS

This section provides a descriptive summary of some methods that have been implemented and tested for Segmentation of text and non-text regions. In some existing system, the text separation has been done based on single character recognition [1][2]. Some system has used robust technique where it first decomposes the object into separate object planes. This technique cannot identify text and it considers text as image only and is time consuming. In some other existing system separation of text and non-text is done using multiplane segmentation approach based on DWT [3][4] (Discrete Wavelet Transform). In these system output, may be a blurred image or may not be noise free. This approach can only be used for lower quality images than jpeg images.

In our proposed system, Segmentation of Textual and non-Textual regions from a document separation is done using segmentation algorithm based on histogram [5] being generated. The input to the proposed model can be of two type's. First, an image created using MS Paint. Secondly, a scanned newspaper image which is subsequent to pre-processing [6] activities for removal of noise. Pre-processing is a method of enhancing the image for better feature extraction. After pre-processing, we obtain a conditioned image input of the document, which implies that the document is noise free.

III. PROPOSED METHOD

The proposed objective of this project is to separate text and non-text regions from a document through segmentation based on histogram generation.

Segmentation is a process of splitting an image in to its constituent's regions or objects. The level to which the sub-division is carried depends on the problem being solved. Image Segmentation algorithms generally are based on one or two basic properties of intensity values are Discontinuity and Similarity. In the first category, the approach is to partition an image based on abrupt changes in intensity such as edges in an image. The principle approaches in the second category are based on partitioning an image in to regions that are similar according to a set of pre-defined criteria, thresholding, region growing, and region splitting and merging are examples of methods in the second category.

The input to the proposed model can be of two types. First, an image created using MS Paint. Secondly, a scanned newspaper image which is subsequent to pre-processing activities for removal of noise. Pre-processing is a method of enhancing the image for better feature extraction. After pre-processing, we obtain a conditioned image input of the document, which implies that the document is noise free.

Target of this project is to translate human identifiable document to machine identifiable codes. This process carried out using MATLAB R2009a software. MATLAB is a high-performance language for technical computing. General scheme for segmenting text and non-text regions from a document is shown in figure (1).

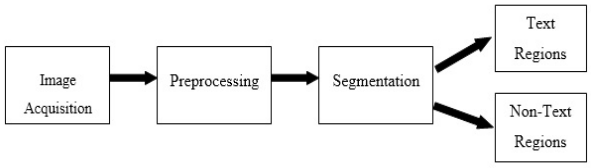

Figure 1. Proposed Model for Segmentation of Textual and Non-Textual Regions from a Document

IV. METHODOLOGY

The methods used to segment text and non-text portions from a document:

A. Horizontal projection
B. Vertical projection
C. Splitting based on a threshold value

A. Horizontal Projection: Horizontal projection is the row wise sum of black pixels of the image under consideration.

Feature extraction: Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features or attributes that are effective in discriminating pattern classes. In this method features are extracted from horizontal projection. The algorithm for horizontal projection is as follows:

Input: A .bmp image or a .jpeg image to the proposed model.

Output: Segmented text and non-text regions.

1. Apply horizontal projection and take the row-wise sum of black pixels.
2. Count the number of continuous sequence of black pixels and store them in an array.
3. Find the biggest value in the array obtained in step 2.
4. Find the position of the biggest value and finally crop that region.

B. Vertical Projection: Vertical projection is the column wise sum of black pixels of the image under consideration.

Feature extraction: Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features or attributes that are effective in discriminating pattern classes. In this method features are extracted from vertical projection. The algorithm for vertical projection is as follows:

Input: A .bmp image or a .jpeg image to the proposed model.

Output: Segmented text and non-text regions.

1. Apply vertical projection and obtain the column-wise sum of black pixels.
2. Plot the sum of black pixels using a histogram.

3. Segment the two portions based on a continuous sequence of column-wise white pixels.
4. Identify which is the text portion and which is the non-text portion based on the following two issues:

- Intensity value.
- Taking the difference between the adjacent pixel's values.

C. Splitting Based on a Threshold Value: In this method, we apply both the horizontal projection and vertical projection to segment text and non-text regions.

Feature extraction: Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features or attributes that are effective in discriminating pattern classes. In this method features are extracted from vertical projection.

The algorithm for splitting based on threshold value is as follows:

Input: A .bmp image or a .jpeg image to the proposed model.

Output: Segmented text and non-text regions.

1. Decide whether to apply horizontal or vertical projection based on a threshold value.
2. Apply the appropriate projection technique.
3. If the obtained result of step 2 still contains both text and non-text portion repeat the steps (1) and (2) again

V. EXPERIMENTAL RESULTS

All experiments were done on Pentium Dual Core 2.10 GHz with 4GB RAM under Matlab R2013a environment. In the experiments, 120 images were scanned document and prepared our own dataset using ms-paint and the size of the images is 490x367 pixels. The satisfactory result has been obtained: the success of detection rate of segmenting text and non-text regions is up to 85%.

The following figures shows the result of different projections as follows: (A) Horizontal Projection, (B) Vertical Projection and (C) Splitting Based on a Threshold Value.

A. Horizontal Projection

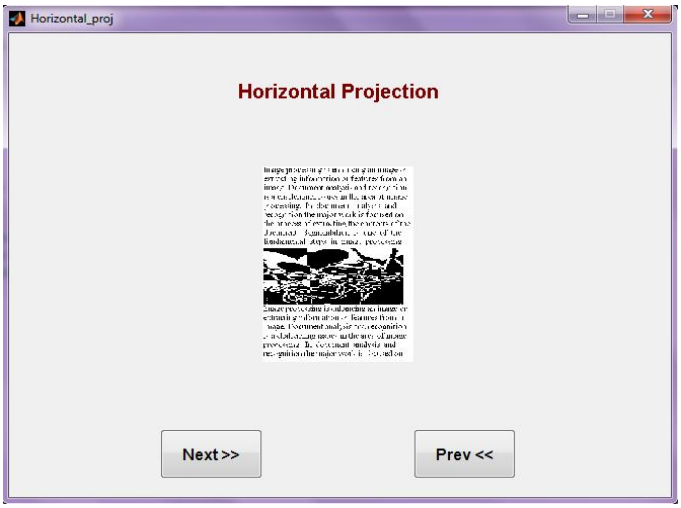

Figure 1.

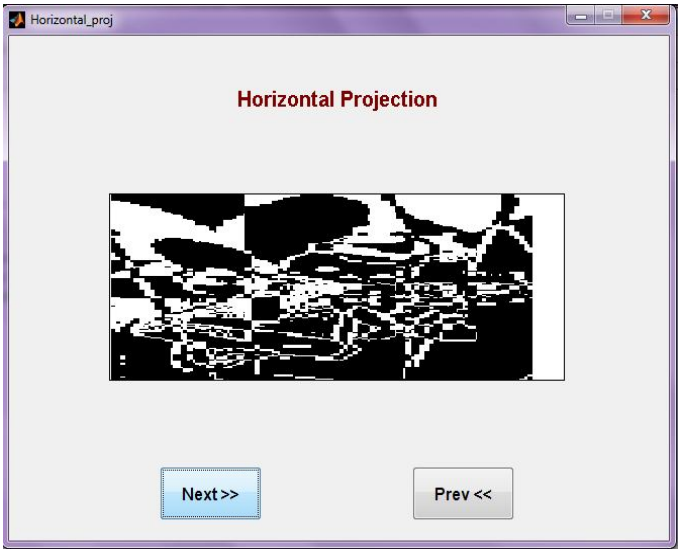

Figure 2.

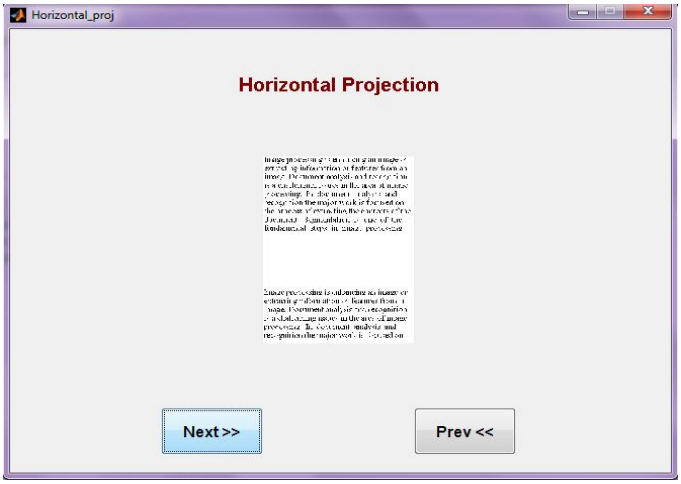

Figure 3.

B. Vertical Projection

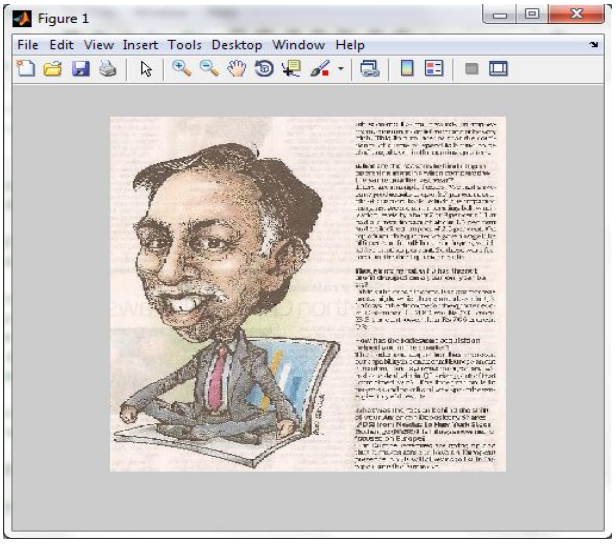

Figure 4.

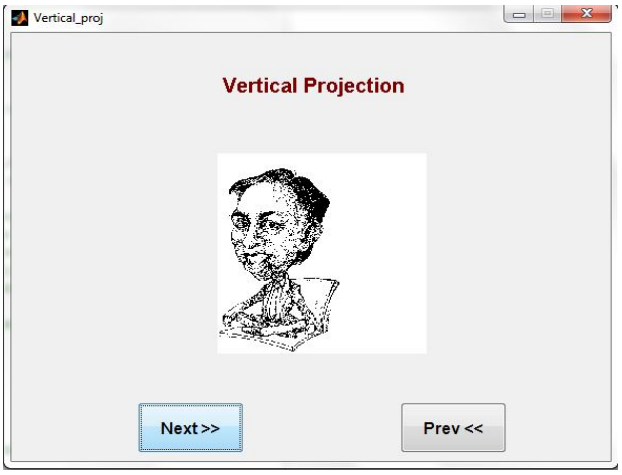

Figure 5.

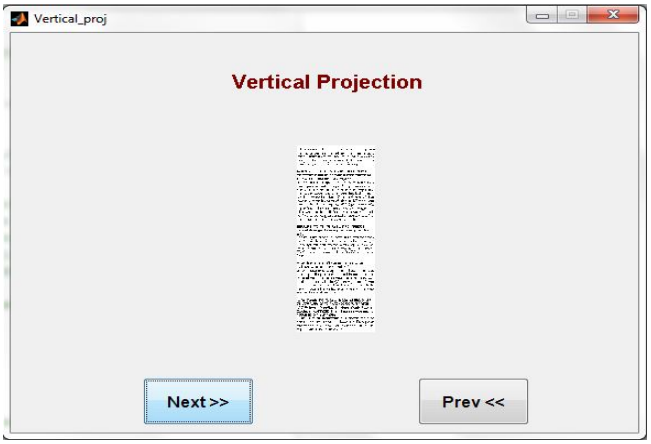

Figure 6.

C. Splitting Based on a Threshold Value

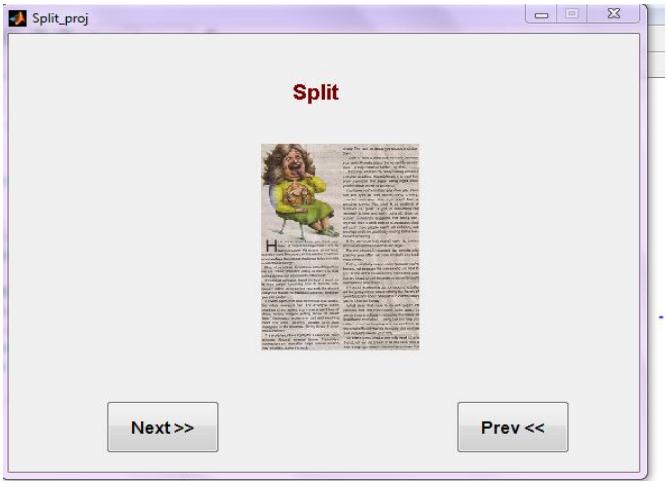

Figure 7.

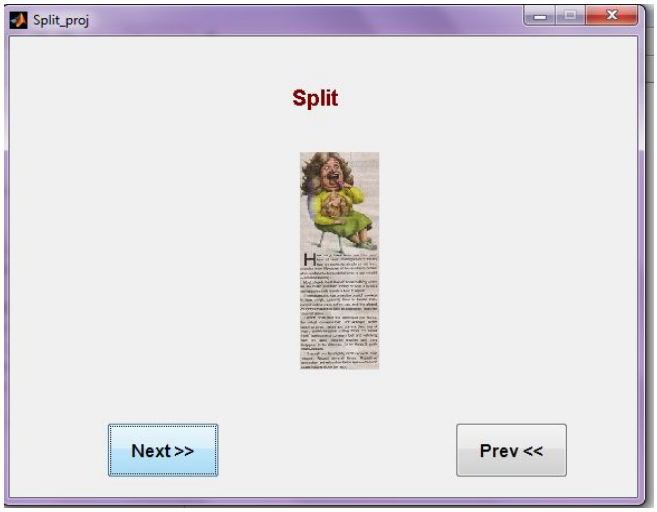

Figure 8.

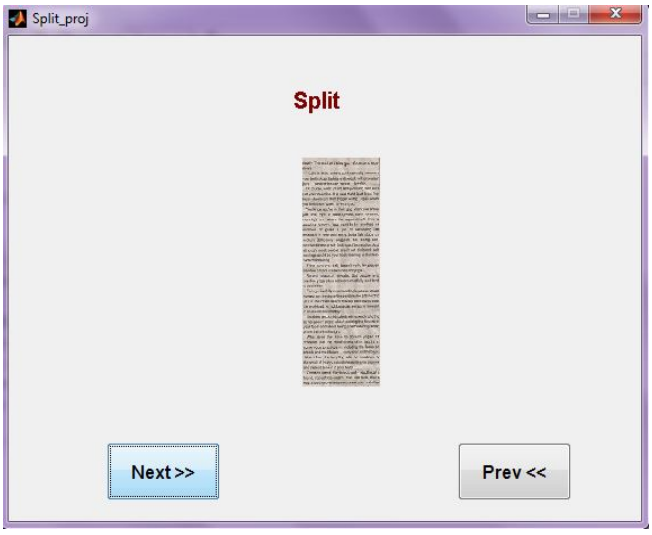

Figure 9.

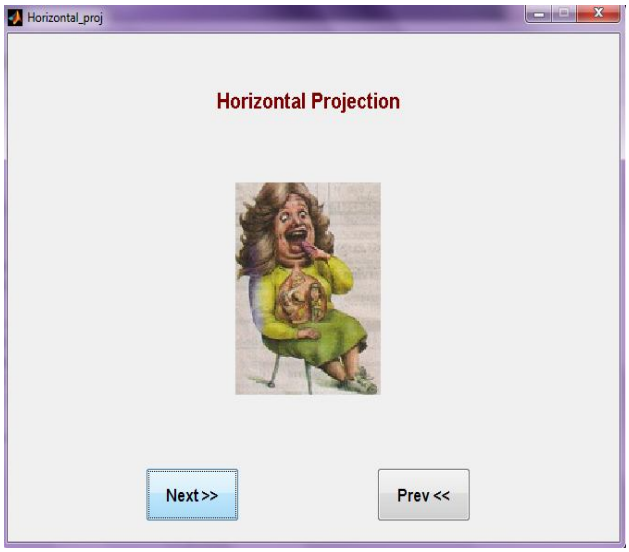

Figure 10.

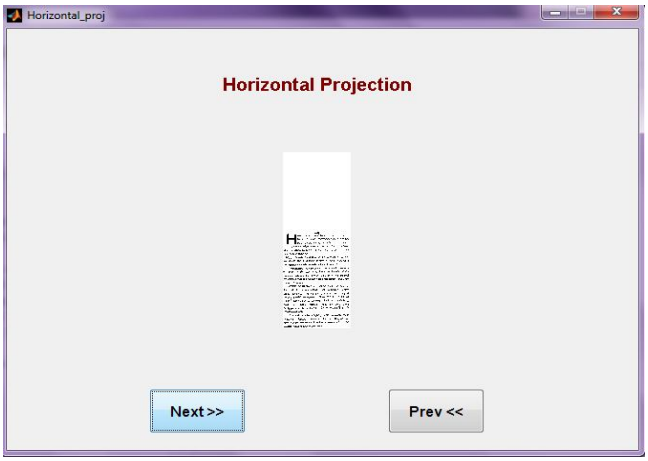

Figure 11.

VI. CONCLUSION

The main task of this project was to implement a system which would be able to separate a text and non-text regions from a document. If a document has multiple segments, then both analysis and recognition becomes severely challenging, as it requires the identification of the text and non-text regions before the analysis of the content could be made. The objective of segmentation of text and non-text regions is to translate human identifiable document to machine identifiable codes.

Our project segmentation of textual and non-textual regions from a document provides various ways like horizontal projection, vertical projection and split & merge projection for identification and separation of textual and non-textual regions from a document based on how the text and image are arranged.

At present in our project the textual and non-textual regions cannot be separated for tilt scanned images, further it can be enhanced for the same. In future, it can be enhanced to identify the numerical figures, which is much complicated as the numerical are written overlapping each other. At present in our project it can identify only high contrast images further it can have implemented for all kinds of images (low and high contrast).

REFERENCES

[1] Yen-Lin Chen, “A Robust Technique for Character String Extraction from Complex Document images”, Asia University.

[2] Neha Gupta, V.K Banga, “Image Segmentation for text extraction” Singapore 2012

[3] James.L.Fisher, “A Rule Based System for Image Segmentation”, presented at IEEE symposium on image segmentation.

[4] Yen-Lin Chen, “A knowledge-based approach for Textual Information Extraction from Mixed Text/Graphics Complex Document Images”, National Taipei University of Technology.

[5] Rajath A N, “An Adaptive Approach: Text Line Extraction from Multi-Skewed Hand Written Documents”, IJCSET, June 2015, Vol 5, Issue 6, ISSN:2231-0711, Page no.158-161.

[6] Rajath A N and Parashiva Murthy B M, ”An Adaptive Approach to Vehicle Number Plate Detection for Indian Style Based”, International Journal of Modern Computer Science (IJMCS), Volume 4, Issue 5, October, 2016, ISSN: 2320-7868 (Online), Page no.: 31-36.