Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Novel Approach for Detection and Extraction of Textual Information from Scanned Document Images and Scene Images

Vandana Gupta [*], Kanchan Singh
Dept. of CSE, ABES-EC,
Ghaziabad, India

Abstract*— *This paper presents a new approach for detection and extraction of text data from both scanned document images and scene images. Text detection and extraction is performed in a four-step approach that consists of the pre-processing which include binarization and noise removal of an image, image segmentation using connected component analysis, feature extraction using variance generation and finally classification by choosing a threshold value of variance property. Experiments conducted on large database of mixed scanned document image and ICDAR databases of scene images demonstrate the validity of this approach.

Keywords*— *Binarization ,Segmentation ,variance,Decesion based Classifier.

I. Introduction

The purpose of document and scene image analysis is to transform the information of a digitized image into an equivalent symbolic representation usable by important information retrieval systems. In order to guarantee an adequate processing of mixed-mode documents and scene images, the discrimination of text and non-text is required.

C. Strouthopoulos and N. Papamarkos [1] proposed a method based on a new set of features combined with a self-organized neural network classifier. The set of features corresponds to the contents and the relationship of 3 3 3 masks, was selected by using a statistical reduction procedure, and provided texture information. Next, a Principal Components Analyzer (PCA) was applied, which results in a reduced number of 'effective' features. Then the final set of features utilized as input vector into a proper neural network to achieved the classification goal. The neural network classifier was based on a Kohonen Self Organized Feature Map (SOFM). Datong Chen et al. [2] proposed a two-step approach that combined the speed of a text localization step, enabling text size normalization, with the strength of a machine learning text verification step applied on background independent features. Text recognition, applied on the detected text lines, was addressed by a text segmentation step followed by a traditional OCR algorithm within a multi-hypotheses framework relying on multiple segments, language modeling and OCR statistics. Sunil Kumar et al.[3] proposed a clustering-based technique for estimating globally matched wavelet filters using a collection of groundtruth images. Fisher classifiers have been used for classification. S.Audithan and RM. Chandrasekaran [4] proposed Haar discrete wavelet transform (DWT) which operated the fastest among all wavelets because its coefficients were either 1 or -1. First, detected edges and then line feature vector graph was generated based on the edge map and the stroke information was extracted. Finally text regions generated and filtered according to line features. Emanuel Indermühle et al. [5] introduced two systems. The first system generated document segmentation. For this purpose, four methods originally developed for machine printed documents: x-y cut, morphological closing, Voronoi segmentation, and whitespace analysis. A state-of-the art classifier then distinguished between text and non-text zones. The second system followed a bottom-up approach that classified connected components. Syed Saqib Bukhari et al. [6] proposed a method in which segmentation approach introduced connected component based classification and self-tunable multi-layer perceptron (MLP) classifier for distinguishing between text and non-text connected components using shape and context information as a feature vector. Marios Anthimopoulos et al. [7] proposed a two-stage system for text detection in video images. In the first stage, text lines detected based on the edge map of the image leading in a high recall rate with low computational time expenses. In the second stage, the result was refined with the help of a sliding window and an SVM classifier trained on features obtained by a new Local Binary Pattern-based operator (eLBP) that describes the local edge distribution. Ming Zhao et al.[8] proposed a classification-based algorithm for text detection with the help of sparse representation with discriminative dictionaries. First, the edges detected by the wavelet transform and scanned into patches by a sliding window. Then, candidate text areas were obtained by applying a simple classification procedure using two learned discriminative dictionaries. Finally, the adaptive run-length smoothing algorithm and projection profile analysis used to further refine the candidate text areas. Lukas Neumann and Jiri Matas [9] proposed method which consisted of (i) departs from a strict feed-forward pipeline and replaces it by a hypotheses- verification framework simultaneously processing multiple text line hypotheses, (ii) used synthetic fonts to train the algorithm eliminating the need for time-consuming acquisition and labeling of real-world training data and (iii) exploits Maximally Stable External Regions (MSERs) which provided robustness to geometric and illumination conditions. Wonder Alexandre et al. [10] proposed a method for localizing text regions within scene images consisting of two major stages. In the first stage, a set of potential text regions was extracted from the input image with the help of residual operators. In the second stage a set of features was obtained from each potential text region and this feature set used as an input to a decision tree classifier in order to label these regions as text or non-text regions. Vinay Raj Hampapur et al.[11] evaluated two methods: Segmentation by dilation and Smart Segmentation that effectively isolate text at different distances, then evaluated three methods namely Correlation, Shape Context and Harris Corner that would discern between text using the user defined text to synthetically generated corresponding template. T. Hoang Ngan Le et al.[12]proposed a adaptive binarization algorithm based on ternary entropy-based approach. In this algorithm first the contrast of intensity was estimated by a grayscale morphological closing operator. A double-threshold was generated by Shannon entropy based ternarizing method to classify pixels into text, near-text, and non-text regions. Chitrakala Gopalan and D. Manjula[13] proposed a statistical unified approach with the help of multi level feature priority (MLFP) algorithm. MLFP feature selection algorithm was evaluated with three common ML algorithms: a decision tree inducer (C4.5), a naive Bayes classifier, and an instance based K-nearest neighbor learner. Alvaro Gonzalez et al.[14] proposed a system which consisted of three main stages: a segmentation stage to find character candidates, a connected component analysis based on fast-to-compute but robust features to accept characters and discard non-text objects, and finally a text line classifier based on gradient features and support vector machines. Mehdi Felhi et al. [15] proposed a skeleton based descriptor to describe the strokes of the text candidates that composed a spatial relation graph, then applied the graph cuts algorithm to label the nodes of the graph as text or non-text. Finally refined the resulted text lines candidates by classifying them using a kernel SVM. Ayatullah Faruk Mollah et al.[16] proposed a method in which a given image was partitioned into blocks that are assigned two types of fuzzy memberships. The membership values post-processed for finer classification as foreground block or background block. Then, a feature-based Multi Layer Perceptron was used to classify the foreground components as text or non-text. Mohamed Benjelil et al. [17] proposed a system based on steerable pyramid transform. The features extracted from pyramid sub-bands served to locate and classify regions into text and non-text in some noise-infected, deformed, multilingual, multi-script document images.

II. Proposed Methodology

This Methodology consists of four steps which are shown in Fig-2.1:

(A)Pre-processing: This process comprises of two steps: image binarization and noise removal. Image binarization is the process that converts an image of up to 256 gray levels to a black and white image i.e binary image.

In this approach for Image binarization a global threshold value is computed to classify all pixels of image such that values above this threshold value as white and all other pixels as black. Noise removal is also done by this threshold value.

(B)Segmentation using connected component analysis: Image segmentation is the process of subdividing a digital image into multiple regions or objects that share similar properties or features according to the set of predefined criteria. The basic use of image segmentation is to find out the location and boundaries of each object in digital image. More accurately, segmentation process assigns a label to each and every pixel in an image such that pixels with the same label share certain visual characteristics. In this approach segmentation is done by connected component analysis.

Connected component analysis: This process defines which pixels are connected to other pixels. A connected component is 8-connected" if diagonally adjacent pixels are considered to be touching; otherwise, it is 4-connected. The types of neighborhood choose affects the number of objects and the boundaries of those objects in binary images. Then connected component labeling is used to identify each object in binary image. In this approach 8 adjacency is used.

After identification of connected object in binary image, region properties are used to measures a set of properties for each labeled region in the label matrix Com. Positive integer elements of Com correspond to different regions. In this process one region property is used for all regions i.e BoundingBox.

BoundingBox: The smallest rectangle containing the each region of label matrix means one best-fit rectangle for one region. after drawing these rectangles a height histogram is drawn which provides numbers of Boundingbox of same height. Then we analyzed that height of all bounding box for text data is different from the height for non-text data. So this histogram provides a common height to identify text and non-text region and this information is used to calculate threshold value of variance property.

(C)Variance property based Feature Extraction: Feature extraction means dimensionality reduction of an image. When the input data size to an algorithm is too large for processing and it seems to be notoriously redundant then the input data will be transformed into a reduced representation set of features also called features vector. Transformation of the input data into the set of features is called *feature extraction*. In this approach variance property is used to feature extraction. This is the block-based property in which the pixel intensity value of text block is higher than the pixels intensity value of non-text block .

Variance Function: This function computes the variance of each column and each row in the input image, or tracks the variance of a sequence of inputs over a period of time.

(D)Classification: In image processing classification is the last step to segregate text and non-text region from an image. This process includes a wide range of Decision-theoretic approaches to the classification of an image. In this paper supervised classification is used in which a common threshold value is computed on the basis of variance graph (row wise and column wise) in which all pixels value above than threshold as text and below than threshold value as non-text. To draw these graph firstly number of rows and number of columns of an image are calculated then Row variance graph is drawn between variance and number of rows in image similarly column variance graph is drawn between variance and number of column in image.

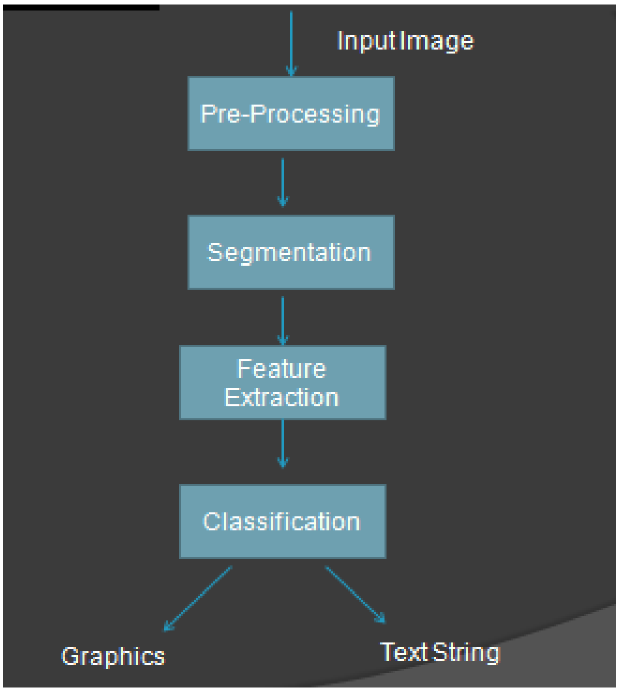


Fig 2.1: Flow Diagram of the Proposed Work

After plotting these graphs a threshold value is computed with the help of height histogram which was calculate in segmentation section such that the variance value greater than threshold values as text and below than as non-text value. Then crop non-text data from image such that text and non-text data segregation is done.

III. IMPLEMENTATION OF METHODOLOGY

Various matlab functions are used to implement this methodology. This section shows step by step implementation:
In first step Binarization and noise removal is done with the help of threshold value. Function used to calculate the threshold value is:

TH=graythresh(I);……………………….(1)

Where I is the original image and TH will store the value of threshold.
And the function for binarization is:

B=im2bw(I,TH);………………………...(2)

Where B is binary image.
In second step segmentation is done by using connected component analysis and bounding box. The function for labeling connected component is:

[Com, N] = bwlabel(B,n)

Where Com is label matrix of same size as B,
B is binary image matrix,
N is number of connected objects in B and
n is the value of pixel connectivity.
if 8-adjacency is used then n can be omitted. Then

[Com, N] = bwlabel(B)………………….(3)

The function to compute region property is:

BOX = regionprops(Com,'BoundingBox')(4)

Where BOX is a structure array to defining measurement of each array,
Com is label matrix and
“BoundingBox’’ is region properties.

In third step features are extracted on based of variance property.The variance function is:

rV=var(B,0,2)……………..……………..(5)
cV=var(B,0,1)……………..……….......(6)

Where B is an input image (i.e binary image), dim is dimension of B along which variance is calculated (where dim =2 for row wise variance and dim=1 for column wise variance) and w is the weight vector where w=0 to use the default normalization. rV and cV stores the values of row variance and column variance respectively.
Function to calculate row and column number of matrix is:

[ro,col]=size(B);…………………………(7)

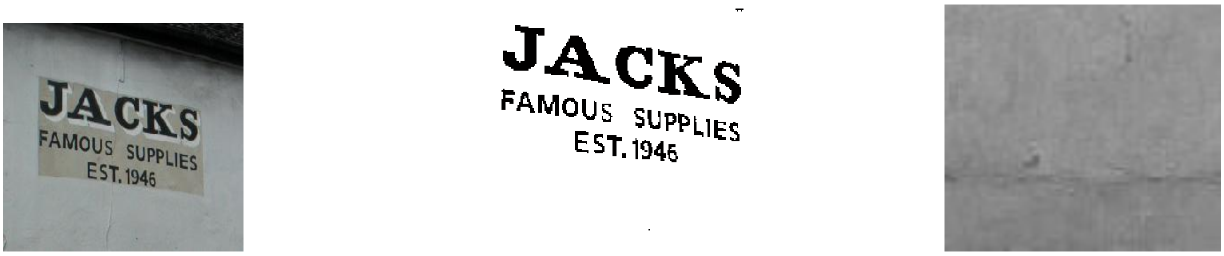
Where ro store number of rows and column stores number of columns.
Function to plotting of the graphs is:

plot(1:ro,rV);…………………………….(8)
plot(1:col,cV);…………………………...(9)

After the completion of the previous step the fourth step is classification. In this step a appropriate threshold value is choosen such that all pixel value above threshold as text and below threshold value as non-text.

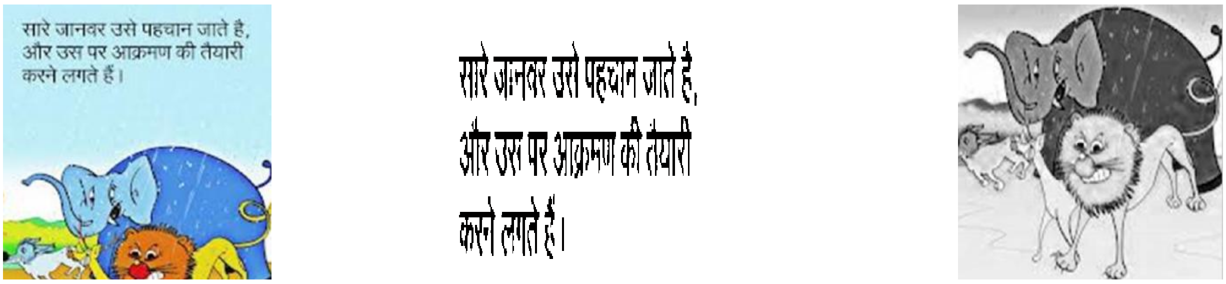
IV. RESULT AND DISCUSSION

We evaluated this proposed approach on the public datasets ICDAR for near about 100 scene images and a database of 40 images for mixed document images. All the images of the ICDAR datasets are natural scene ones and were captured using a digital camera under different luminosity conditions. It is evident from the results that the performance of the proposed algorithm is satisfactory on both types of images as shown in Fig 3.1 and 3.2.



(a)Original Scene image (i)Text extraction from image(a) (ii)Non-Text extraction from image(a)

Fig: 3.1



(b)Original Document image (i)Text extraction from image (b) (ii)Non-Text extraction from image(b)

Fig: 3.2

V. CONCLUSION

A robust and effective algorithm for separating text from mixed documents image and scene image has been presented. The algorithm accepts the text in various font sizes. Even low quality printings where characters are split into several connected components can be analyzed. The main component of the proposed system is variance generation of the image to detect text and non-text data then compute a common threshold value to extract text data from images. The algorithm has been validated on ICDAR dataset for scene images and own dataset of mixed document image.

REFERENCES

[1] C. Strouthopoulos and N. Papamarkos,“Text identification for document image analysis using a neural network,” Image and Vision Computing 16, PII S0262-8856(98)00055-9, pp. 879-896, 1998.

[2] Datong Chen, Jean-Marc Odobez, Herve Bourlard, “Text detection and recognition in images and vedio frames”, Pattern Recognition 37 (2004) 595 – 608, doi:10.1016/j.patcog.2003.06.001,2004.

[3] Sunil Kumar, Rajat Gupta, Nitin Khanna, Santanu Chaudhury, and Shiv Dutt Joshi, “Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model,” IEEE Trans. On image processing, Vol. 16, No. 8, Digital Object Identifier 10.1109/TIP.2007.900098, 2007.

[4] S.Audithan and RM. Chandrasekaran, “Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform, ”European Journal of Scientific Research, ISSN 1450-216X ,Vol.36, No.4 , pp.502-512,2009.

[5] Emanuel Indermühle, Horst Bunke, Faisal Shafait and Thomas Breuel,” Text versus non-Text Distinction in Online Handwritten Documents,” ACM, 978-1-60558-638-0/10/03, pp.22–26, 2010.

[6] Syed Saqib Bukhari, Mayce Ibrahim Ali Al Azawi and Faisal Shafait “Document Image Segmentation using Discriminative Learning over Connected Components,” ACM, 978-1-60558-773-8/10/06, 2010.

[7] Marios Anthimopoulos, Basilis Gatos and Ioannis Pratikakis, “A two-stage scheme for text detection in video images,” Image and Vision Computing 28, pp .1413–1426, doi:10.1016/j.imavis.2010.03.004,2010.

[8] Ming Zhao, Shutao Li and James Kwok , “Text detection in images using sparse representation with discriminative dictionaries,” Image and Vision Computing 28 ,pp. 1590-1599, doi:10.1016/j.imavis.2010.04.002,2010.

[9] Lukas Neumann and Jiri Matas, “ A method for text localization and recognition in real-world images,” In proc. of ACCV, pp. 8-12, 2010.

[10] Wonder Alexandre, Luz Alves and Ronaldo Fumio Hashimoto, ‘’ Text Regions Extracted from Scene Images by Ultimate Attribute Opening and Decision Tree Classification,” In proc. Of SIBGRAPI,DOI 10.1109/SIBGRAPI.2010.55,2010.

[11] Vinay Raj Hampapur, Tahrina Rumu and Umit Yoruk,” Key-word Guided Word Spotting In Printed Text”, In proc. of image processing project, 2011.

[12] T. Hoang Ngan Le, Tien D. Bui and Ching Y. Suen, “Ternary Entropy-based Binarization of Degraded Document Images Using Morphological Operators,” In proc. Of International Conference on Document Analysis and Recognition, DOI 10.1109/ICDAR.2011.32,2011.

[13] Chitrakala Gopalan and D. Manjula, “Statistical modeling for the detection, localization an extraction of text from heterogeneous textual images using combined feature scheme,” SIViP, 5:165–183,2011.

[14] Alvaro Gonzalez, Luis M. Bergasa, J. Javier Yebes and Sebasti´an Bronte, “Text Location in Complex Images,” In proc. of International Conference on Pattern Recognition,978-4-9906441-0-9,2012.

[15] Mehdi Felhi, Nicolas Bonnier and Salvatore Tabbone,“ A Skeleton Based Descriptor for Detecting Text in Real Scene Images,” In proc. of International Conference on Pattern Recognition,pp. 11-15, 2012.

[16] Ayatullah Faruk Mollah and Subhadip Basu and Mita Nasipuri, “Text Detection from Camera Captured Images Using a Novel Fuzzy-based Technique,” In proc. of International Conference on Emerging Applications of Information Technology, 978-1-4673-1827-3/12, 2012.

[17] Mohamed Benjelil, Rémy Mullot and Adel M. Alimi, “Page segmentation based on Steerable Pyramid features,” In proc. of International Conference on Frontiers in Handwriting Recognition, DOI 10.1109/ICFHR.2012.253,2012.