Scene Text Detection using the Regression Tree Technique

Bhumi B. Patel[1] Neeta Chaudasama[2]
[1,2]Department of Computer Engineering
[1,2]Parul Institute of Engineering Institute of Technology, Limda, Gujarat, India

Abstract— Text detection and localization in natural scene images is important for content-based image analysis. This problem is challenging due to the complex background, the non-uniform illumination, the variant of text font, size and line orientation. In this paper, we present a hybrid approach to robustly detect and localize texts in natural scene images. A text region detector is designed to estimate the text existing confidence and scale information in image pyramid, which help segment candidate text components by local binarization. To efficiently filter out the non-text components, a conditional random field (CRF) model considering unary component properties and binary contextual component relationships with supervised parameter learning is proposed. Finally, text components are grouped into text lines/words with a learning-based energy minimization method. Since all the three stages are learning-based, there are very few parameters requiring manual tuning. Experimental results evaluated on the ICDAR2005 competition dataset show that our approach yields higher precision and recall performance compared with state-of-the-art methods. We also evaluated our approach on a multilingual image dataset with promising results.

Key words: Conditional Random Field (CRF), Connected Component Analysis (CCA), Text Detection, Text Localization
General Terms: Text detection

I. INTRODUCTION

Image processing is a physical process used to convert an image signal into a physical image. Fig.1 shows, Image acquisition is the first process in image processing that is used to acquire digital image. Image enhancement is the simplest and most appealing areas of digital image processing. The idea behind enhancement techniques is to bring out detail that is obscured, or simply to highlight certain features of interest in an image. Recognition is the process that assigns a label to an object based on its description. This is the act of determining the properties of representing the region for processing the images. Information Extraction (IE) is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases, this activity concerns processing, human language texts by means of Natural Language Processing (NLP). Recent activities in a multimedia document processing like automatic annotation and concept extraction out of images/audio/video could be seen as information extraction. [2] [3]

Existing system presented a hybrid approach to robustly detect and localize texts in natural scene images by taking advantages of both regions-based and CC-based methods. This system consists of three stages are the preprocessing stage, the connected component analysis stage and text grouping.

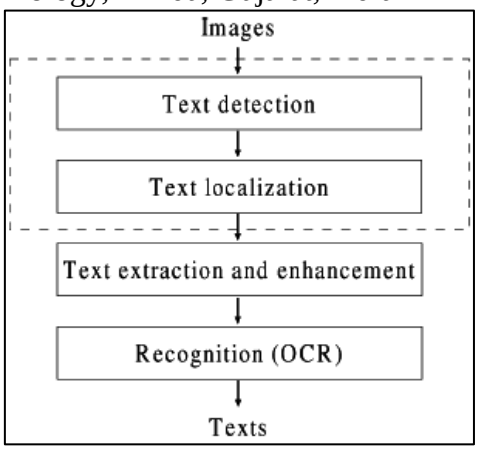

Fig. 1: Flow chart

At the preprocessing stage, a text region detector is designed to detect text regions in each layer of the image pyramid and project the text confidence and scale information back to the original image, scale-adaptive local binarization is then applied to generate candidate text components. At the connected component analysis stage, [4] [5] [7] a CRF model combining unary component properties and binary contextual component relationships is used to filter out non-text components. At the last stage, neighboring text components are linked to a learning-based minimum spanning tree (MST) algorithm and between-line/word edges are cut off with an energy minimization model to group text components into text lines or words. And also describes the binary contextual component relationships, in addition to the unary component properties, are integrated in a CRF model, whose parameters are jointly optimized by supervised learning. But this approach fails on some hard-to-segment texts. Although the existing methods have reported promising localization performance, there still remain several problems to solve. For region-based methods, the speed is relatively slow and the performance is sensitive to text alignment orientation. On the other hand, CC-based methods cannot segment text components accurately without prior knowledge of text position and scale. Here, designing fast and reliably connected component analyzer is difficult since there are many non-text components which are easily confused with texts when analyzed individually.

II. RELATED WORKS

Most regions-based methods are based on observations that text regions have distinct characteristics of non-text regions such as the distribution of gradient strength and texture properties. Generally, a region-based method consists of two stages: 1) text detection to estimate text existing confidence in local image regions by classification, and 2) text localization to cluster local text regions into text blocks, and text verification to remove non-text regions for further processing.

An earlier method proposed by Wu *et al.* uses a set of Gaussian derivative filters to extract texture features from local image regions. With the corresponding filter responses, all image pixels are assigned to one of three classes ("text", "non text" and "complex background"), then c-means clustering and morphological operators are used to group text pixels into text regions.

Li *et al.*proposed an algorithm for detecting texts on video by using first- and second-order moments of wavelet decomposition responses as local region features classified by a neural network classifier. Text regions are then merged at each pyramid layer and further projected back to the original image map.

Recently, Weinman *et al* as a CRF model for patch-based text detection. This method justifies the benefit of adding contextual information to the traditional local region-based text detection methods. Their experimental results show that this method can deal with texts of varying scales and alignment orientations.

To speed up text detection, Chen and Yuille [5] proposed a fast text detector using a cascade AdaBoost classifier, whose weak learners are selected from a feature pool containing gray-level, gradient and edge features. Detected text regions are then merged into text blocks, from which text components are segmented by local binarization. Their results on the ICDAR 2005 competition dataset show that this method performs competitively and is more than 10 times faster than the other methods. Unlike region-based methods, CC-based methods are based on observations that texts can be seen as a set of connected components, each of which has distinct geometric features, and neighboring components have close spatial and geometric relationships. These methods normally consist of three stages: 1) CC extraction of segment candidate text components from images; 2) CC analysis to filter out non-text components using heuristic rules or classifiers; and 3) post-processing to group text components into text blocks (e.g., words and lines).

The method of Liu *et al.* extracts candidate CCs based on edge contour features and removes non-text components of wavelet feature analysis. Within each text component region, a GMM is used for binarization by fitting the gray-level distributions of the foreground and background pixel clusters.

Zhang *et al* presented a Markov random field (MRF) method for exploring the neighboring information of components. The candidate text components are initially segmented with a mean-shift process. After building up a component adjacency graph, a MRF model integrating a first-order component term and a higher-order contextual term is used for labeling components as "text" or "non-text".

For multilingual text localization, Liu *et al.* proposed a method which employs a GMM to fit third-order neighboring information of components using a specific training criterion: maximum minimum similarity (MMS). Their experiments show good performance on their multilingual image data sets.

III. PROPOSED SYSTEM

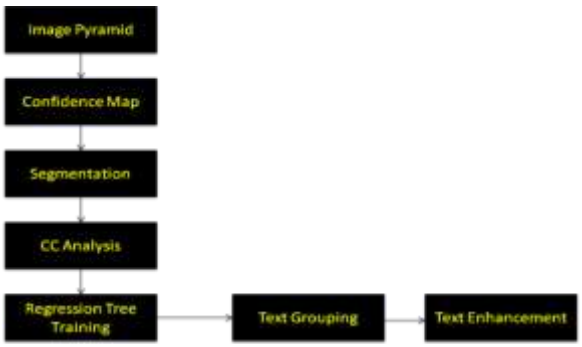

Fig. 2: Flow of Proposed Work

A. Image Pyramid

Image pyramid is a collection of decreasing resolution images arranged in the shape of a pyramid. And region text detector is designed to detect text in each layer by using the kernel function.

Fig. 3: Original image

Fig. 4: Image Pyramid

B. Confidence Map

By probability calibration, the text confidence map for each layer of image pyramid is obtained. Then, all the pixel confidence and scale values at different pyramid layers are projected back to the original image for calculating the final text confidence and scale maps.

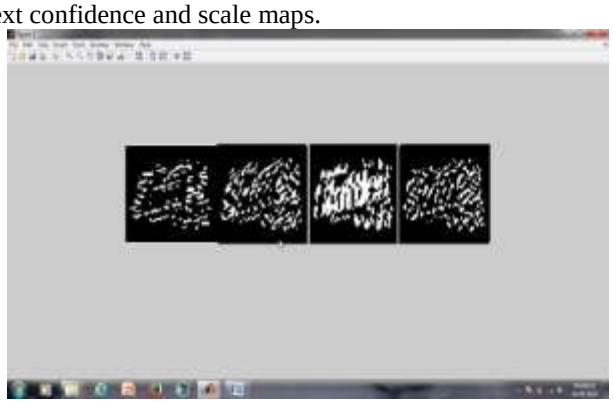

Fig. 5: Each resolution image resized to original image size

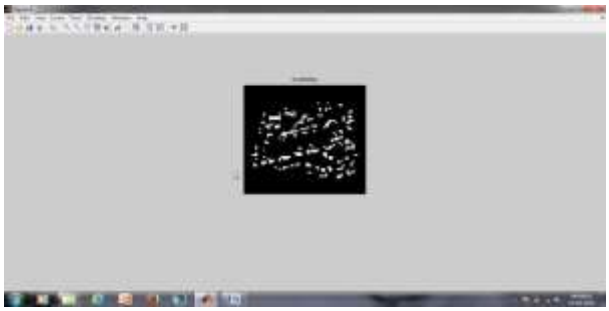

Fig. 6: Scale Map image

C. Segmentation

To segment the candidates, from the gray level image, a binarization is adopted due to its high efficiency and non-sensitivity to image degrading. To segment candidate CCs from the gray-level image, a Niblack's local binarization algorithm is adopted due to its high efficiency and non-sensitivity to image degrading. The formula to binaries each pixel is defined as [2]:

$$b(x) = \begin{cases} 0, & \text{if } gray(x) < \mu_r(x) - k \cdot \sigma_r(x); \\ 255, & \text{if } gray(x) > \mu_r(x) + k \cdot \sigma_r(x); \\ 100, & \text{otherwise.} \end{cases}$$

Where μ(x) and ϵ(x) are the intensity mean and standard deviation (STD) the pixels within a-radius window centered on the pixel and the smoothing term. And here we have used morphological dilation operation.

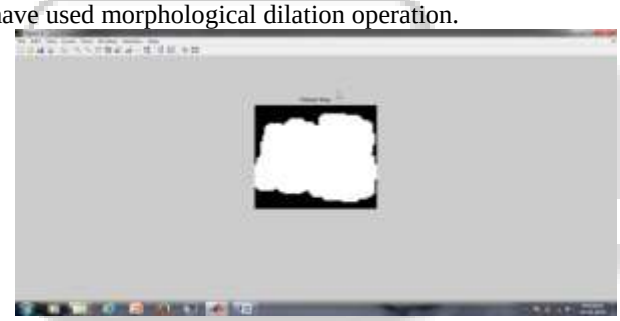

Fig. 7: Dilated Map image

Fig. 8: Outlined original image

D. Connected Component Analysis:

This module presents the connected component analysis (CCA) stage using a CRF model combining unary component properties and binary contextual component relationships is used to filter out non-text components. Conditional random field (CRF) model is proposed to assign candidate components as one of the two classes ("text" and "non-text") by considering both unary component properties and binary contextual component relationships. CRF is a probabilistic graphical model which has been widely used in many areas such as natural language processing. Next considering that neighboring text components normally have similar width or height, build up a component neighborhood graph by defining a component linkage rule. And also use the CRF model to explore contextual component relationships as well as unary component properties. During the test process, to alleviate the computational overhead of graph inference, some apparent non-text components are first removed by using thresholds on unary component features. The thresholds are set to safely accept almost all text components in the training set.

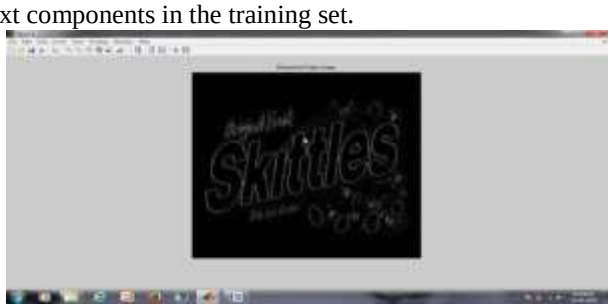

Fig. 9: Sharpened Edge image

E. Regression Tree Training

To group text components into text regions are lines and words, a learning-based method of clustering nearing components into a tree with a minimum spanning tree (MST) algorithm and cutting off between-line (word) edges with an energy minimization model is designed. Cluster text components in a tree with MST based on a learned distance metric, which is defined between two components as a linear combination of some features. With the initial component tree built with the MST algorithm, between-line/word edges need to be cut to partition the tree into sub Trees, Each of which corresponds to a text unit. Finally, text words corresponding to partitioned sub trees can be extracted and the ones containing two small components are removed as noises. With the initial component tree built with the MST algorithm, between-line/word edges need to be cut to partition the tree into sub trees, each of which corresponds to a text unit (line or word).

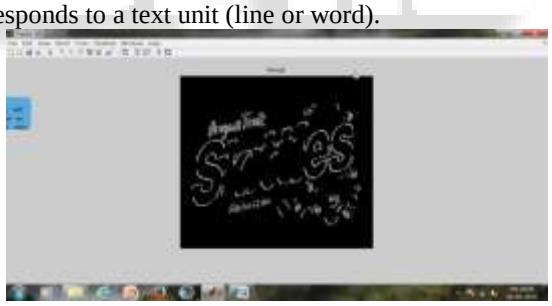

Fig. 10: Result

F. Text Grouping Method:

1) Text Line Partition:

A method to formulate the edge cutting into the tree is proposed as a learning-based energy minimization problem. In the component tree, each edge is assigned one of two labels: "linked" and "cut", and each sub tree corresponding to a text line are separated by cutting the "cut" edges. The objective of the proposed method is to find the optimal edge labels such that the total energy of the separated sub trees is minimal.

2) Text Word Partition:

For comparing our system with previous methods which reported word localization results, further partition text lines into words using a similar process as line petition. The major difference lies in the word-level features, which are defined as: 1) word number; 2) component centroid distances of cut edges; 3) component bounding box

distances of cut edges; 4) bounding box distances between words separated by cut edges; 5) the ratio between the component centroid distance of the cut edge and the average component centroid distance of the edges within separated words; and 6) bounding box distance ratio between the cut edge and edges within separated words.

3) Text Enhancement

Finally, text words corresponding to partitioned sub-trees can be extracted and the ones containing too small components are removed as noises. The bounding box is drawn to partition the text into lines and words in the original image itself.

Fig. 11: Text Grouping and Enhancement

IV. RESULT AND CONCLUSION

Text locating in natural scene image with complex background is a difficult, challenging, and important problem. In this paper, an accurate text region extraction algorithm based on two methods with gray-information is presented. The proposed methods work very well in a text region in natural images. Texts in natural scene images contain very important information about location information and road signs. And regression tree technique got very well result in this training process. And in this paper we have improved accuracy, speed, and efficiency very well.

Images	First image	Second image	Third image
Time	125s	104s	100s

Table 1: Result analysis of Time parameter

REFERENCES

[1] M. Swami Das, B. HimaBindhu, A. Govardhan[2012], "Evaluation of Text Detection and Localization Methods in Natural Images" [1].

[2] X. L. Chen, J. Yang, J. Zhang, and A.Waibel, "Automatic detection and recognition of signs from natural scenes," IEEE Trans. Image Process., vol. 13, no. 1, pp. 87–99, Jan. 2004 [2].

[3] S. L. Feng, R. Manmatha, and A. McCallum, "Exploring the use of conditional random field models and HMMs for historical handwritten document recognition," in Proc. Washington, DC, 2006, pp. 30–37 [3].

[4] S. M. Lucas, "ICDAR 2005 text locating competition results," in Proc.8th Int. Conf. Document Analysis and Recognition (ICDAR‟05), Seoul, South Korea, 2005, pp. 80–84 [4].

[5] Xin Zhang, Fuchun Sun, Lei Gu, "A Combined Algorithm for Video Text Extraction",2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010 [5].

[6] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1631–1639, 2003[6].

[7] A Hybrid Approach to Detect and Localize Texts in Natural Scene Images Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, Senior Member [7].

[8] Y.-F. Pan, X. W. Hou, and C.-L. Liu, "A robust system to detect and localize texts in natural scene images," in Proc. 8th IAPR Workshop on Document Analysis Syetems (DAS‟08), Nara, Japan, 2008, pp. 35–42[8].

[9] Y.F. Pan, X. Hou, C.L. Liu, Text Localization in Natural Scene Image base on Conditional Random Field International Conference on Document Analysis and Recognition, (2009)[9].

[10] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian approach to multi-oriented text detection in video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 412 –419, feb. 2011[10]

