Detecting and Localizing Color Text in Natural Scene Images Using Region Based & Connected Component Method

Mohanabharathi.R, [1] Surender.K, [2] Selvi.C[3]
[1, 2, 3] *Asst. Professor /Department of Computer Science and Engineering Selvam College of Technology, Namakkal.*

Abstract: *Large amounts of information are embedded in natural scenes which are often required to be automatically recognized and processed. This requires automatic detection, segmentation and recognition of visual text entities in natural scene images. In this paper, we present a hybrid approach to detect color texts in natural scene images. The approaches used in this project are region based and connected component based approach. A text region detector is designed to estimate the probabilities of text position and scale, which helps to segment candidate text components with an efficient local binarization algorithm. To combine unary component properties and binary contextual component relationships, a conditional random field (CRF) model with supervised parameter learning is proposed. Finally, text components are grouped into text lines/words with a learning-based energy minimization method. In our proposed system, a selective metric-based clustering is used to extract textual information in real-world images, thus enabling the processing of character segmentation into individual components to increase final recognition rates. This project is evaluated on natural scene image dataset.*

Keywords: Conditional random field (CRF); connected component analysis (CCA); text detection; text localization.

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 descriptors. This is the act of determining the properties of represented 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 multimedia document processing like automatic annotation and concept extraction out of images/audio/video could be seen as information extraction. [2] [3]

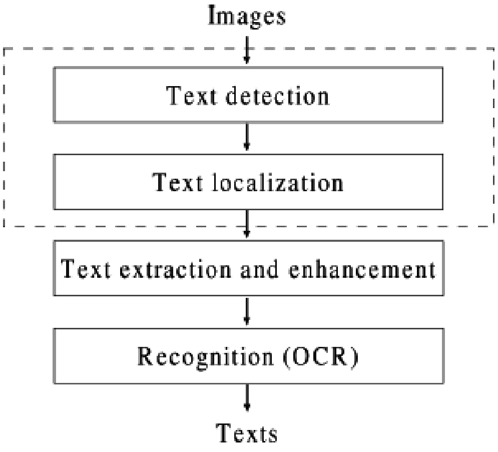


Fig.1.Text information extraction

Existing system presented a hybrid approach to robustly detect and localize texts in natural scene images by taking advantages of both region-based and CC-based methods. This system consists of three stages are the pre-processing stage, the connected component analysis stage and text grouping. At the pre-processing 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 with 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 reliable connected component analyzer is difficult since there are many non-text components which are easily confused with texts when analyzed individually.

This paper is organised as follows: Section II briefly reviews the related work. Section III describes the preprocessing of image. Section IV explains the connected component analysis using CRF model. Section V describes the text line/word grouping method. Clustering based method for text extraction and character segmentation is discussed in section VI. Experimental results and conclusion are presented in section VII.

II. RELATED WORKS

Most region-based methods are based on observations that text regions have distinct characteristics from 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.* [44] 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.* [16] proposed an algorithm for detecting texts in 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.* [14] use a CRF model for patch-based text detection. This method justifies the benefit of adding contextual information to traditional local region-based text detection methods. Their experimental results show that this method can deal with texts of variable 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 [15] 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 to 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.* [17] extracts candidate CCs based on edge contour features and removes non-text components by 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.* [19] 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 [18] 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 datasets.

III. PRE-PROCESSING

In this module, preprocessing stage of the overall process is discussed. 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. To extract and utilize local text region information, a text region detector is designed by integrating a widely used feature descriptor: Histogram of oriented gradients (HOG) and waldboost classifier to estimate the text confidence and the corresponding scale, based on which candidate text components can be segmented and analyzed accurately. Initially, the original color image is converted into a gray level image. To measure the text confidence for each image patch in a window, no matter it is accepted or rejected. [2] [3] [9] [10] The posterior probability of a label y_i, y_i ε{'text', 'non-text'}conditioned on its detection state , s_i, s_i ε{'accepted', 'rejected'} at the stage t, can be estimated based on the Bayes formula as defined as

$$\begin{aligned} P_t(y_i|s_i) &= \frac{P_t(s_i|y_i)P_t(y_i)}{\sum\limits_{y_i} P_t(s_i|y_i)P_t(y_i)} \\ &= \frac{P_t(s_i|y_i)P_{t-1}(y_i|\text{accepted})}{\sum\limits_{y_i} P_t(s_i|y_i)P_{t-1}(y_i|\text{accepted})}, \end{aligned}$$

Where all the stage likelihoods P_t (s_i/y_i)are calculated on a validation dataset during training. The text scale map is used in local binarization for adaptively segmenting candidate CCs and the confidence map is used later in CCA for component classification. The formula to binaries each pixel x is

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

Here $\mu_r(x)$ and $\sigma_r'(x)$ are mean and standard deviation with radius r. Figure 3 shows the example of preprocessing stage. They calculate the radius from the text scale map which is more stable under noisy conditions. After local binarization assume that within each local region, gray-level values of foreground pixels are higher or lower than the average intensity.

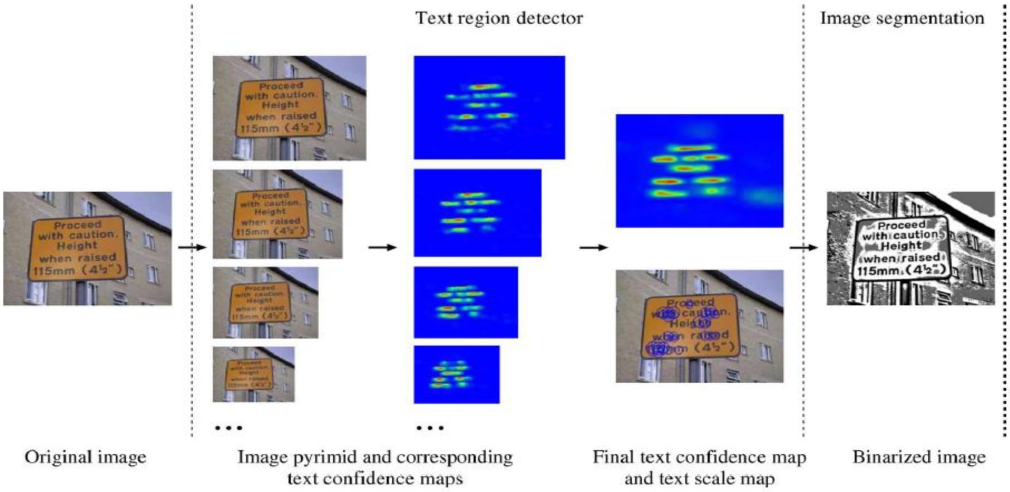

Text region detector
Image segmentation
Original image
Image pyrimid and corresponding text confidence maps
Final text confidence map and text scale map
Binarized image

Fig.3.Example of preprocessing stage

IV. 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) [4] [7] is proposed model 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 computation 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.

V. TEXT GROUPING METHOD

To group text components into text regions are lines and words, a learning-based method by 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 into 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 subtrees, each of which corresponds to a text unit. Finally, text words corresponding to partitioned subtrees can be extracted and the ones containing too 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 subtrees, each of which corresponds to a text unit (line or word).

A .Text Line Partition

A method to formulate the edge cutting in 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 subtree 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 subtrees is minimal. The total text line energy is defined as

$$E(L) = \sum_{i=1}^{N} W_{line} \cdot F_i ,$$

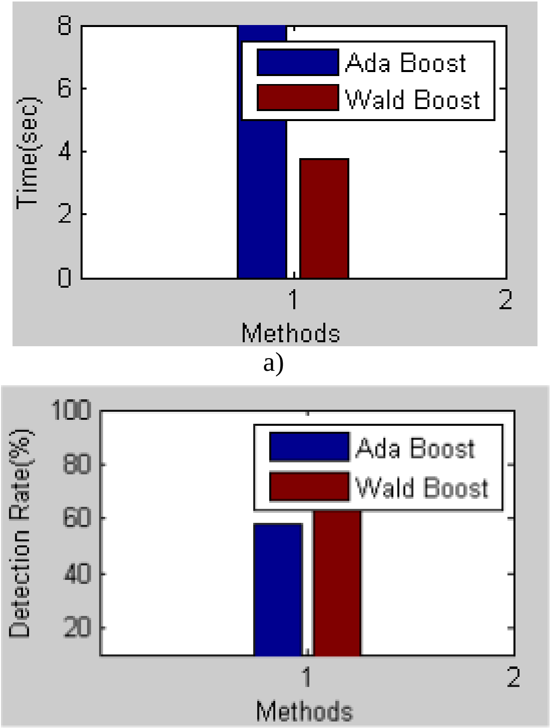
Where N is the number of subtrees (text lines), F_i is the feature vector of a text line, and W_{line} is the vector of combining weights.

B. 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 partition. 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.

VI. SELECTIVE METRIC-BASED CLUSTERING USING LOG–GABOR FILTERS

This module discuss about the selective metric based clustering using log-Gabor filter. Hence, our selective metric-based clustering is integrated into a dynamic method suitable for text extraction and character segmentation. This method uses several metrics to merge similar color together for an efficient text-driven segmentation in the RGB color space. However, color information by itself is not sufficient to solve all natural scene issues; hence complement it with intensity and spatial information obtained using Log–Gabor filters, thus enabling the processing of character segmentation into individual components to increase final recognition rates. Our selective metric-based clustering uses mainly color information for text extraction and our system fails for natural scene images having embossed characters. In this case, foreground and background have the same color presenting partial shadows around characters due to the relief but not enough to separate textual foreground from background in a discriminative way as displayed.



a)

b)

Fig.3. Comparison between adaboost and waldboost classifier a) Execution Time b) Detection Rate

Gray-level information with the simultaneous use of a priori information on characters could be a solution to handle these cases. Next we propose a new text validation measure M to find the most textual foreground cluster over the two remaining clusters. Based on properties of connected components of each cluster, spatial information is already added at this point to find the main textual cluster. The proposed validation measure, M, is based on the largest regularity of connected components of text compared to those of noise and background. And also we use Log–Gabor filters that present globally high responses to characters. Hence, in order to choose efficiently which clustering distance is better to handle text text extraction, we perform an average of pixel values inside each mask. The mask which has the highest average is chosen as the final segmentation.

VII. RESULT AND CONCLUSION

From the results its incurred that waldboost classifer has better execution time and detection rate of text, when compared with previously used adaboost classifier in preprocessing stage .Hence this classifier can be used for text recognition to be integrated with text localization for complete text information extraction.

REFERENCES

[1] D. T. Chen, J.-M. Odobez, and H. Bourlard, “Text detection and recognition in images and video frames,” *Pattern Recogn.*, vol. 37, no. 5, pp. 595–608, 2004.

[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.

[3] X. R. Chen and A. L. Yuille, “Detecting and reading text in natural scenes,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’04)*, Washington, DC, 2004, pp. 366–373.

[4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’05)*, San Diego, CA, 2005, pp. 886–893.

[5] 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.

[6] J. Gllavata, R. Ewerth, and B. Freisleben, “Text detection in images based on unsupervised classification of high-frequency wavelet coefficients,” in *Proc. 17th Int. Conf. Pattern Recognition (ICPR’04)*, Cambridge, U.K., 2004, pp. 425–428.

[7] J. M. Hammersley and P. Clifford, *Markov Field on Finite Graphs and Lattices*, 1971, unpublished. [10] X.-B. Jin, C.-L. Liu, and X. Hou, “Regularized margin-based conditional log-likelihood loss for prototype learning,” *Pattern Recogn.*, vol. 43, no. 7, pp. 2428–2438, 2010.

[8] Y. Jin and S. Geman, “Context and hierarchy in a probabilistic image model,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’06)*, New York, NY, 2006, pp. 2145–2152.

[9] B.-H. Juang and S. Katagiri, “Discriminative learning for minimum error classification,” *IEEE Trans. Signal Process.*, vol. 40, pp. 3043–3054, 1992.

[10] K. Jung, K. I. Kim, and A. K. Jain, “Text information extraction in images and video: A survey,” Pattern Recogn., vol. 37, no. 5, pp. 977–997, 2004.

[11] 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.

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

[13] V.Wu, R. Manmatha, and E. M. Riseman, “Finding text in images,” in Proc. 2nd ACM Int. Conf. Digital Libraries (DL’97), New York, NY, 1997, pp. 3–12.

[14] J. Weinman, A. Hanson, and A. McCallum, “Sign detection in natural images with conditional random fields,” in Proc. 14th IEEE Workshop on Machine Learning for Signal Processing (MLSP’04), São Luís, Brazil, 2004, pp. 549–558.

[15] 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.

[16] H. P. Li, D. Doermann, and O. Kia, “Automatic text detection and tracking in digital video,” IEEE Trans. Image Process., vol. 9, pp.147–156, Jan. 2000.

[17] Y. X. Liu, S. Goto, and T. Ikenaga, “A contour-based robust algorithm for text detection in color images,” IEICE Trans. Inf. Syst., vol. E89-D, no. 3, pp. 1221–1230, 2006.

[18] X. B. Liu, H. Fu, and Y. D. Jia, “Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images,” Pattern Recogn., vol. 41, no. 2, pp. 484–493, 2008.

[19] D.-Q. Zhang and S.-F. Chang, “Learning to detect scene text using a higher-order MRF with belief propagation,” in Proc. IEEE Conf. ComputerVision and Pattern Recognition Workshop s (CVPRW’04),Washington, DC, 2004, pp. 101–108.

[20] 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 onDocument Analysis Syetems (DAS’08), Nara, Japan, 2008, pp. 35–42.