Hybrid Approach to Extract Text in Natural Scene Images

Prabhu Kumar C.[1] Sanjana T. S.[2] Sneha D.[3] Soujanya B.[4] Supritha M. Y.[5]
[1]Assistant Professor
[1,2,3,4,5]Department of Computer Science & Engineering
[1,2,3,4,5]PESITM Shivamoga, India

Abstract— Text Extraction from natural scene images has been done with various methodologies. Most of the existing systems mainly use color and edges for detecting the text. We propose a two stage hybrid text extraction approach by combining texture and CC-based information. Text in the image is detected and localized using first and second order statistical texture features. In the next stage CC extraction is used to segment candidate text components from the localized text region. Finally morphological operations and heuristic filters are used to filter out non text components. Experimental results show that the proposed approach detects, localizes and extracts text from natural scene images efficiently and also can handle variations in size, fonts and orientation.

Key words: Natural Scene Images; Statistical Features; Text Localization; Text Extraction; Connected Component; Morphological Operation; Texture Analysis; Heuristic Filters

I. INTRODUCTION

The content-based image analysis techniques have been receiving more and more attention. Among all contents in image text has great interest as it contains important and useful information in our surroundings, e.g. street signs, name plates, advertisements. Automatic recognition and translation of these texts has many applications such as license plate localization [1], Scene text recognition and translation will be of great help for foreign travelers and visually impaired people. It can also be used in content-based image indexing [13], automatic forms reading [6].Some other applications are address block localization, identification of parts in industrial automation, robot navigation. However the segmentation and recognition of text from document images is quite successful, extraction of text from the natural scene images is a challenging task. Common problem for text extraction are variations in font style, size, orientation as well as the complex background. Moreover, camera based images can be subjected to numerous possible degradations such as blur, uneven lighting, low resolution, which makes it more difficult to recognize any text from the background noise.

There have been a number of surveys about text extraction in the literature [4, 8, 10, 21].The reported methods can be roughly categorized into three groups. The first category uses connected component analysis, which is based on observations that texts can be seen as a set of connected components. The second category is based on edges [14], which assume high contrast differences between the text and background.

The third category is based on textures [7] which are based on the observations that texts in images have distinct textural properties that can be used to discriminate them from background. Three important features of edges, Edge strength, Edge density and Edge orientation variance are used by Liu and Samara bandu [12]. A feature map is obtained using these three properties. Magnitude of the second derivative of intensity is used as a measurement of edge strength. Edge density is calculated based on the average edge strength within a window. By employing non-linear weight mapping the proposed method distinguishes text regions from texture-like regions, such as window frames, wall patterns etc. Laplacian operator is used for text detection by Phan et al. [16].Then pixels are classified into text and non-text clusters by K-means clustering algorithm. Finally false positives are eliminated using projection profile analysis and empirical rules.

Kwang in Kim et al. [5] proposed a method which uses the combination of SVM and CAMSHIFT. SVM is used for analyzing the textural properties of texts. Texture features are computed by the intensities of the raw pixels which are fed directly to the SVM. Text regions are identified by applying a CAMSHIFT algorithm to the results of the texture analysis. The proposed method encounters problems classifying very small text or text with a low contrast. Yi-Feng Pan et al. [20] designed a text region detector to detect text regions in each layer of the image pyramid. Then scale-adaptive local binarization is applied to generate candidate text components. A conditional random field (CRF) model is used for filtering non-text components. Finally text components are grouped into text lines/words with learning based energy minimization method. This method fails in localizing text in low intensity images.

Harr discrete wavelet transform and K-means clustering is used by Narasimha Murthy and Kumaraswamy. [15]. for extracting text from image. For more accurate classification of text and non-text area Morphological operations are included. Usually most of the algorithms take gray scale image as input, but in this approach RGB color image is used as input. The prominent edges are detected using Harr wavelet transform. Statistical features; mean, standard deviation and energy are estimated. Means clustering is used to partition a data set into cluster according to some defined distance measure of intensities. Then non-text regions are removed using morphological operations. The proposed method gives valid text localization result for the image with uniform background, but fails to locate the text when background is non-uniform. The proposed algorithm is sensitive to skew and direction of placement of text.

From the above analysis we can see that each system has its advantage and disadvantage. This motivates us to develop a system which is a combination of different approaches, so we get the advantage of each method. The rest of the paper is organized as follows: In section 2, We propose text localization based on texture features and text extraction method using cc-based technique are presented, Experimental results are presented in section 3,Conclusions and future work are discussed in section 4.

II. PROPOSED SYSTEM

Many earlier works [11, 20] have shown that a combination of different methods can improve the performance of text extraction algorithm. This motivates us to combine texture feature in combination with CC-based method. Proposed approach system architecture is shown in figure 1.

A. Text Localization

The aim of the text localization is to estimate probabilities of the text position in an image. Text in natural scene images often incorporated in numerous objects. In the pre-processing stage initially if the input image is color image, it is converted into gray scale. The color components may differ in a text region, while having an almost constant intensity. So, the intensity image Y is processed in the next steps of the algorithm rather than the color components R, G and B. In the next stage of pre-processing stage constant background is suppressed so that text regions can be detected accurately and with less computation. This is done by using high pass filter in the DCT domain [2].The DCT co-efficient's globally map the periodicity of an image and can be a quite efficient solution for constant background suppression.

The DCT co-efficient values are computed using equation (1).The DC component is the first entry in the DCT matrix. The constant background is removed by applying high pass filter given by equation (4). Then inverse DCT as defined in equation (5), is applied to obtain the background suppressed image. In the pre-processed image most of the unwanted details are removed, only gray level discontinuities belonging to the text and edges remain for further processing.

$$\beta_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A(m,n) \cos\frac{\pi(2m+1)p}{2M} \cdot \cos\frac{\pi(2n+1)q}{2N} \quad (1)$$

$0 \le m \le M-1$

$0 \le n \le N-1$

Where,

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le p \le M-1 \end{cases} \quad (2)$$

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{\frac{2}{n}}, & 1 \le q \le N-1 \end{cases} \quad (3)$$

$$H(u,v) = \begin{cases} 0, & (u,v) = (1,1) where\ u = 1,\dots,\dots,8\ and\ v = 1,\dots,8 \\ 1, & otherwise \end{cases} \quad (4)$$

$$A = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p \alpha_q \beta_{pq} \cos\frac{\pi(2m+1)p}{2M} \cdot \cos\frac{\pi(2n+1)q}{2N} \quad (5)$$

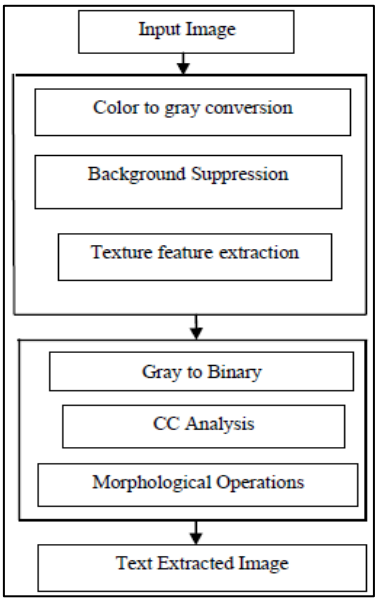

Fig. 1: Flow of the Proposed Method

Human vision can identify text of foreign languages without having to understand them or identifying individual characters because text has unique texture that differentiates it from the rest of the scene. In the proposed method two basic feature groups are used for extracting textural properties of the blocks. These features are calculated in the spatial domain and the statistical nature of texture is taken into account. The variance which is a measure of the spread of the values of intensities about the mean is calculated using equation (6), where A is the input image matrix; r and c are number of rows and columns. In order to approach higher classification accuracies it is necessary to consider the spatial dependence relationship of pixels with one another. This is achieved by computing a set of gray-tone spatial-dependence probability distribution also known as gray-level co-occurrence matrix (GLCM) for each image block. Contrast which is the measure of local variations in the gray-level is calculated from GLCM using equation (7).

$$\text{Variance} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} (A(i,j) - \mu)^2}{(r*c)-1} \quad (6)$$

$$\text{Contrast} = \sum_{i,j} |i-j|^2 p(i,j) \quad (7)$$

Discriminative functions checks whether the blocks feature vector satisfies the threshold value or not. Only those blocks which satisfy both discriminative functions are classified as text blocks. Text blocks connected in rows and columns are merged to obtain text regions. Four pairs of co-ordinates of the boundary blocks are determined by the maximum and minimum co-ordinates of the top, bottom, left and right points of the corresponding blocks. Finally the detected text blocks are merged together spatially and text regions are enclosed within a bounding box.

B. Text Extraction

In this stage from the localized text region, the text characters are separated from the background. Text extraction is critical and essential step as the efficiency of the OCR depends upon the accuracy of the text extraction system. Here we are using connected component approach for extracting characters from the localized text region. The localized text regions are converted into binary by applying suitable threshold value

using Otsu's technique. Since already text regions are located independent global threshold are used for conversion. In this step the grayscale images are converted into black and white images of distinguishing background and foreground objects. Texture based method localizes the text regions accurately by extracting texture information while text components are determined accurately using CC-based method. Connected component labeling is applied to the binary image which groups the pixels belonging to the same object, based on pixel connectivity. Here; the connected component algorithm proposed by Suzuki et al. [19] is employed. The algorithm scans the input binary image from top left corner. When encountered with a pixel value of 1,it checks the adjacent left, right, top and bottom pixels, if they are also of value '1' then its co-ordinates will be recorded and pixel value is set 0'.This procedure repeats in reverse direction. The recursive checking is continued until there are no more points with a pixel value of '1' around the recorded points, from which the first cluster of objects is identified. This process is repeated for the entire image. The generated connected components of the binary image may include some too large and/or too small connected components which are obviously not text. These non-character connected components are required to be detected and removed without removing any characters. Morphological filters and heuristic rules are used to remove false positives. Morphological operation opening is applied to remove objects that are too small or too big, and isolated regions are removed, making the assumption that characters are not alone. In the CCA the information of all the objects in the image, such as area, height, widths are given by a simple labeling process. These values do not require additional computation as they are already computed. Further heuristic filters are used to filter non-text CCs based on the height, weight and aspect ratio. The output of this stage is a binary image with only text characters.

III. Results & Discussion

In order to demonstrate the performance of the proposed method, we analyzed several types of images in which text has different font, size, color and orientation. Here we have presented evaluation on public data set: ICDAR 2011 robust reading competition data set and own image data set collected in indoor and outdoor conditions using digital camera and mobile phones. To evaluate the performance of the algorithm on other than English, we also collected images with text in regional language kannada. The resolution of these images ranged from 960X1280 to 2048X1536.All the images were in JPEG format. The proposed system has been implemented using MATLAB and operated on a 2.53GHz computer with windows-7 OS. We proposed a hybrid approach by combining texture-based and CC- based method. The combination of the two approaches takes the advantage of both approaches. Specifically, texture-based methods can extract texture information accurately to detect text regions while CC-based methods can filter out non-text components accurately.

Main idea is to keep good compromise between computation time and final result. Hence initially background suppression is performed before extracting texture features, which reduces the computation time and also increases the efficiency of texture feature extraction from each block. Most of the existing work on text localization using texture, extracts four to six texture features by second order statistics. These techniques gives better result than CC or edge based methods but computationally expensive. Finding threshold value for each texture feature is tedious. Hence here we propose a combination of first and second order statistics for texture feature extraction. We calculate only two texture feature, thus reduces the computational cost and at the same time provides results better than using only second order statistics.

Any priori information regarding the font size or format of the text in the image are not used. Figure 2 shows the examples of text localization results from both data sets. It can be seen from the results that most of the text blocks are well detected despite variations in font size and style. Our method works successfully in the cases where text is randomly oriented at different angles as shown in Figure 2c.

In the next stage the characters are extracted from the localized region using cc-based technique and morphological operations.CC-based method works accurately with prior information, as text is already detected in a more or less closing bounding box, text becomes the more relevant information in the new image. This increases the accuracy of CC-based method. Figure 3 shows some examples of text extraction results.

The performance of the text extraction results are efficiently evaluated by precision and recall rate. For a group of images precision rate and recall rat are computed for each image and then averaged over all images. Precision and recall rate are calculated using formulas proposed by Chitrakala gopalan. And Manjula, D [3] as given in (8) and (9) .Where CDP (Correctly Detected Pixels) is the number of pixels matching between output image and ground truth image. Table I shows the PR and RR for the test data base images. We can see that the PR and RR of the proposed algorithm are remarkably high; however, some false positives may be detected due to periodical pattern or structural features that look like characters such as tree, branches, and leaves in complex images.

$$\text{Precision Rate} = CDP/CDP+FP \quad (8)$$

$$\text{Recall Rate} = CDP/CDP+FN \quad (9)$$

Data-Set	Precision rate (%)	Recall rate (%)	Time
ICDAR	91	85	5-6 secs
Own-Dataset	87	78	10 secs

Table 1: Precision & Recall Rate of the Proposed Method for Different Data Set

IV. Conclusions

In this paper, a method to extract text from natural scene images based on combination of texture and CC-based technique is proposed. The system employed the texture features for localizing the text regions and characters are extracted using connected component and morphological operations. The whole scheme is successfully evaluated on ICDAR data base and in-house data base containing images with different text font, size, orientation and languages.

Fig. 2(a), Fig. 2(b), Fig. 2(c): Text Localization Examples (From Left To Right: Original Image, Results Of Background Suppression, Merging Of Text Blocks And Localized Text Region).

Fig. 3(a), Fig. 3(b), Fig. 3(c): Text Extraction Examples (From Left to Right: Text Region Localized Image, Text Extracted Image)

The results show that the proposed method gives high precession rate and recall rate for the heterogeneous images. Although the proposed system reported encouraging performance, it still needs further improvements. False positives still exists in some cases where the structural patterns in image matches with the text texture such as tree branches, leaves. Further code optimization of our system is needed for enhancing the speed.

REFERENCES

[1] Arth, C., Limberger, F. and Bischof, h. 'Real-time license plate recognition on an Embedded DSP-platform', IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'07), pp.1-8, 2007.

[2] Angadi, S.A. and Kodabagi, M.M. 'A Texture Based Methodology for Text Region Extraction from Low Resolution Natural Scene Images', International Journal of Image Processing, Vol. 3.Issue.5, 2010.

[3] Chitrakala gopalan. And Manjula, D. 'Statistical modeling for detection, localization and extraction of text from heterogeneous images using combined feature scheme', Springer-Verlag London, Vol 5, pp. 165-183, 2011.

[4] Jung, K., Kim, I. and Jain, A.K. 'Text Information extraction in images and video: A Survey', Pattern recognition, Vol. 37, no.5, pp. 977-997, 2004.

[5] Kwang in Kim. Keechul Jung. And Jin Hyung Kim. 'Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, 2003.

[6] Kavallieratou, E., Blcan, D., Popa, M. and Fakotakis, N. 'Handwritten text localization in skewed documents', International Conference on Image Processing, pp.1102-1105, 2001.

[7] Kim, K.J., Jung, K. and Kim, J.H.'Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm' IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 25,No. 12, pp.1631-1639,2003.

[8] Kumuda, T. and Basavaraj, L.'Text Extraction from Natural Scene Images using Region Based Methods-A Survey', in Proceedings of ACEEE International conference on Signal Processing and Image Processing,pp.412-416,2014.

[9] Kumuda, T. and Basavaraj, L.'Detection and Localization of Text from Natural scene images using Texture Features', 2015 IEEE International Conference on Computational Intelligence And Computing Research,pp. 739-742,2015.

[10] Liang, J., Doermann, D. and Li, H. P. 'Camera-based analysis of text and documents: A survey', Int. J. cument Anal. Recogn, Vol. 7, No.2-3, pp. 84–104, 2005.

[11] Liu, Y., Goto, S. and Ikenaga, T. 'A contour-based robust algorithm for text detection in color images', IEICE Transaction Information System, E89-D (3), pp 1221-1230, 2006.

[12] Liu, X. and Samarabandu, J. 'Multiscale edge-based text extraction from complex images', IEEE Int. Conf. Multimedia Expo, pp. 1721–1724.

[13] Lienhart, R. and Effelsberg, W. 'Automatic text segmentation and text recognition for video indexing', Multimedia system, pp 69-81, 2000.

[14] Lyu, M., Song, J. and Cai, M.'A Comprehensive method for multilingual video text detection, localization and extraction', IEEE transaction in circuits and systems for video technology, Vol. 15, No. 2, pp. 243-255,2005.

[15] Narasimha Murthy K N. and Kumaraswamy, Y, S. 'A novel method for efficient text extraction from real time images with diversified background using Haar Discrete Wavelet Transform and K-means clustering', IJCSI, Vol. 8, Issue 5, No. 3, 2011.

[16] Phan, T.Q., Shiva kumara, P. and Tan, C.L. 'A Laplacian method for video text detection', ICDAR 09', pp. 66–70, 2009.

[17] Robert M., Haralick. Shanmugam, K. and Its'hak Dinstein. 'Textural Features for Image Classification', IEEE Transactions on Systems, Man and Cybernetics, Vol. smc-3.No. 6, pp.610-621, 1973.

[18] Shehzad Muhammad Hanif. & Lionel Prevost. 'Texture based text detection in natural scene images: A help to blind and visually impaired persons', Conference &

Workshop on Assistive Technologies for People with vision & Hearing Impairments, 2007.

[19] Suzuki, K., Horiba, I. and Sugie N.'Linear-time connected-component labeling based on sequential local operations', Computer vision and image understanding, pp 1-23,2003.

[20] Yi-feng pan., Xinwen Hou. And Cheng-lin liu. 'A Hybrid Approach to Detect and Localize Texts in Natural scene images', IEEE transactions on image processing, Vol 20, No 3, 2011.

[21] Zhang, J. and Kasturi, R. 'Extraction of text objects in video documents: Recent progress', in Proc. 8th IAPR workshop on document analysis systems (DAS'08), Nara, Japan, pp. 1-13, 2008.

