Text Localization and Extraction in Natural Scene Images

Miss Nikita Aher
M.E. Student, MET BKC IOE,
University of Pune,
Nasik, India.
e-mail: nikitaaher@gmail.com

Abstract**—** Content based image analysis methods are receiving more attention in recent time due to increase in use of image capturing devices. Among all contents in images, text data finds wide applications such as license plate reading, mobile text recognition and so on. The Text Information Extraction (TIE) is a process of extraction of text from images which consists of four stages: detection, localization, extraction and recognition. Text detection and localization plays important role in system's performance. The existing methods for detection and localization, region based and connected component based, have some limitations due difference in size, style, orientation etc. To overcome the limitations, a hybrid approach is proposed to detect and localize text in natural scene images. This approach includes steps: pre-processing connected component analysis, text extraction.

Keywords*- CCA, CCA based method, Pre-processing, Region Based method, Text detection, Text localization, Text information extraction (TIE).*

*****

I. Introduction

With the increase in use of digital image capturing devices, content based image indexing process is receiving attention. Content based image indexing means the process of labeling the images based on their content. Image content can be divided in two categories: perceptual and semantic. Perceptual content covers the attributes such as color, intensity, texture, shape and their temporal changes, whereas semantic content means objects, events and their relation. Among all semantic content, text within an image is of interest because: 1) text is used for describing content of image, 2) text can be extracted easily and 3) text founds wide applications.

A TIE [2] system is a system used to extract text from images or videos. The input to TIE is a still image or sequence of images. There are five stages in TIE, which are: i) detection, ii) localization, iii) tracking, iv) extraction and enhancement, v) recognition. The flowchart given in figure 1 shows the flow of text information extraction system and also the relationship between the five stages:

Text detection is a method to determine the presence of text in given frame. Text localization determines the location of text in the image and generates bounding boxes around the text. Tracking is performed to reduce processing time for text localization. Extraction segments text components from background. Recognition of extracted text component is done using OCR. Among these stages, performance of detection and localization has effect on overall system's performance.

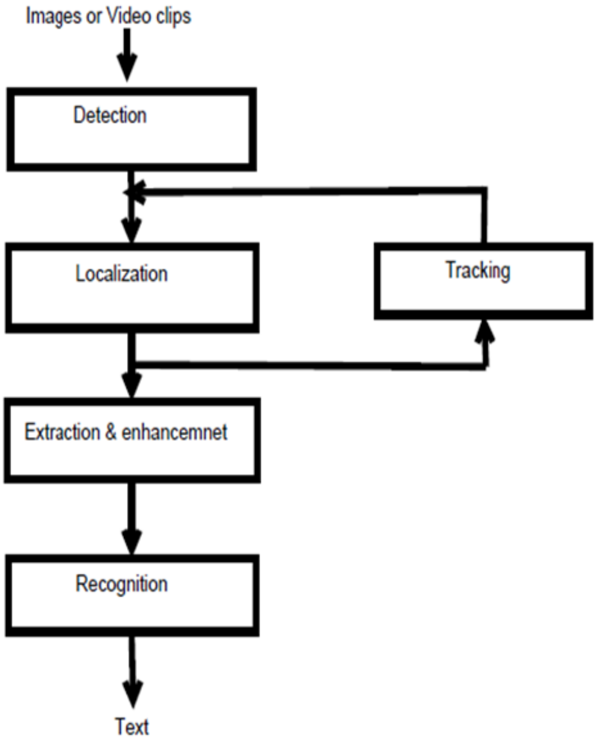

Fig 1.TIE

The existing methods of localization and detection can be roughly categorized into region based method and connected component based method. The region based method use the properties of color or gray scale in a text region or their difference with the corresponding properties of the background. On the other hand, the connected component based method directly segment candidate text components by edge detection or color clustering. Here the located text components can be directly used for recognition which reduces the computation cost.

The existing method faces several problems. For region based method, the performance is sensitive to text alignment, orientation and text density and also speed is relatively slow. On the other hand CC based method cannot segment text component accurately, also reliable CC analyzer is difficult to design since there are many non-text components which are easily confused as text components.

To overcome the above difficulties, the hybrid approach is proposed to robustly detect and localize texts in natural scene images by taking advantage of both regions based and CC based methods.

II. LITERATURE SURVEY

Region based consists of two stages: 1.Text detection: estimate text likelihoods. 2. Text localization: text region verification and merging.

An earlier method proposed by Wu et al. [3] is composed of following steps: 1) a texture segmentation method is used to find where the text is likely to occur. 2) Strokes are extracted from the segmented text regions. These strokes are processed to form rectangular bounding boxes around the corresponding text strings. 3) An algorithm which cleans up the background and binaries the detected text is applied to extract text from regions enclosed by the bounding boxes in input image. 4) The extracted text can be passed through OCR engine. This method uses a Gaussian derivative filters to extract texture features from local image regions.

The method of Zhong et al. [4] localizes captions in JPEG compressed images and MPEG compressed video. Segmentation is performed using distinguishing texture characteristics of text regions. The Discrete Cosine Transform (DCT) coefficients of images blocks are used to capture texture feature. It consists of two basic steps: first is detection of candidate caption regions in DCT compressed domain and second is the post processing to refine the potential text regions.

The method by Kim et al. [5] classifies pixel located at center into text or non-text by analyzing its textual properties using a Support Vector Machine (SVM) and then applies continuously adaptive mean shift algorithm (CAMSHIFT) to results of texture classification to obtain text chip. The combination of CAMSHIFT and SVM produces efficient results for variable scales text.

Whereas connected component (CC) based methods have steps: 1. CC extraction, 2. CC analysis and 3. CC grouping.

The method of Liu et al. [6] extracts candidate CCs based on edge contour features and removes non-text components by wavelet feature analysis. The flow of algorithm is: 1) robust edge detection, 2) Candidate text region generation, 3) Text region verification. 4) Binarizing text region using Expectation maximization algorithm.

Zhu et al. [7] uses Non-linear Niblack (NLNibalck) method to decompose gray image into candidate CCs. Then every candidate CC is fed into a series of classifiers and each classifier will test one feature of this CC. If one of the CC is rejected by any of classifier then it is considered as a non-text CC and need no further judgment. Finally, the CCs passing through all classifiers will be processed by a post processing procedure and form the final segmentation result.

Zhang et al. [8] used a Markov random field (MRF) method for exploring the neighboring information of components. The mean shift process is used for labeling candidate text components as text or non-text using component adjacency graph.

III. SYSTEM OVERVIEW AND WORKING FLOW

The proposed system consists of 3 stages as shown in figure 2. At pre-processing stage a text region detector is designed to detect text regions. Then in connected component analysis non-text component are filter out by formulating component analysis into component labelling problem. At final stage the text extraction is done using OCR which gives final output as text.

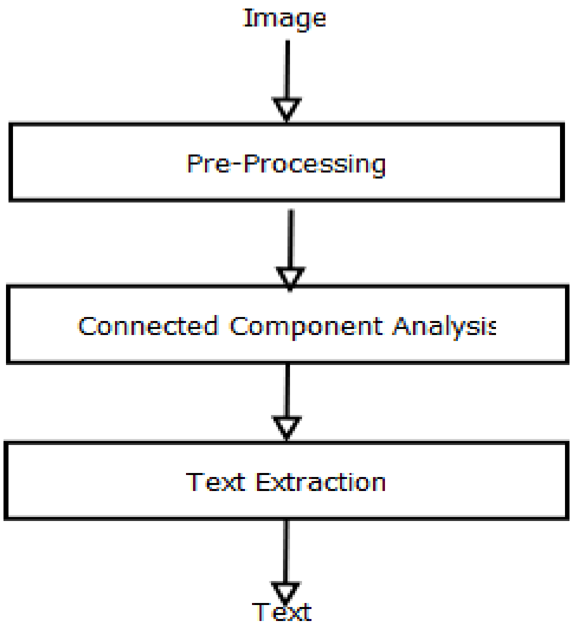

Fig 2.Flowchart of Proposed System

A. *Pre-Processing*

A text region detector is designed using feature descriptor: Histogram of Oriented Gradients (HOG) [9] and a Wald boost [10] classifier to calculate text confidence and scale value based on which segmentation of candidate text is done. The output of Wald boost classifier is translated to posterior probability. The posterior probability of a label x_i, x_i € {'text', 'non-text'} conditioned on its detection stage d_i, d_i € {'accepted', 'rejected'} at the stage t, is calculated based on bayes formula as defined as:

$$P_t(x_i / d_i) = \frac{P_t(d_i / x_i)P_t(x_i)}{\sum_{x_i} P_t(d_i / x_i)P_t(x_i)}$$
$$= \frac{P_t(d_i / x_i)P_{t-1}(x_i / accepted)}{\sum_{x_i} P_t(d_i / x_i)P_{t-1}(x_i / accepted)} \quad (1)$$

The text scale map obtained is used in local binarization for segmenting candidate CCs and confidence map is used in CCA for component classification. Basically in this step first image is converted to gray image and is then converted to binarized image, on which waldboost algorithm is applied.

B. *Connected Component Analysis*

The filters are used to label components as text or non-text. The filters perform labelling of objects in image obtained after applying waldboost algorithm. It colours each separate object using different colour. The image processing filter treat all non black pixels as object's pixels and all black pixel as background. Initially the image is divided into blocks and these blocks are filtered.

Then the text extraction is performed to obtain the text from image.

C. *Text Extraction*

The text extraction is done using OCR, that is, Optical Character Recognition technique. OCR extracts text and layout information from document image. OCR is integrated using MODI, that is, Microsoft Office Document Imaging library which is contained in Office 2003. The input given to OCR is output of connected components algorithm.

IV. Mathematical Modelling and Implementation Strategy

A. *Luminance Method*

The proposed method called Luminance [18] is used to convert the colour image provided as input into gray image. This method work on each pixel. In this method the colour property of pixel is taken as input. The colour property of pixel has three values namely, red, green and blue. The gray value ranges from 0 to 255. The gray value is obtained by using formula:

$$Gray = (\mathrm{Re}\, d \times 0.2125 + Green \times 0.7154 + Blue \times 0.0721) \quad (2)$$

The coefficient used for red, green and blue are as per ITU-R recommendation .

B. *Histogram of Oriented Gradients*

The method of Histograms of Oriented Gradients (HOG) [9] is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. This is implemented by dividing the image window into small regions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation. The steps for HOG are:

i. Compute gradients.
ii. Weighted into spatial and orientation cell.
iii. Contrast normalizes over overlapping spatial blocks.
iv. Collect HOG's over detection window.

C. *Waldboost Classifiers*

The Wald Boost [10] classifier executes the SPRT test using a trained strong H_t with a sequence of thresholds classifier $\theta_A^{(t)}$ and $\theta_B^{(t)}$. If H_t exceeds the respective threshold, a decision is made. Otherwise, the next weak classifier is taken. If a decision is not made within T cycles, the input is classified by thresholding H_T on value γ specified by user. Also the strong classifier H_T and thresholds $\theta_A^{(t)}$ and $\theta_B^{(t)}$ are calculated in learning process. The algorithm for Wald Boost classification is:

```
1. Given: $h^{(t)}, \theta_A^{(t)}, \theta_B^{(t)}, \gamma$.
2. Output: a classified object $x$.
3. For $t = 1....T$.
4. If $H_t(x) \geq \theta_B^{(t)}$, then classify $x$ to the class
   +1 and terminate.
5.If $H_t(x) \leq \theta_A^{(t)}$, then classify $x$ to the class
   -1 and terminate.
6. end.
7. If $H_T(x) \geq \gamma$, then classify $x$ as +1,
   otherwise classify as -1 and terminate.
```

V. Implementation Platform

Microsoft's implementation of the C# [15] specification is included in Microsoft Visual Studio suite of products. It is based on the ECMA/ISO specification of the C language, which Microsoft also created. Microsoft's implementation of C# uses the library provided by .NET framework. The .NET framework is a software development kit that helps to write applications like windows applications, web application or web services. The .NET framework 3.5 is used for proposed system. The main C# namespace used for programming is drawing which contains all classes needed to draw on form such as Bitmap, Brush, Font, Graphics, Image, etc.

VI. RESULTS GENERATED

A. Data Set

The input to the system will be an image. The image will be in any format such as BMP (Bit Map Format), GIF (Graphic Interchange Format), PNG (Portable Network Graphics), JPEG (Joint Photographic Experts Group), etc. The given image will be first converted to BMP image and then further processing will be done. The standard data set used is ICDAR 2011

B. Result Set

The output obtained after applying the BT709 i.e. luminance method is known as gray as shown in figure 4. Then after applying filter on gray image, the binarized image is obtained as shown in figure 5. Then the waldboost algorithm is applied and resulting image is as shown in figure 6. Then the connected component algorithm is applied which colors the object with different colors as shown in figure 7. Finally the OCR is applied which extracts the text from image and stores in text file.

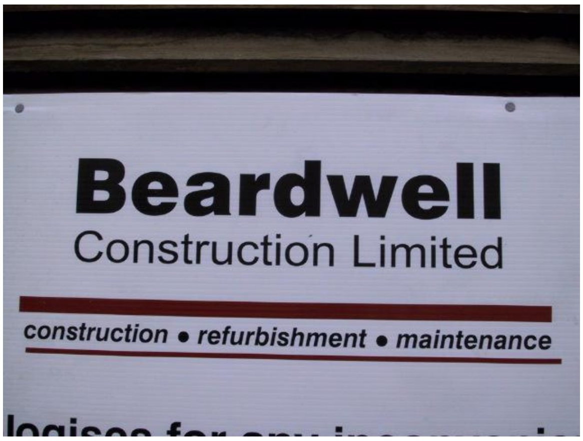
Fig 3.Original Image

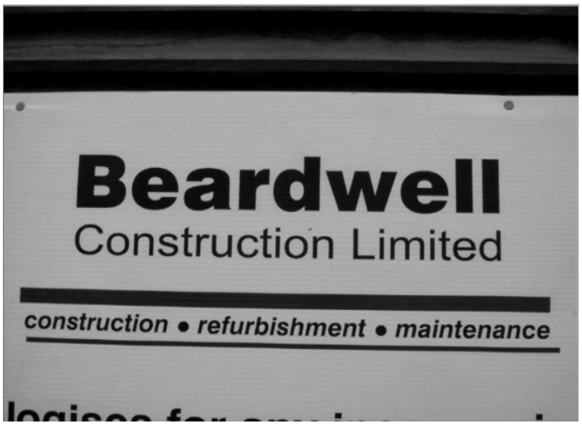
Fig 4.Gray Image

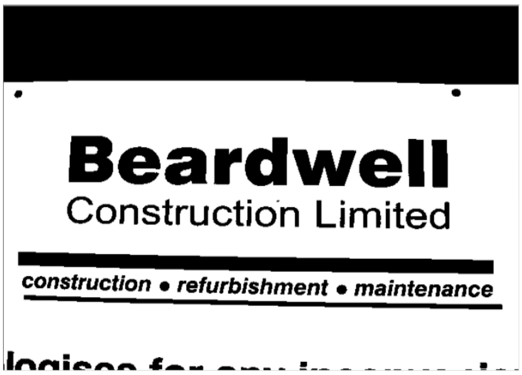
Fig 5.Binarized Image

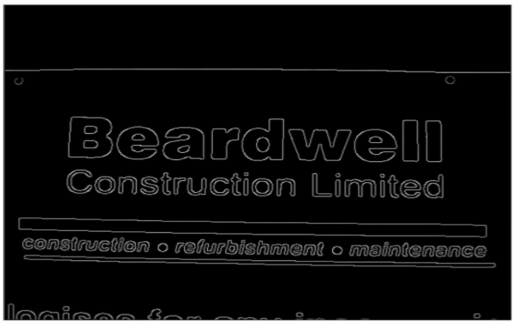
Fig 6.Output of Waldboost

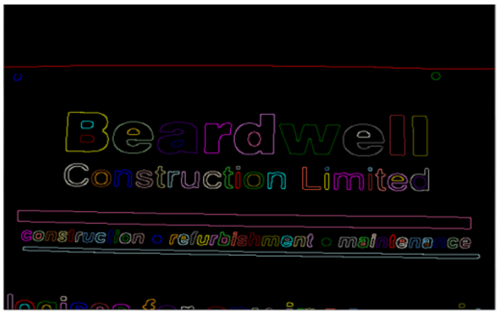
Fig 7.Output of connected component

C. Result Analysis

The below table shows the percentage of characters that were extracted from image. The results are obtained on ICDAR 2011 dataset.

Table 1 Number of Characters Extracted in Percentage

Sr No.	Percentage of characters extracted (%)
1.	77.7
2.	91.8
3.	66.6
4.	100

VII. CONCLUSION

Thus the proposed system consists of 3 main steps: pre-processing connected component analysis and text extraction. The output of pre-processing is a binarized image. Then in CCA, the labeling of component is done as text or non-text. Then finally in text extraction the text components are extracted.

REFERENCES

[1] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images", in *IEEE transactions on Image Processing*, Vol. 20, No. 3, March 2011.

[2] 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. 77–997, 2004.

[3] 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.

[4] Y. Zhong, H. J. Zhang, and A. K. Jain, "Automatic caption localization in compressed video", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, No. 4, pp. 385–392, 2000.

[5] 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] 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.

[7] K. H. Zhu, F. H. Qi, R. J. Jiang, L. Xu, M. Kimachi, Y. Wu, and T. Aizawa, "Using Adaboost to detect and segment characters from natural scenes", in *Proc. 1st Conf. Caramera Based Document Analysis and Recognition (CBDAR'05)*, Seoul, South Korea, 2005, pp. 52–59.

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

[9] 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.

[10] J. Sochman and J. Matas, "WaldBoost – Learning for time constrained sequential detection", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, 2005, pp. 150–156.

[11] A. Turtschi, J. Werry, G. Hack, J. Albahari, "C# .NET Web Developers Guide" by Syngress Publication.

[12] Anil K. Jain, "Fundamentals of Digital Image Processing", Published by Prentice Hall of India Private Limited.

[13] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing", Pearson Publication, Printed by India Binding House, 2009.

Miss. Nikita B. Aher is post graduate student of Computer Engineering at MET Bhujbal Knowledge City, Nasik under University of Pune. She have completed her BE from same college.