Detect and Localize Text in Natural Image using Wavelet Transformation

Maunil Patel[1], Aniruddhsinh Dodiya[2]

[1]Information Technology Department, L. D. College of Engineering
Gujarat Technological University, Ahmedabad Gujarat India

[2]Information Technology Department, L. D. College of Engineering
Gujarat Technological University, Ahmedabad Gujarat India

Abstract: ***With increasing use of capturing device in recent year. Content based image serves as important clues for many image based application. Scene text contains significant and beneficial information. Detection and localization of scene text is used in many applications. In this paper, a new approach is developed which locate text in different backgrounds. However, variation of text due to differences in size, style, orientation and alignment, as well as low image contrast and complex background make the problem of automatic text localization extremely challenging. The text localization algorithm system is designed to locate text in different kinds of images. Firstly, the color image is converted into grayscale image, and apply segmentation algorithm for segment gray scale image. After that, Haar Wavelet Transform (DWT) is employed. Haar DWT decompose image into four sub image coefficients, one is average and other three are detail. Which is find the approximately text area in the given image (confident region). After that apply binary and unary properties of conditional random field (CRF) on confident region image for connected component labeling finally, using some specific condition, the text is obtained in bounding box.***

Keywords: Wavelet Transform, text detection, text localization.

1. Introduction

Texts in images provide highly condensed information about the contents of the images. Although texts provide important information about images, it is not an easy problem to detect and segment them. Texts localization is not easy for the following reasons. First of all, text sizes may change from small too big and text fonts may vary in a wide range as well. Secondly, texts present in an image may have multiple colors and appear in a much cluttered background. The existing methods of text detection and localization can be roughly categorized into two groups: region-based and connected component (CC)-based. Region-based methods attempt to detect and localize text regions by texture analysis. It is find estimate text region in image and then neighboring text regions are merged to generate text blocks. Because text regions have distinct textural properties from non-text ones, these methods can detect and localize texts accurately even when images are noisy.

CC-based methods directly segment candidate text components by edge detection. The non-text components are then pruned with heuristic rules or classifiers. Since the number of segmented candidate components is relatively small, CC-based methods have lower computation cost and the located text components can be directly used for recognition. 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. Difficult to designing fast and reliable connected component analyzer, since there are many non-text components which are easily confused with texts.

To overcome the above difficulties, we present detect and localize texts in natural scene images with wavelet transform. We design a text region detector to estimate the probabilities of text position, which help segment candidate text components with an efficient local binarization algorithm. For connected component labeling we design combination of unary and binary properties of the conditional random field which connect text each other. Finally, text components are localizing with use of bounding box. Figure 1 show the flowchart of proposed system.

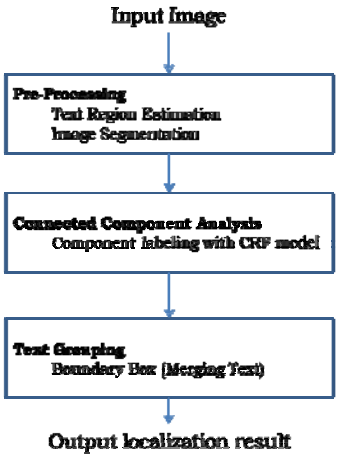

Figure 1: Flow Chart of Proposed Model

2. Related Work

Many efforts have been made for text extraction and recognition in image. Chung-Wei Liang and Po- Yueh Chen in their paper DWT Based Text Localization presents an efficient and simple method to extract text regions from static images or video sequences. They implemented Haar Discrete Wavelet Transform (DWT) with morphological operator to detect edges of candidate text regions for isolation of text data from the documented video image.

A Video Text Detection and Recognition System presented by Jie Xi, Xian-Sheng Hua, Xiang-Rong Chen, Liu Wenyin, and Hong-Jiang Zhang proposed a new system for text information extraction from news videos. They developed a method for text detection and text tracking to locate text areas in the key-frames.

Xian-Sheng Hua, Pei Yin, Hong-Jiang Zhang in their paper efficient video text recognition using multiple frame integration presented efficient scheme to deal with multiple frames that contain the same text to get clear word from isolated frames.
C´elineThillou and Bernard Gosselin proposed a thresholding method for degraded documents acquired from a low-resolution camera. They use the technique based on wavelet denoising and global thresholding for non uniform illumination. In their paper Segmentation-based binarization for color-degraded images they described the stroke analysis and character segmentation for text segmentation. They proposed the binarization method to improve character segmentation and recognition.

S. Antani and D. Crandall in their paper Robust Extraction of Text in Video describes an update to the prototype system for detection, localization and extraction of text from documented video images. Rainer Lienhart and Frank Stuber presented an algorithm for automatic character segmentation for motion pictures in their paper 'Automatic text recognition in digital videos', which extract automatically and reliably the text in pre-title sequences, credit titles, and closing sequences with title and credits. The algorithm uses a typical characteristic of text in videos in order to enhance segmentation and recognition
Zhong et al. used a CC-based method, which uses color reduction. They quantize the color space using the peaks in a color histogram in the RGB color space. This is based on the assumption that the text regions cluster together in this color space and occupy a significant portion of an image. Each text component goes through a filtering stage using a number of heuristics, such as area, diameter, and spatial alignment. The performance of this system was evaluated using CD images and book cover images.

Kim segments an image using color clustering in a color histogram in the RGB space. Non-text components, such as long horizontal lines and image boundaries, are eliminated. Then, horizontal text lines and text segments are extracted based on an iterative projection profile analysis. Kim et al. used cluster-based templates for filtering out non-character components for multi-segment characters to alleviate the difficulty in defining heuristics for filtering out non-text components.

3. Pre Processing

To extract and utilize local text region information, a text region detector is designed to estimate the text confidence and the corresponding scale, based on which candidate text components can be segmented and analyzed accurately.

3.1 Image Segmentation

To segment candidate CCs from the gray-level image, the segmentation evaluation is always difficult as it is, for a part, subjective. Most of time, it is impossible to have a ground truth to be used with a representative measure. We define as clearly as possible what *properly segmented* means: the character must be readable; it must not be split or linked with other features around it. The thickness may vary a little provided that its shape remains correct. Figure 2 shown the example of image segmentation.

- Niblack's local binarization algorithm is adopted due to its high efficiency and non-sensitivity to image degrading. The formula to binarize each pixel is defined as

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

Where $\mu_r(x)$ and $\mu_r(y)$ are the intensity mean and standard deviation (STD) of the pixels within a -radius window centered on the pixel x and the smoothing term k is empirically set to 0.4. Different from other methods, where the radius of windows is fixed or chosen based on some simple rules, such as the gray-level STD, we calculate the radius from the text scale map which is more stable under noisy conditions. After local binarization, because we assume that within each local region, gray-level values of foreground pixels are higher or lower than the average intensity, connected components with 0 or 255 value are extracted as candidate text components and those of value 100 are not considered further. Figure 2 show the resultant image of image segmentation.

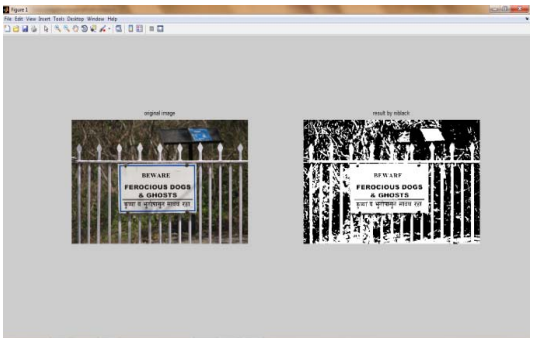

Figure 2: (a) original image (b) Segmented image

3.2 Text Region Detector and confidence map

If the input image is a gray-level image, the image is processed directly starting at discrete wavelet transform. If the input image is colored, then its RGB components are combined to give an intensity image. Usually, color images are normally captured by the digital cameras. The pictures are often in the Red-Green-Blue color space. Intensity image Y is given by:
Y = 0.299R + 0.587G + 0.114B

Image Y is then processed with 2-d discrete wavelet transform. The Y is actually Value component of the Hue-Saturation-Value (HSV) color space. The RGB color image and its grayscale image. In this step, there is conversion from RGB color space into HSV color space, after that Value component is extracted from HSV color space using above expression. The noise of the image is reduced by using a median filtering that is applied on the above

grayscale image. After this filtering step, a great part of noise will be removed while the edges in the image are still preserved.

3.2.1 Haar discrete wavelet transforms

We are using Haar discrete wavelet transform which provides a powerful tool for modeling the characteristics of textured images. Most textured images are well characterized by their contained edges. It can decompose signal into different components in the frequency domain. We are using 2-d DWT in which it decomposes input image into four components or sub-bands, one average component(LL) and three detail components(LL, HL, HH). The detail component sub-bands are used to detect candidate text edges in the original image. Using Haar wavelet, the illumination components are transformed to the wavelet domain. This stage results in the four LL, HL, LH and HH sub image coefficients. The traditional edge detection filters can provide the similar result as well but it cannot detect three kinds of edges at a time. Therefore, processing time of the traditional edge detection filters is slower than 2-d DWT. The reason we choose Haar DWT because it is simpler than that of any other wavelets. Figure 3 show the approximate text area in image.

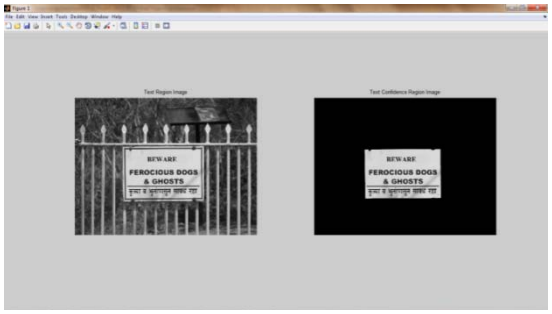

Figure 3: (a) original image (b) text region detector and confidence map

4. Connected Component Analysis

In connected component analysis (CCA) stage, a CRF model combining unary component properties and binary contextual component relationships is used to filter out non-text components. Here, we present a conditional random field (CRF) 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.

4.1 Brief introduction of CRF

Conditional random fields (CRFs) are discriminative graph models which are designed for labeling tasks such as text identification and document image segmentation. The motivation to use CRFs to label the text region from video frames arises from the spatial inter-dependencies of different areas in images. For example, text blocks are sequential from left to right. By considering neighboring information of blocks, isolated noises among text blocks can be easily removed which leads to more satisfactory labeling results. Unlike other generative graphical models such as Markov random fields (MRFs) which require specifying the likelihood function; CRFs have a more flexible formulation. More formally, let $X = \{xi\}$ be the observed features from candidate blocks, and $Y = \{yi\}$ be random variables over corresponding labels. The joint distribution over the label yi given an observation xi has the form:

$$p(y_i \mid x_i) \propto \exp\left(\lambda A(y_i, X) + \mu \sum_{(i,j)\in E} I(y_i, y_j, X)\right) \quad (1)$$

where function $A(\cdot)$ is called associated potential which measures the confidence of label yi with observations, function $I(\cdot)$ is interaction potential which tends to smooth labels over entire graph G, λ and μ are parameters that control the influence from observations and neighboring nodes to center node i, and (i, j) €E means neighboring nodes of node I that are connected by edges E in the graph G.

In our work, we use the topology for our CRFs. By making a Markov assumption, each gray node yi in the hidden layer exclusively corresponds to a detected block in the image and connects to its four nearest neighbors and their corresponding observations. In real images, the neighbor system of blocks is determined by Euclidean distance between them but may not necessarily be located as a grid. To integrate the predicted confidence of blocks into CRFs framework, we define the associated potential as:

$$A(y_i, X) = \sum_{j\in N} p_j \exp\left(-|d_{i,j} \cdot \cos(\theta_{i,j})|\right) \quad (2)$$

Where j runs over neighbors of node i including itself, pj is the posterior estimated by the SVM for node j, di, j is the spatial Euclidean distance between node i, j, and θ is the angle between centers of node i and j. The idea behind equation 2 is that if two neighboring nodes are close to each other and their separation is mostly horizontal, they have more influences on each other.

4.2 Properties of CRF

A. Unary Component Features

To characterize single component's geometric and textural properties, we use different types of unary component features such as Normalized width and height, Aspect ratio, Compactness.

B. Binary Component Features

To characterize the spatial relationship and geometric and textural similarity between two neighboring component and, we use different types of binary component features such as Shape difference, Overlap ratio, Scale ratio, Gray-level difference.

5. Text Grouping

Text grouping is adjacent letters in order to form words. This task is one of the most difficult of text (word) extraction. In order to analyze the performance of a text extraction algorithm it is commonly recommended to

compute the precision and recall rates. The problem is that these performance parameters are so dependent on correctly classified words. There is several works which try to solve this issue. While the method proposed in is effective but too complicated because of training data necessity, the method proposed in is simpler but not effective. To merge adjacent letters in words we propose to use the following process which is based on the computation of distances between bounding boxes (BB) of letters detected in the previous step. The parameters used in this merging letters process are illustrated in Fig. 5. (B1, B2) represents the coordinates of the center of the two BBs of connected component.

"B1 (y1a)" and "B1 (y2a)" (respectively, "B2 (y1a)" and "B2 (y2a)") represent the coordinates of the first BB (of the second BB) in vertical direction, "Width1" and "Width2" represent the width of the two BBs studied. "Distance" represents the distance between the centroids of the two BBs considered along the horizontal direction.

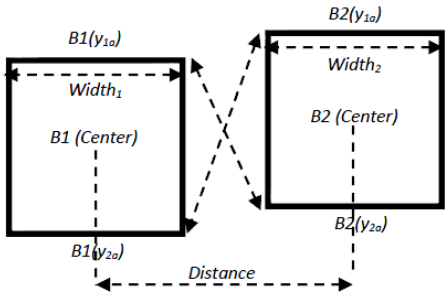

Figure 4: Parameters used in merging process

This first step of merging is based on a merging of letters along almost horizontal line. Here we have limited our study to text images whose letters are relatively well aligned. The conditions for merging letters in detected regions are defined as below:

- [B2 (y2a) >B1 (y1a)] & [B2 (y1a) <B1 (y2a)]
- [Distance< 0.7 × Max (Width1, Width2)]

Any pair of BB which supplies both above conditions is then merged in this step. Figure 6 show the bounding box (rectangle box) on each text in image.

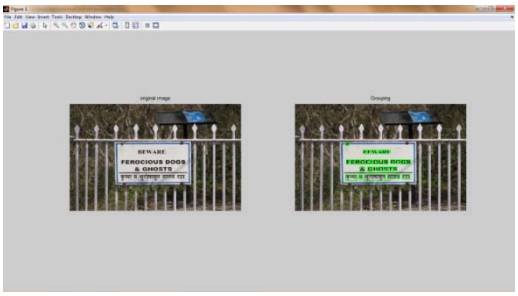

Figure 5: (a) Original image (b) localize text area in image

6. Conclusion

The given images to be convert to gray scale image and find segmented image from the gray scale image. Segmented image is decomposed by using Wavelet Transform. It will decompose the original image into four frequency sub bands for improve the contrast and resolution of the image and find approximation text area in the image with help of connected component graph. Then with the use of bounding box localize text in image. After that detect and localize texts by integrating region information into a robust CC-based method.

References

[1] H. Demirel and G. Anbarjafari, Image Resolution Enhancement by Using Discrete and Stationary Wavelet Decomposition, IEEE Trans. on Image Processing, Vol.20, May-2011,No. 4, pp 1458-1460 .

[2] 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.

[3] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 256–268, 2002.

[4] 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.

[5] YuZhong, Hongjiang Zhang, and Anil K. Jain, "Automatic Caption Localization in Compressed Video", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, (4) (2000) 385-392.

[6] C.M. Lee, and A. Kankanhalli, Automatic Extraction of Characters in Complex Images, International Journal of Pattern Recognition Artificial Intelligence, 9 (1) (1995) 67-82.

[7] 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.

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

[9] Yu Zhong, KalleKaru, and Anil K. Jain, Locating Text In Complex Color Images, Pattern Recognition, 28 (10) (1995) 1523-1535.

[10] 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.

[11] Y.-F. Pan, X. W. Hou, and C.-L.Liu, "Text localization in natural scene images based on conditional random field," in *Proc. 10th Int. Conf. Document Analysis and Recognition (ICDAR'09)*, Barcelona, Spain, 2009, pp. 6–10.

[12] S. Shetty, H. Srinivasan, M. Beal, and S. Srihari, "Segmentation and labelling of documents using conditional random fields," in *Proc. Document Recognition and Retrieval XIV, Proc. SPIE*, San Jose, CA, Jan. 2007, pp. 6500U-1–11.

[13] 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.

[14] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data," in *Proc.*

18th Int. Conf. Machine Learning (ICML'01), San Francisco, CA, 2001, pp. 282–289.

[15] Chung-Wei Liang and Po-Yueh Chen, "DWT based Text Localization", International Journal of Applied Science and Engineering:2004