Text Extraction from Complex Natural Images

Manoj Kumar, GueeSang Lee
Department of Computer Science, Chonnam National University

ABSTRACT

The rapid growth in communication technology has led to the development of effective ways of sharing ideas and information in the form of speech and images. Understanding this information has become an important research issue and drawn the attention of many researchers. Text in a digital image contains much important information regarding the scene. Detecting and extracting this text is a difficult task and has many challenging issues. The main challenges in extracting text from natural scene images are the variation in the font size, alignment of text, font colors, illumination changes, and reflections in the images. In this paper, we propose a connected component based method to automatically detect the text region in natural images. Since text regions in mages contain mostly repetitions of vertical strokes, we try to find a pattern of closely packed vertical edges. Once the group of edges is found, the neighboring vertical edges are connected to each other. Connected regions whose geometric features lie outside of the valid specifications are considered as outliers and eliminated. The proposed method is more effective than the existing methods for slanted or curved characters. The experimental results are given for the validation of our approach.

Keywords*: text detection, natural image, character segmentation*

1. INTRODUCTION

Digital cameras mounted on various handheld devices have become very popular. The manufacturers of these devices are seeking to embed various useful technologies into such devices. The recognition of text in natural scenes is one of these important technologies. Text information in any kind of digital image is helpful to analyze the image in detail and provide a better understanding of the information present in it. Automated processes of understanding the text from images have many computational challenges. The extraction of text from images is complicated due to various factors, such as the variation of the light intensity, alignment of text, color, font size, and sometimes due to the camera angle. Text information from natural images provides valuable information, such as the name of the place, sign board information, e.g. road signs. This information can be extracted and used to search for additional facts. As mentioned above, finding the text area in natural images is a difficult task and is hindered by divers issues, such as dull lighting, reflections, the shape and size of the text, graphical images in between the text, brand logos, font style, noise in the background and out of focus images. In addition, the text size, color and orientation remain unpredictable. Many researchers have attempted to find solutions to these kinds of problems based on many different detection and segmentation methods. Extraction methods can be divided into two categories, region based and texture based. The common texture based method is robust, but the high complexity of the texture is the main problem when processing complex images. The color based method involves finding a uniform background and is helpful in investigating text candidate regions in a uniform color background [8]. Lim [2] computes the features of edge information and identifies the text region based on filters and block information. The drawback in the color based approach is that it cannot detect text from an image if the graphics and text have a similar color and shape. Clustering based approaches have been implemented [4], but they are more sensitive to noise and color variations.[1]The proposed histogram based analysis is suitable only if text is present in a straight line, whereas if the text is tilted or arranged in curves, it is quite difficult to locate it in full view. Edge based histogram analysis to locate license plate is proposed by [11]. Texture based method using frequency information like DCT is illustrated in [9]. Techniques based on the stroke filter [6] provide a fast and effective method of finding the text area by detecting strokes such as structures present in the images. However, there is a high probability of considering logo and other graphics images as stroke like structures and detecting the wrong areas as text. The edge based [4] method of text extraction makes use of connected components and morphological operations. As edge detection is more robust for text and complex backgrounds, it gives a poor detection rate, as mentioned in the experimental results section.

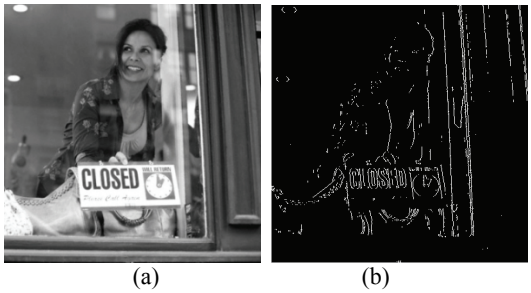
In this paper, we explore the edge based method and solve the problem of highly complex backgrounds using a modified pixel grouping method along with a pre-processing step. The general assumption for identifying text characters in the image is that they are arranged closely together in the horizontal direction and have high contrast against the background with a continuous intensity change along the horizontal line. This makes the text region more dominant and project out from complex backgrounds. To obtain this high contrast region, we first applied a global binarization method. To get the unique characteristics of the continuous vertical strokes of text, we sought inspiration from the edge based text detection method in [4] to extract all possible vertical edges from the input image. After extracting these vertical edges, we used morphological operations [4] to connect all of the vertical edges that are

* *Corresponding author. E-mail : gslee@jnu.ac.kr*
Manuscript received Sep. 01, 2009 ; accepted Apr. 15, 2010

arranged closely in an acceptable manner. Once all the edges are connected, we eliminate false positives based on the geometrical properties. We tested our algorithm with several complex natural scene images and the proposed method was found to outperform the method in [4] and give high positive rates for text arranged in a curved shape and with titled angles.
The rest of the paper is organized as follows. Section 2 explains the pre-processing steps. In Section 3, we introduce the edge detection process and section 4 illustrates the edge grouping method. Searching for the text region is explained in section 5. The experimental results are discussed in section 6. Finally, the conclusion and a discussion of future work are given in section 7.

2. PRE-PROCESSING

Texts in natural scene images typically have a complex background. Figure 1(a) shows a gray scale natural scene image with text embedded in a signboard surrounded by different objects. The vertical edge map generated from the gray scale image in figure 1 (a) is shown in figure 1(b). The edge map shows all of the possible edges in the image according to the intensity variation along the vertical lines. A close observation along the text line reveals the difficulties of the region grouping method proposed in [4].



(a) (b)

Fig.1. (a) Gray Scale Image, (b) Corresponding Edge Image.

In figure 1 (b), the gaps between edges along the text line are lesser then Maximum Interval parameter proposed by [4]. So there is a high chance of grouping text characters along with the long vertical edge of the signboard and glass window. As a result, an unexpected grouping will occur and make the process text detection process more complicated. The presence of vertical edges in natural scene images is unpredictable, so we propose a different method in this paper. Initially, the input RGB image is converted to a gray scale image using the following equation, as shown in figure 2(b).

0.2989 * R + 0.5870 * G + 0.1140 * B

Fig.2. (a) Original color image (b) Corresponding Gray image.

Where R, G, B are the 3 color components of red, green and blue, of the color image. Many techniques have been developed to binarize gray images.
The simplest way to binarize a gray scale image is by choosing the best threshold value and classifying all of the pixels with values above this threshold as white and the others as black. In our experiment, we use a global binarization method to determine the global threshold. There are many types of global binarization methods, of which Otsu's method is the most popular [12].

By using the histogram analysis of the image with less illumination changes, Otsu's method tries to find a single threshold value for the whole image and binarize the image effectively, thus functioning rapidly and giving a good result, as shown in figure 4(a). The binarization creates more connected components in the background region and some in the text and non text regions.

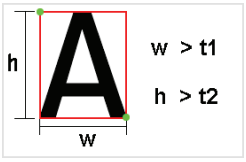


Fig. 3. (a) Connected component of binary image with Bounding Box.

Using the properties of the region, the height (H) and width (W) of each connected component are analyzed with the help of the bounding box defined by the coordinates of the CC, as shown in figure 3. As mentioned above, the backgrounds will have a larger number of grouped pixels forming a maximum height and width. Therefore, taking this into account, we set the threshold values, T1 and T2, so as to eliminate some of the non text regions from the binary image. Figure 4(b) shows the resulting image after applying geometric rules to the binary image.

(a) (b)

Fig. 4. (a) Binarization Result. (b) Candidate text region.

3. EDGE DETECTION PROCESS

Even though the previous process eliminates some of the not text regions whose height and width exceed the threshold value, there could still be some non text regions similar to text with the same shape and size as text characters. Morphological operations such as erosion and dilation are frequently used to distinguish text from noise. However, choosing the proper structuring element (SE) required to use this type of operation is very difficult and, if the size of the structuring element is smaller than that of the text characters, there is a possibility of disconnecting the text strokes. Therefore, we apply a different

method of distinguishing the text region from the noise using the edge information. Edge detection plays a vital role in the text detection process and helps to locate the text region precisely, even if the text has high illumination changes and intensity variations. As text characters in any image have a high contract against the background and are arranged close to each other in a horizontal direction, they have unique characteristics. To extract these unique characteristics, we apply Sobel Vertical Edge filter. This detects vertical strokes along the horizontal line of the text region, as shown in figure 5.

Fig.5. Result of Vertical edge detection.

Even though figure 5 shows the text region with more closely packed shorter and longer edges, selecting these regions is not easy. To make the search processes easy and robust, we analyze the edges and apply an edge grouping method to connect all of these edges together and form a connected region.

4. EDGE GROUPING

Even though vertical edge detection shows the prominent features of the text region, it is still difficult to locate the text region using this information. The main assumption regarding text regions is that they are arranged in the horizontal direction with even gaps between each pair of characters. Using this assumption as a key feature, we search for the vertical edges and group these edges to form a group of connected components. This creates a strip of connected components of text characters arranged in a single row. As mentioned above, there are different kinds of morphological operations, but choosing the right structural element is a difficult task. Therefore, in order to group these edges together, we apply a different method, as illustrated in figure 6, based on some predefined set of rules.

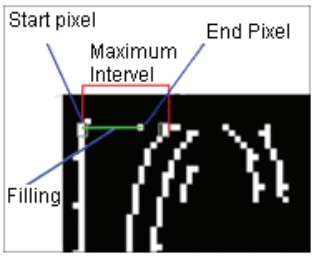


Fig.6 illustrates the process of edge grouping.

The proposed edge grouping method operates as follows. From the input image, two sets of pixels, the start and end pixels, are selected between the predefined Maximum Interval (M) and the gaps between these two sets of pixels are filled with intensity value 255 (white pixels). To group every pixel in the image, the image is scanned from left to right and top to bottom. Any pixel in the input image is marked as the Start Pixel (SP) and the index is shifted to maximum interval as shown in red line in figure 6. If a pixel is found between start pixel and Maximum Interval it is marked as the End Pixel (EP). Now the pixels in-between these two pixels (SP) and (EP) are given intensity value (255). This draws a line and forms a connection between the start and end pixels, as shown in figure 6 in green color. The same process is repeated for all pixels in the image and forms a group of edges, as shown in figure 7.

Fig.7. Result after grouping edges.

In most natural scenes, the text characters are arranged close to each other with an even gap between each pair of characters. Even if there is a variation in the font size, the vertical edge detection process finds all possible edges from the beginning to the end of the character. This creates a minimal gap between the edges of one character and the next one. Therefore, a constant maximum interval of 25 pixels helps to find the pixels of an adjacent character even if the font size varies. This maximum interval limits the constraint placed on the font size and orientation and connects all of the text strokes.

5. FINDING THE TEXT REGION

After the grouping process, there could be many connected components which contain text and non text regions and the process of searching for text in these connected components is tedious. In order to make our search simpler, we make the assumption that the texts are arranged along a horizontal line with a unique distance between each pair of characters and form a strip with constant height. Taking this information into account, we use the geometric properties of each connected component to find the text region. The image is scanned from top to bottom and the height and width of each connected component is measured using a bounding box. A connected component will be considered as a text region if it satisfies the following rules.

1. Its width and height should not be too small or too large.

2. The width of each connected component should be greater than 4 times its height.

As images have different font sizes, the threshold for the second rule is set dynamically. We use the character height as a factor to calculate the threshold dynamically. The above rules enable some of the minor noise to be eliminated and the text region to be recovered from the grouped image. As shown in figure 8 texts from complex natural scene with different orientation are extracted successfully.

(a) (b)

Fig.8. (a) Detected text candidates region. (b) Final output.

6. EXPERIMENT RESULTS

We propose a different method using global binarization and edge grouping to detect the text in natural scene images. This method was tested with different kinds of images, including sign boards, advertisements, name tags, name plates, book images, etc. The precision and recall of 70 images, which were calculated in terms of the number of words, are 0.9 and 0.89, respectively. The precision is defined as the number of correct estimates divided by the total number of estimates. The estimates refer to the detection result found by the proposed method. The recall is defined as the number of correct estimates divided by the total number of targets in the ground truth. The targets refer to the detections which are supposed to be included in the final result. However, the proposed method is limited to some standard font sizes, while images with a very large font size or small font size of less than 12 pixels are not taken into account in this method. In this experiment, we encountered some difficulties in detecting characters that are more widely spaced and images with high illumination changes. These two kinds of images give a higher false positive rate. Figure 9 shows some experimental results for images from our database.

7. CONCLUSIONS AND FUTURE WORD

The proposed method gives a high success rate for complex natural scene images. However, this method could still be extended to images with high illumination changes. Therefore, in the future, we plan to modify our approach and focus on solving the problems encountered during this experiment.

(a) (b) (a) (b)

Fig.9. (a) Original Image (b) Experimental results

ACKNOWLEDGEMENT

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2010-C1090-1011-0008).

REFERENCES

[1] T.N. Dinh, J.H. Park and G.S. Lee, "Low-Complexity Text Extraction in Korean Signboards for Mobile Applications," *Proc. IEEE International Conference on Computer and Information Technology*, 2008, pp. 333-337.

[2] P. Shivakumara, W. Huang and C.L. Tan, "Efficient Video Text Detection using Edge Feature," *Proc. International conference Pattern Recognition*, 2008, pp.8-11.

[3] Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang and S. Tang, "A Novel Image Text Extraction Method Based on K-Means Clustering," *Proc. International conference on Information System,* 2008, pp.185-190.

[4] P. Dubey, "Edge Based Text Detection for Multi-purpose Application," *Proc. International Conference on Signal Processing*, 2006, pp.16-20.

[5] C. Li, X.Q.Ding and Y.S.Wu , "Automatic text location in natural scene Images," *Proc. International conference of Document Analysis and Recognition*, 2001, pp.1069-1073.

[6] X. Li, W. Wang, S. Jiang, Q. Huang and W. Gao, "Fast and effective text detection," *IEEE International Conference on Image Processing*, 2008, pp.969-972.

[7] Q. Liu, C. Jung, S.K. Kim, Y.S. Moon and J.Y. Kim, "Stroke Filter for Text Localization in Video Images," *IEEE International Conference on Image Processing*, 2006, pp.1473-1476.

[8] S.A.R. Jafri, M.Boutin and E.J. Delp, "Automatic text area segmentation in natural images," *Proc. IEEE International Conference on Imaging Processing*, 2008, pp.3196-3199.

[9] A.C. Rodríguez, J.H. Kim, S.H. Kim and Y.B. Fernández,

"English to Spanish Translation of Signboard Images from Mobile Phone Camera," *submitted to IEEE Transactions on PAMI*,2009.

[10] X. Liu and J.Samarabandu, "Multiscale Edge-Based Text Extraction from Complex Images," *Proc. International Conference of Multimedia and Expo,* 2006, pp.1721-1724.

[11] F. Faradji, A.H. Rezaie, and M. Ziaratban, "A Morphological-Based License Plate Location," *IEEE International Conference on Image Processing*, 2007, pp.57-60.

[12] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Systems Man and Cybernetics*, *vol.09, Jan*. 1979, pp.62-66.

Manoj Kumar

He received the B.S. degree in Computer Science from Kongunadu Arts &Science College, Coimbatore, Tamilnadu, India, in April 2001.

Currently, he is pursuing M.S degree in Computer Science at Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and mobile applications.

Gueesang Lee

He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.