Text Information Extraction And Analysis From Images Using Digital Image Processing Techniques

Partha Sarathi Giri
Department of Electronics and Communication, M.E.M.S, Balasore, Odisha

Abstract – Text data present in images contain useful information for automatic explanation, indexing, and structuring of images. Extraction of this information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image. However variations 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 extraction extremely challenging in the computer vision research area. The proposed methods compare two basic approaches for extracting text region in images: edge-based and connected-component based. The algorithms are implemented and evaluated using a set of images that vary along the dimensions of lighting, scale and orientation. Accuracy, precision and recall rates for each approach are analyzed to determine the success and limitations of each approach.

Keywords: Text information extraction; Text detection; Text localization

I. INTRODUCTION

Text Extraction from image is concerned with extracting the relevant text data from a collection of images. Recent studies in the field of image processing show a great amount of interest in content retrieval from images and videos. This content can be in the form of objects, color, texture, shape as well as the relationships between them. The semantic information provided by an image can be useful for content based image retrieval, as well as for indexing and classification purposes [4, 10]. As stated by Jung, Kim and Jain in [4], text data is particularly interesting, because text can be used to easily and clearly describe the contents of an image. Since the text data can be embedded in an image or video in different font styles, sizes, orientations, colors, and against a complex background, the problem of extracting the candidate text region becomes a challenging one [4]. Current Optical Character Recognition (OCR) techniques can only handle text against a plain monochrome background and cannot extract text from a complex or textured background [7]. As stated text have some common distinctive characteristics in terms of frequency and orientation information, and also spatial cohesion. Spatial cohesion refers to the fact that text characters of the same string appear close to each other and are of similar height, orientation and spacing [7]. Two of the main methods commonly used to determine spatial cohesion are based on edge [1, 2] and connected component [3] features of text characters. Among them, text within an image is of particular interest as (i) it is very useful for describing the contents of an image; (ii) it can be easily extracted compared to other semantic contents, and (iii) it enables applications such as keyword-based image search, automatic video logging, and text-based image indexing.

1.1. Text in images

Texts usually have different appearance changes like font, size, style, orientation, alignment, texture, color, contrast, and background. All the changes will make the problem of text extraction become complicated and difficult. Text in images can exhibit many variations with respect to the following properties:

Size: Although the text size can vary a lot, assumptions can be made depending on the application domain.

Alignment: The characters in the caption text appear in clusters and usually lie horizontally, although sometimes they can appear as non-planar texts as a result of special effects. This does not apply to scene text, which can have numerous perspective distortions. Scene text can be aligned in any direction and can have geometric distortions.
Inter-character distance: Characters in a text line have a uniform distance between them.

Color: The characters in a text line are subject to have the same or similar colors. This property makes it possible to use a connected component-based approach for text detection. Many of the research informed till date has focused on finding ‘text strings of a single color’. However, video images and other complex color documents can contain ‘text strings with more than two colors (polychrome)’ for effective visualization, i.e., different colors within one word.

Motion: The same characters usually exist in consecutive frames in a video with or without movement. This property is used in text tracking and enhancement. Scene text can have arbitrary motion due to camera or object movement.

Edge: Most caption and scene texts are designed to be easily read, thereby resulting in strong edges at the boundaries of text and background.

Compression: Many digital images are recorded, transferred, and processed in a compressed format. Thus, a faster TIE system can be achieved if one can extract text without decompression.

II. TEXT INFORMATION EXTRACTION (TIE)

A TIE system receives an input in the form of a still image or a sequence of images. The images can be in gray scale or color, compressed or uncompressed, and the text in the images may move or may not. The problem arises due to TIE system can be divided into the following sub-problems: (i) detection (ii) localization (iii) tracking (iv) extraction and enhancement (v) recognition (OCR) shown in Fig. 1. Text detection refers to the determination of the presence of text in a given sequence of images. Text localization is the process of determining the location of text in the image and generating bounding boxes around the text. Text tracking is performed to reduce the processing time for text localization and to maintain the integrity of position across adjacent frames. Although the precise location of text in an image can be indicated by bounding boxes, the text needs to be segmented from the background to facilitate its recognition. That means, the extracted text image has to be converted into a binary image and enhanced before it is fed into an OCR engine. Text extraction is the stage where the text components are segmented from the background. Text Enhancement of the extracted text components is required because the text region usually has low resolution and is horizontal to noise. Thereafter, the extracted text images can be transformed into plain text using OCR technology.

Fig. 1. Architecture of a TIE system.

2.1 Text extraction techniques

To implement, test, and compare and contrast two approaches for text region extraction in images, and to discover how the algorithms perform under variations of lighting, orientation, and scale transformations of the text. The algorithms are from Liu and Samarabandu in and Gillavata, Ewerth and Freisleben in. The comparison is based on the accuracy of the results obtained, and precision and recall rates.

2.1.1 Edge based text region extraction

Edge-based methods focus on the 'high contrast between the text and the background'. The edges of the text boundary are identified and merged, and then used to filter out the non-text regions. Usually, an edge filter is used for the edge detection, and a smoothing operation or a morphological operator is used for the merging stage. The basic steps of the edge-based text extraction algorithm are explained, and diagrammed in Figure 2.1. (1) Create a Gaussian pyramid by convolving the input image with a Gaussian kernel and successively down-sample each direction by half. (2) Create directional kernels to detect edges at 0, 45, 90 and 135 orientations. (3) Convolve each image in the Gaussian pyramid with each orientation filter. (4) Combine the results of step 3 to create the Feature Map. (5) Dilate the resultant image using a sufficiently large structuring element to cluster candidate text regions together. (6) Create final output image with text in white pixels against a plain black background.

The procedure for extracting a text region from an image can be broadly classified into three basic steps: detection of the text region in the image, localization of
the region, and creating the extracted output character image.

![Block diagram for edge based text extraction](image)

Figure 2: Block diagram for edge-based text extraction

### 2.1.1.1 Detection

Among the several textual properties in an image, edge-based methods focus on the "high contrast between the text and the background". The edges of the text boundary are identified and merged, and then several methods are used to filter out the non-text regions. In this section, the region with the possibility of text for a given image is detected. A Gaussian pyramid is created by successively filtering the input image with a Gaussian kernel of size 3x3 and down sampling the image in each direction by half. Down sampling refers to the process whereby an image is resized to a lower resolution from its original resolution. A Gaussian filter of size 3x3 will be used. Each level in the pyramid corresponds to the input image at a different resolution. These images are then convolved with directional filters at different orientation kernels for edge detection in the horizontal (0°), vertical (90°), and diagonal (45°, 135°) directions.

![Sample Gaussian pyramid](image)

Figure 2.2 (a) Sample Gaussian pyramid with 4 levels (b) Each resolution image resized to original image size (c) Sample resized image of the pyramid after convolution with 0° kernel

After convolving the image with the orientation kernels, a feature map is created. A weighting factor is associated with each pixel to classify it as a candidate or non-candidate for text region. A pixel is a candidate for text if it is highlighted in all of the edge maps created by the directional filters. Thus, the feature map is a combination of all edge maps at different scales and orientations with the highest weighted pixels present in the resultant map.

### 2.1.1.2 Localization

The process of localization involves further enhancing the text regions by eliminating non-text regions. One of the properties of text is that usually all characters appear close to each other in the image, thus forming a cluster. By using a morphological dilation operation, these possible text pixels can be clustered together, eliminating pixels that are far from the candidate text regions. Dilation is an operation which expands or enhances the region of interest, using a structural element of the required shape and size. The process of dilation is carried out using a very large structuring element in order to enhance the regions which lie close to each other. In this algorithm, a structuring element of size [2x2] has been used. Figure 3.3 below shows the result before and after dilation of an original image. The resultant image after dilation may consist of some non-text regions or noise which needs to be eliminated.

![Dilation results](image)

Figure 2.3. (a) Original Image (b) Before dilation (c) After dilation

### 2.1.1.3 Character extraction

The common OCR systems available require the input image to be such that the characters can be easily parsed and recognized. The text and background should be monochrome and background-to-text contrast should be high.

![Character extraction results](image)

Figure 2.4. (a) Original Image (b) Result

### 2.1.2 Connected Component based text region extraction

### 2.1.3 Connected Component based methods use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image. A geometrical analysis is needed to merge the text components using the spatial arrangement of the components so as to alter out non-text components and mark the boundaries of the text regions. The basic steps of the connected component
Based methods text extraction algorithm are explained, and diagrammed in Figure 2.5. (1) Convert the input image to YUV color space. The luminance(Y) value is used for further processing. The output is a gray image. (2) Convert the gray image to an edge image. (3) Compute the horizontal and vertical projection profiles of candidate text regions using a histogram with an appropriate threshold value. (4) Use geometric properties of text such as width to height ratio of characters to eliminate possible non-text regions. (5) Binarize the edge image enhancing only the text regions against a plain black background. (6) Create the Gap Image using the gap-filling process and use this as a reference to further eliminate non-text regions from the output.

Figure 2.5. Basic Block diagram for Connected Component based text extraction

2.1.3.1 Pre-Processing

The input image is pre-processed to make easy detection of text regions. As proposed in the image is converted to the YUV color space (luminance + chrominance), and only the luminance(Y) channel is used for further processing. The Y channel refers to brightness or intensity of the image whereas the U and the V channels refer to the actual color information. Since text present in an image has more contrast with its background, by using only the Y channel, the image can be converted to a grayscale image with only the brightness contrast information present.

2.1.3.2 Detection

In this process, the connected-component based approach is used to make possible text regions stand out as compared to non-text regions. Every pixel in the edge image is assigned a weight with respect to its neighbors in each direction. This weight value is the maximum value between the pixel and its neighbors in the left (L), upper (U) and upper-right (UR) directions. The algorithm uses these three neighbor values to detect edges in horizontal, vertical and diagonal directions. The resultant edge image obtained is sharpened in order to increase contrast between the detected edges and its background, making it easier to extract text regions. Figure 2.6(c) shows the sharpened edge image for the Y Channel gray image.

2.1.3.3 Localization

The sharpened edge image is considered as the input intensity image for computing the projection profiles, with white candidate text regions against a black background. The vertical projection profile shows the sum of pixels present in each column of the intensity or the sharpened image. Similarly, the horizontal projection profile shows the sum of pixels present in each row of the intensity image. These projection profiles are essentially histograms where each bin is a count of the total number of pixels present in each row or column. Candidate text regions are segmented based on adaptive threshold values, T1 and T2, calculated for the vertical and horizontal projections respectively. Only regions that fall within the threshold limits are considered as candidates for text. The value of threshold T1 is selected to eliminate possible non text regions such as doors, window edges etc. that have a strong vertical orientation. Similarly, the value of threshold T2 is selected to eliminate regions which might be non-text or long edges in the horizontal orientation.

\[
T_1 = \frac{\text{Mean(Vertical projection profile)}}{60}
\]

\[
T_2 = \frac{\text{Mean(Vertical projection profile)} + \text{Mean(Vertical projection profile)}}{20}
\]

Figure 2.6. (a) Original image (b) Y channel for test image (c) sharpened edge image for the Y Channel

2.1.3.4 Enhancement and Gap Filling

The geometric ratio between the width and the height of the text characters is considered to eliminate possible non-text regions. This ratio value will be defined after experimenting on different kinds of images to get an average value. In this paper, regions with minor to major axis ratio less than 20 are considered as
candidate text regions for further processing. Next a gap image will be created which will be used as a reference to refine the localization of the detected text regions. If a pixel in the binary edge image created is surrounded by black (background) pixels in the vertical, horizontal and diagonal directions, this pixel is also substituted with the background value.

Figure 2.9 Result obtained by connected component based text detection for test image

Figure 2.9(a) Original image (b) Result

III. EXPERIMENTAL RESULTS

The experimentation of the proposed algorithm was carried out on a data set consisting of different images such as indoor, outdoor, posters etc. These test images vary with respect to scale, lighting and orientation of text in the image. Currently the data set consists of number of images, Hindi character images and variations of 2 indoor poster images as well as outdoor images, taken from a digital camera. The complete lists of test images are summarized in Table 1. The performance of each technique has been evaluated based on its precision and recall rates obtained.

![Graph showing precision and recall rates](image)

Precision rate takes into consideration the false positives, which are the non-text regions in the image and have been detected by the algorithm as text regions. Recall rate takes into consideration the false negatives, which are text words in the image, and have not been detected by the algorithm. Thus, precision and recall rates are useful as measures to determine the accuracy of each algorithm in locating correct text regions and eliminating non-text regions. Experiments have been performed to test the proposed system and to measure the accuracy of the system. The results from edge based and connected component based algorithm are shown in Figure-3.1 and Figure-3.2 respectively.

![Image showing experimental results](image)

IV. CONCLUSIONS

Text locating in natural scene image with complex background is a difficult, challenging, and important problem. In this paper, an accurate text region extraction algorithm based on two methods with gray information is presented. The proposed methods work very well on text region in simple images. Texts in natural scene images contain very important information about location information and road signs. One of the further studies is to design the verifying extraction text region by SVM and HMM, and then to design the recognizer system for extraction text regions.

V. REFERENCES


ISSN (Print) : 2319 – 2526, Volume-2, Issue-1, 2013