Text Extraction from Natural Scene Images using Region based Methods-A Survey

1Kumuda T and 2L Basavaraj
1 Adichunchanagiri Institute of Technology, Dept of E& C, Chikmagalur, India.
2 A T M E, Dept of E& C, Mysore, India.
Kumuda.thmmaiiah@gmail.com and basavaraj.atme@gmail.com

Abstract—Text embedded in image provides high-level semantic information for automatic annotation, indexing and retrieval. Text extraction involves detection, localization, extraction, enhancement and recognition of the text from the given image. However, variations of the 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. Research on the location and extraction of the texts in complex background has important significance in current information age. A large number of techniques have been proposed to address this problem. The proposed methods were based on morphological operators, wavelet transform, artificial neural networks, skeletonization operation, region based, histogram techniques etc. All these techniques have their benefits and restrictions. This paper provides the performance comparisons of several existing methods proposed by researchers in extracting the text from natural scene image.

Index Terms—complex background, text detection, text localization, text extraction, text enhancement

I. INTRODUCTION

Wireless communication, mobile devices and cameras are becoming a part of daily life. This gives rise to new applications and opportunities in the field of digital image processing. One of these new research areas in the field of computer vision and pattern recognition is camera-based text recognition and extraction. Text extraction in natural scene images refers to algorithms and techniques that are applied to extract text from the camera captured natural scene images. Text that appears in these images contains important and useful information. While the segmentation and recognition of text from document images is quite successful, detection of coloured scene text is a new challenge for all camera based images. Common problems for text extraction from camera based images are the lack of prior knowledge of any kind of text features such as colour, font, size and orientation as well as the location of the probable text regions. Text extraction from camera based scene images is a very difficult problem because it is not always possible to precisely define the features of text in a coloured scene image due to the wide variations in possible formats; for example, geometry (location and orientation), colour similarity, font and size. Moreover, camera based images can be subjected to numerous possible degradations such as blur, uneven lighting, low resolution and contrast which makes it more difficult to recognise any text from the background noise. Text extraction from images have many useful applications in document analysis, detection of vehicle license plate, analysis of article with tables, maps, charts, diagrams, keyword based image search, identification of parts in industrial automation.

DOI: 03.AETS.2014.5.377
© Association of Computer Electronics and Electrical Engineers, 2014
content based retrieval, name plates, object identification, street signs, text based video indexing, video content analysis, page segmentation, document retrieving, address block location etc. According to the features utilized, text extraction methods can be categorized into two types: region-based and texture-based. Region-based methods use the properties of the color or gray scale in a text region or their differences with the corresponding properties of the background. These methods can be further divided into two types: connected component (CC)-based and edge based. These two approaches work in a bottom-up fashion. First the cc’s or edges are identified and then merging these, bounding boxes for text are obtained. CC 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 filter out non-text components and mark the boundaries of the text regions. 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 heuristics are used to filter out the non-text regions. Texture-based methods use the observation that text in images has distinct textural properties that distinguish them from the background. The techniques based on Gabor filters, Wavelet, FFT, Spatial variance, etc., can be used to detect the textural properties of a text region in an image.

II. Evaluation Of State Of The Art

A number of methods for text extraction using region based techniques have been published in recent years. Use of multiple features and cascade AdaBoost classifier for text detection is described by Yi-Feng Pan et al. [8]. In text localization step text lines are generated using a window grouping method, integrating text line competition analysis. Then within each text line, local binarization is used to extract candidate connected components (CCs) and non-text CCs are filtered out by Markov Random Fields (MRF) model. Three modifications in the adaboost classifier system is made. 1. During text detection stage, Histogram of oriented gradient and multiscale local binary pattern features are used to build up candidate feature pool 2. Introducing a text line computation technique based on relaxation labelling to filter out incorrect text lines around correct ones. 3. Adopting a connected component analysis approach to filter out non-text cc’s based on Markov Random Fields model. The text detection consists of two stages - pre-processing and region analysis. At pre-processing step, the image is first transformed from RGB to gray-level space. Then an image pyramid is formed by rescaling the gray-level image by nearest neighbour interpolation. In the region analysis step, window sampling, feature extraction, feature integral map generation, and window classification are adopted sequentially to detect candidate text windows from the image pyramid. Text localization also consists of two steps: text line generation and text extraction. At text line generation step, a window grouping approach is used to group the detected windows into candidate text lines, and then by using the text line competition analysis, the incorrect lines around the correct ones are filtered out. Connected components are extracted from each text line region by local binarization in text extraction step, and then a connected component analysis approach based on MRF model is employed to filter out non-text components and localize text lines accurately.

Three text extraction methods based on intensity information for natural scene images are proposed by JiSoo Kim [2]. The first method is gray value stretching and binarization by an average intensity of the image. Split and Merge approach is the second one, which is one of well-known algorithms for image segmentation. The combination of the first two method results in third method. In GIA method first, the input color image is converted to a gray image. Then a median filter and a contrast stretching are applied to the gray image. To the resulting image a high pass filtering and an opening operations are applied. Finally an edge image is extracted by Laplacian operator. Then connected components and their bounding boxes are extracted, and their locations, size, and aspect ratio are determined. The text region extraction consists of binarization, long line and noise removal, and candidate text region extraction. In Split/Merge method, the split algorithm starts with the entire image as a single region, split the region into four sub regions, repeats these steps until no further splits take place. Merge Process Considers any two or more neighbouring sub regions, merge the n regions into a single region, Repeat these steps until no further merges take place. Homogeneous segment regions whose width and height exceed max threshold are removed; same is with the homogeneous segment regions less than min threshold. Small holes in the components and connects components separated by small gaps, are closed using standard dilation algorithm during the dilation process.

Xiaoqing Liu [3] described an approach to detect text not only with printed document images but also with natural scene text. They proposed a multiscale edge-based text extraction algorithm-which uses the three important properties of edges, so that detect and extract text in complex images can be detected and extracted.
automatically. The three important properties used by the algorithm are: edge strength, density and variance of orientations. Magnitude of the second derivative of intensity is used as a measurement of edge strength. Based on the average edge strength within a window, edge density is calculated. Variances of orientations are evaluated using four orientations, where 0° denotes horizontal direction, 90° denotes vertical direction, and 45° and 135° are the two diagonal directions, respectively. A compass operator with a convolution operation results in four oriented edge intensity images. Gaussian pyramids are used to generate multiscale images, which successively low-pass filter and down-sample the original image in both vertical and horizontal directions, which results in reduced image. Then these multiscale images are simultaneously processed by the compass operator as individual inputs. Text regions are localized using clustering method. A morphological dilation operator are used to connect the very close regions together while leaving those whose position are far away to each other. A morphological dilation operator with a 7×7 square structuring element is used. Two constraints are used to filter out non-text blobs, the first constraint is used to filter out all the very small isolated blobs where as the second constraint is used to filter out those blobs whose widths are much smaller than corresponding heights. Then the retaining blobs are enclosed in boundary boxes. Finally a thresholding algorithm is used which segments the text regions into white characters in a pure black background.

A robust connected-component (CC) based method for automatic detection and segmentation of text in real-scene images is suggested by Zhu Kai-hua [4]. In this approach first, a Non-Linear Niblack method (NLNiblack) is proposed to decompose the image into gray candidate CCs. Then, classifiers trained by Adaboost algorithm are used in cascade, to which CCs are fed. Each classifier in the cascade responds to one feature of the CC. 12 novel features are proposed, which are insensitive to scale, noise, text language and text orientation. The CCs passing through the cascade are considered as text components and are used to form the segmentation result. Non-CCs of the image are discarded and more computation is spent on processing promising text-like CCs. The method is composed of three stages. After decomposing the image into a set of CCs, all candidate CCs are classified into 2 categories, text or non-text. Then 12 novel features are proposed to expose the intrinsic characteristics of text CCs. The first three features are geometric features. They are area ratio, length ratio and aspect ratio. Four shape regularity features are Holes, Contour Roughness, Compactness and Occupy are used to suppress noise with irregular shape but have strong texture response. Two stroke statistics features are used. The last two spatial coherence features exploit the spatial coherence information to filter out the non-text CCs. In the last stage, the CCs passing through the whole classifier cascade will be processed by a post processing procedure and form the final segmentation result by Non-Linear Niblack method.

Standard deviation based method is used for edge detection by Keshava Prasanna [9]. In the proposed approach edges are detected using color reduction technique, and a standard deviation base method and new connected component properties are used for the localization of text regions. Edges in all directions are detected by using Standard deviation technique. First the images are transformed into HSV color space. Only the intensity data (V channel of HSV) is used during further processing. Noise is reduced by using a median filtering operation on the (intensity) band, then a contrast-limited Adaptive Histogram Equalization is applied for contrast enhancement. Xiaqian Liu [6] proposed an effective approach to locate text in images based on connected component analysis. An image is converted into two complementary binary images by multi-scale adaptive local thresholding operator. Then, connected components (CCs) are extracted from both of them, which ensure that bright or dark text in contrast background can be detected. Further, connected components belongs to characters are identified with the help of stroke features, and the obtained candidate components are further checked on the word level by using a graph to represent spatial relation of different components. Finally, localization of text regions is performed by searching the collinear maximum group over the graph. Localized text regions in two images are refined and merged in post-processing. A set of multi-scale sliding windows are used. For an image, the sliding window moves from left to right and from top to bottom, and the size is gradually zoomed in during the scanning process. By using one-time scanning an image is binarized. Morphological “close” operation is performed on the binary image to remove certain scattered and tiny regions. Inverse of the binary image is obtained to get another complementary image and connected components extraction and text localization are conducted on both of them. Text verification is carried on two levels. One is the character level, based on the features of text stroke structure and the other is the word level based on the spatial relation among character connected components. The obtained sets of connected components are further grouped into words based on spatial location relationship and then the corresponding scene text regions are located.

Céline Mancas-Thillou [7] proposed a method in which similar colors are merged together for an efficient text-driven segmentation in the RGB color space, then it is complemented with spatial and intensity
information which is obtained using Log–Gabor filters, thus enabling the processing of character segmentation into individual components to increase final recognition rates. Orientation and color magnitude are used together. Using log-Gabor filters color variations are combined with spatial information. Unsupervised segmentation algorithm is used with 3-means clustering, where two clusters belong to textual foreground and background, while the third one is a noisy cluster. New text validation measure is used to find the most textual foreground cluster over the two remaining clusters. Fang Liu et al. [1] proposed approach mainly includes two steps. In the first step, a density based clustering method is employed to segment candidate characters by integrating color feature of character pixels and spatial connectivity. In most images, colors of pixels in one character are commonly non-uniform due to the noise. So in the second step a new histogram segmentation method is proposed to obtain the color thresholds of characters. Finally, non-characters are filtered by prior knowledge and texture-based method.

Xu-cheng yin et al [10] designed a fast and effective pruning algorithm to extract maximally stable extremely regions (MSER) as character candidate using the strategy of minimizing regularized variations. The single-link clustering algorithm is used to group the character candidates into text candidates, where automatically clustering threshold and distance weights are learned by a novel self training distance metric learning algorithm. The posterior probabilities of text candidates corresponding to non-text are estimated with a character classifier and text candidates with high non-text probabilities are eliminated. They proposed a self-training distance metric learning algorithm that can learn distance weights and clustering threshold automatically which clusters the character candidates into text candidates. Then a character classifier is used to estimate the posterior probabilities of text candidates corresponding to non-text and remove text candidates with high non-text probabilities. Rong-chichang [5] proposed a connected component based text detection and extraction method. Images are subjected to otsu thresholding, canny edge detection and connected component labelling to obtain candidate text blocks. A fast connected component algorithm enables noise filtering to obtain the candidate texts and their features. AdaBoost classifier training is used to categorize text and non-text characters. Finally connected component fusion method is used to confirm the correctness of the text block, compensating for the problem of possible omissions.

III. PERFORMANCE EVALUATION

The precision and recall rates have been computed based on the number of correctly detected words in an image in order to further evaluate the efficiency and robustness. The precision rate (1) is defined as the ratio of correctly detected words to the sum of correctly detected words plus false positive. False positive are those regions in the image, which are actually not characters of text, but have detected by the algorithm as text regions.

\[
\text{% Precision Rate} = \frac{\text{correctly detected words}}{\text{correctly detected words} + \text{False negatives}} \times 100 \quad (1)
\]

The Recall rate (2) is defined as the ratio of correctly detected Words to the sum of correctly detected words plus false negatives. False negatives are those regions in the image, which are actually text characters, but have been not detected by the algorithm.

\[
\text{% Recall Rate} = \frac{\text{correctly detected words}}{\text{correctly detected words} + \text{False negatives}} \times 100 \quad (2)
\]

Performance analysis of various region-based text extraction techniques is summarized in Table I.

IV. CONCLUSION

In this paper a comparison study of region-based text extraction techniques are presented. There are many applications of a text extraction such as vehicle license plate detection, text based image indexing and retrieval, text based video indexing, keyword based image search, assistance to visually impaired persons, street signs, name plates, robot navigation etc. There are number of algorithms and methods for text extraction from images which uses different attributes related to text such as size, font, color, intensity, connected components, edges, contrast etc. Every approach has its own benefits and restrictions. Even though there are number of methods, there is no single unified approach that fits for all the applications because of deviation in text. This paper also exposed a performance comparison table of different region-based technique that was proposed earlier for text extraction from complex images.
TABLE I PERFORMANCE ANALYSIS OF REGION-BASED TEXT-EXTRACTION TECHNIQUES

<table>
<thead>
<tr>
<th>S.No</th>
<th>AUTHOR</th>
<th>YEAR</th>
<th>METHODS USED</th>
<th>PRECISION RATE (%)</th>
<th>RECALL RATE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HAM-88.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GIA-75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SMA-69.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HAM-76.2</td>
</tr>
<tr>
<td>2.</td>
<td>Xiaoqing Liu et al. [3]</td>
<td>2006</td>
<td>Multi-scale edge based algorithm. Compass operator, feature map generation</td>
<td>96.6</td>
<td>91.8</td>
</tr>
<tr>
<td>5.</td>
<td>Keshava prasanna et al. [9]</td>
<td>2011</td>
<td>Color reduction technique, standard deviation base method, new connected component properties</td>
<td>88.6</td>
<td>92.2</td>
</tr>
<tr>
<td>6.</td>
<td>Xiaoquan Liu et al. [6]</td>
<td>2012</td>
<td>Multi-scale adaptive local thresholding operator, stroke features, collinear maximum group over the graph</td>
<td>65</td>
<td>63</td>
</tr>
<tr>
<td>10.</td>
<td>Xu-cheng yin et al [10]</td>
<td>2013</td>
<td>Pruning algorithm to extract MSERS, single-link clustering algorithm, self-training distance metric learning algorithm</td>
<td>68.5</td>
<td>82.6</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their constructive comments. Also I would like to thank my guide Dr. L Basavaraj, Principal, A T M E 30th his support. This research was supported in part by VVCE, Mysore, India.

REFERENCES

[10] Xu-cheng yin, Xuwang yin, Kaizhu huang, and Hong-wei hao, “Robust text Detection In Natural Scene Images”, IEEE transactions on pattern analysis and machine intelligence,0162-8828/13/$31.00©2013 IEEE.