TEXT DETECTION AND EXTRACTION FROM IMAGES – A COMPREHENSIVE SURVEY

Sheethal S1 and J V Gorabal2

12Department of Computer Science Engg, Sahyadri College of Engg. & Management

Abstract - Text extraction plays a very important role in today’s world by extracting valuable information from the images. The text present in images contains some useful information for automatic annotation, identification, indexing and retrieval. Extraction of this information involves detection, localization, tracking, binarization extraction and finally text recognition. However there might be variations in the text due to the differences in the orientation, size, style, alignment and also because of the low contrast images with a complex background. Due to the rapid growth in multimedia documents, many researchers have contributed on text extraction from images. Many methods were proposed for text extraction. This paper discusses various schemes proposed earlier for extracting text from images. This paper also discusses the performance metrics required to evaluate the performance of the system.

Keywords - Text extraction, Text detection, Local thresholding, learned dictionaries, connected components.

I. INTRODUCTION

Text extraction from images is basically extracting the relevant text data from image. Due to the rapid evolution in technology there has been digitization in all areas. Lots of resources are available in electronic medium. Many existing paper based collections, scanned documents, pictures, book covers, magazines, pamphlets, educational, business cards etc are converted into images. Some of these images impose many challenging research issues in text extraction and recognition. Extraction of text from images have many useful applications in detection of vehicle license plate, document analysis, analysis of article with tables, keyword based image search, identification of parts in industrial automation, content based retrieval, object identification, name plates, street signs, document retrieving, page segmentation, text based video indexing address block location etc.

Due to the evolution of the multimedia documents and growing requirement, studies show a great amount of interest in efficient extraction of text. Intensive research projects are performed for text extraction in images. Many techniques are suggested for text extraction. The proposed methods are based on canny edge detection algorithm, Morphological Component Analysis, DCT, sparse representations etc.

Section 2 gives a detailed survey about the different methods that have been adopted for text detection and extraction in the past few years. Section 3 gives the different metrics with respect to the performance evaluation of the system. Section 4 gives a conclusion to the different approaches which are proposed in this paper.

II. LITERATURE REVIEW

A large number of approaches have been proposed during the course of period for extracting the text from the images. The existing work on text extraction from images can be classified according to different criteria. This paper classifies methods according to the different types of image, analyzes those algorithms and discusses them.

P. Nagabhushan [1] proposed an efficient approach to detect and extract the text in complex background colour document images. The proposed method uses an algorithm called the canny edge detector algorithm to detect the edges. Once the edges are obtained, the dilation operation is
performed on them. It was found that it created holes in most of the connected components. These correspond to character strings. Connected components without holes were eliminated. Other non-text components were eliminated by computing and analyzing the standard deviation of each connected component. There was an unsupervised local thresholding which was devised to perform foreground segmentation in detected text regions. At the end the noisy text regions were identified and were processed to further enhance the quality of retrieved foreground.

Thi [2] describes an approach for novel text extraction from graphical document images. The proposed method uses Morphological Component Analysis (MCA) algorithm. This algorithm describes an advancement of sparse representation framework. It shows two appropriately chosen discriminative over complete dictionaries. Two discriminative dictionaries were based on undecimated wavelet transform and curvelet transform.

Angadi [3] proposed an automated method to detect and extract text prior to further image analysis. The proposed methodology uses Discrete Cosine Transform (DCT) based high pass filter to remove and prevent the dissemination of the constant background. The processed image was divided into 50x50 block and then the texture feature matrix was computed. A discriminant function which was anew was used to classify text blocks. The detected text blocks were merged back together to obtain new text like regions. Finally, the refinement phase was a post processing step used to improve the detection accuracy. This phase used to cover small portions of missed text present in adjacent undetected blocks and unprocessed regions. The proposed method comprises of 5 phases; Background removal/suppression in the DCT domain, texture features computation on every 50x50 block and obtaining a feature matrix D, Classification of blocks, merging of text blocks to detect text regions, and refinement of text regions. This methodology had been conducted on 100 indoor and outdoor low resolution natural scene images containing text of different size, font, and alignment with complex backgrounds containing Kannada text and English text. The approach also detected nonlinear text regions and can be extended for text extraction from the images of other languages with little modifications.

Thanh-Ha do[4] proposed an approach to extract text regions from graphical documents. In this method, first empirically two sequences are constructed of learned dictionaries for the text and graphical parts respectively. Then sparse representations are computed for all different sizes and non-overlapped document patches in these learned dictionaries. Each patch can be classified based on these representations. It is classified into the text or graphic category by comparing its reconstruction errors. Same-sized patches in one category are then merged together to define the corresponding text or graphic layers which are combined to create a final text/graphic layer. Finally, in a post-processing step, text regions are further filtered out by using some learned thresholds.

Ranjini [5] talks about English text extraction from blob in comic image. Detecting text and extracting text from comic images helps to preserve the text and formatting during the conversion process and it provides very fine quality of text from the printed document. Initially, a pre-processing is done on the image. In the pre-processing step the RGB images are converted into a binary image by applying the threshold values between 0 to 1.Then the image is subjected to balloon detection. CCL algorithm is applied to the noise removed RGB images for detecting the connected components in the image which helps to detect the connected components in the image often it is used for Balloon detection. Text blob extraction is performed on them which are used to identify text blobs from non text blobs. To avoid the false detection and to reduce the complexity the text blobs need to be identified exactly. The identification is done, based on the features of blob size. Finally, the text is recognised and extracted by using the optical character recognition method.

Karim Sobottka [6] proposed a significant automatic approach for text location and identification on coloured book and journals. In the pre-processing step a clustering algorithm is applied to reduce the amount of small variations in colours. For extracting text hypothesis two methods have been developed. One method is based on a top-down analysis which uses successive splitting of image regions. The other method is a bottom-up region growing algorithm. At the end, both methods are combined to distinguish between text and non-text elements. Text elements are
binarized using automatically extracted information about text colour. The binarized text regions can be used as input for a conventional OCR module. The proposed method is not restricted to cover pages, but can be applied to the extraction of text from other types of colour images as well.

S.Audithan [7] formulated a computationally fast method to extract text like regions from printed documents. It uses Haar discrete wavelet transform to detect edges of candidate text regions. Thresholding technique is used to remove non-text edges. Morphological dilation operator is used to connect the isolated candidate text edge and then a line feature vector graph is generated based on the edge map. This method explains an improved canny edge detector to detect text pixels. The stroke information extracts the spatial distribution of edge pixels. Finally text regions are generated and filtered according to line features.

Syed Saqibet [8] described a significant approach for curled textline information which is extracted from grayscale camera-captured document images. This approach is based on differential geometry, which uses local direction of gradients and second derivatives as the measure of curvature. The grayscale textline is enhanced by using multi-oriented multi-scale anisotropic Gaussian smoothing. Detection of central lines of curled textlines is found using ridges. Hessian matrix is used for finding direction of gradients and derivatives. By using this information, ridges are detected by finding the zero-crossing of the appropriate directional derivatives of smoothed image. This method is robust against high degrees of curl and requires no post-processing.

III. PERFORMANCE EVALUATION

There are several performance evaluations to estimate the algorithm for text extraction. Some of the metrics to evaluate performance are precision, recall and f-measure. To order to evaluate the efficiency and robustness of an algorithm, precision, recall and f-measure are computed based on the number of correctly detected characters in an image.

3.1. Precision

Precision is the percent of all relevant documents that is returned by the search.

3.2. Recall

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

3.3. F-measure

The F measure is defined as the weighted mean of the precision and recall.

IV. CONCLUSION

Text extraction can have a wide range of applications. It can be used for keyword based image search, text indexing and retrieval, page segmentation, object identification, technical paper analysis, vehicle license detection and recognition etc. A number of methods have been proposed over a last few years to extract text from images. These methods consider different attributes of text in image such as size, font, colour, alignment, contrast, orientation etc. This paper provides a broad study of the algorithms proposed earlier for text extraction. Every method has its own benefits and restrictions. Even though there are many algorithms that were proposed earlier, there is no single algorithm that gives the best performance. The future work mainly concentrates on developing an algorithm for fast text extraction from an image and which best fits the performance.

REFERENCES

[1] Nirmala Shivananda and P. Nagabhushan, “Separation of Foreground Text from Complex Background in Color Document Images”, IEEE Seventh International Conference on Advances in Pattern Recognition, 2009.
[2] Thai V. Hoang, S. Tabbone(2010) ,“Text Extraction From Graphical Document Images Using Sparse Representation” in Proc. Das, pp 143–150.
[3] S. A. Angadi and M. M. Kodabagi, “Text Region Extraction from Low Resolution Natural Scene Images using Texture Features”,IEEE 2010.
[4] Thanh-Ha Do, Salvatore Tabbone, Oriol Ramos-Terrades, “Text/graphic separation using a sparse representation with multi-learned dictionaries”, 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.

[5] Dr. M. Sundaresan and S. Ranjini, “Text Extraction from Digital English Comic Image Using Two Blobs Extraction Method”, IEEE Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME) March 21-23, 2012.

[6] Karin Sobottka, Horst Bunke and Heino Kronenberg (2009), “Identification Of Text On Colored Book And Journal Covers”, ICDAR.

[7] Sachin, Grover, Kushal Arora, Suman K. Mitra (2009), “Text Extraction From Document Images Using Edge Information”, IEEE India Council Conference.

[8] Syed Saqib Bukhari, Thomas M. Breuel, Faisal Shafait (2009), “Textline Information Extraction From Grayscale Camera-Captured Document Images”, ICIP Proceedings Of The 16th IEEE International Conference On Image Processing, pp: 2013 – 2016.