TEXT RETRIEVAL SYSTEM: A Review

Dr. S.BHUVANESWARI[1] Prof. G.ASHA[2], Prof. S.SUNDARAMATHI[3], Prof. S. SASIKALA[4], Prof. S.SUMATHI[5]
ASSISTANT PROFESSOR[12345],
Department of BCA.
Annai College of Arts and Science,
Kovilacheri, Kumbakonam, India.

***Abstract*— segmentation is the process of subdividing a digital image into multiple sets of pixels. Image segmentation is thus inevitable. Segmentation is very much helpful for text-based images extracting specific information like line, word or even character from the entire image. The extracted information can be a line or a word or even a character. This paper discusses various methodologies to segment a text based image at various levels of segmentation and also it serves as guidance for readers working on the text based segmentation. First, the need for segmentation is discussed and the various factors affecting the segmentation process are also discussed. Finally, the available techniques with their advantages and disadvantages are reviewed, along with directions for future perspective. The discussion is made on the handwriting recognition since this area requires more advanced techniques for efficient information extraction and to reach the ultimate goal of machine simulation of human reading.**

***Index Terms*—Handwritten text, printed text, levels of segmentation, segmentation methods, text document image analysis.**

1. INTRODUCTION

Text detection and recognition is an area of research which aims to develop a system to read the text from images automatically. Due to the digitization, most of the resources such as historical manuscripts, land records, old books, journals, newspapers, even hand written text documents are converted to images. Automatically detecting and extracting text from these digitized images present many challenging and recognize due to their variation in terms of size, font, style, orientation, alignment, contrast, background etc.

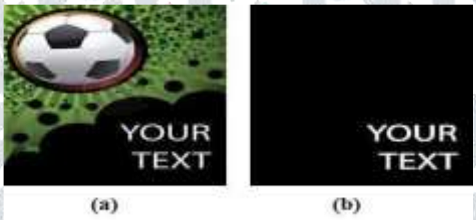

Fig. 1. a) Text in image b) Extracted text.

Fig. 1 (a) shows a text image. Automatic methods for text extraction from images aim to detect the characters based on the general properties of text pixels, namely: text contains a number of edges, text width is larger than height, and text is usually of uniform size. Text size is a major factor which is usually of uniform size and texture property of text is irregular and weak [1]. These variations turn the process of word detection complex and difficult [2]. In the case of handwritten manuscripts, differently from machine printed, the complexity of the problem even increases. Since handwritten text can vary greatly depending on the user skill, disposition and even cultural background. Here, we present a method to segment text lines based on morphology and histogram. Morphological operations are used to produce a binary image [3], as an initial step in the process of text line extraction from video images containing text information. In their application, a not precise box containing the region of the text is used as output of the system to identify machine printed text in different video contexts. An important fact in relation to image analysis based on contrast is that this characteristic is robust in relation to changes in illumination and it is invariant to different image transformations such as scaling, translation and skewing. Once the page document has been preprocessed, a technique based on projection profiles is applied. In [4] the commonly used printed document segmentation technique, projection profiles can also be applied for handwritten documents. Work in [5] proposed segmentation approach which is divided in 3 levels and utilizes projection profiles along the Y and X axes alternately. In this work, new method is proposed to detect the hand written text automatically. The proposed method is based on morphological operation and histogram projection profiles.

Hand written text documents are common today, for example, old precious documents are mostly in handwritten format. Robust detection of text from these documents is a challenging problem.

Optical character recognition (OCR) is used to transform the images of either handwritten or printed document to a machine readable and editable format. In general, all OCR systems have the following stages: image preprocessing, segmentation, extraction of features and finally recognition of characters. The results of each of these stages are greatly affected by the performance of the previous stages. To make the results of the subsequent stages more accurate, segmentation plays an important role. The extraction of region of interest from the given image is termed as segmentation. In the segmentation of scene text images, first we extract the lines then the words and finally the characters. Segmentation of characters from a document is still a open challenge in developing efficient OCR systems.

2. RELATED WORKS

The text in an image gives detailed information about a scene, which is helpful in a wide range of applications, such as indexing, image understanding, monitoring/controlling the movement of an object, and human-computer related applications. This section discusses the work done so far related to the detection of text from images automatically both from simple and complex backgrounds. The aim of text detection is to identify candidate text regions in a given input image.

Text detection methods are broadly divided into three kinds: connected component based, texture based, and edge based. Though the connected component based method locates the text very easily, it fails in the complex background [6]. The computational complexity is more in the case of texture based method in the classification stage, and appearance of the text like regions may confuse the detection process [7]. The performance of edge based methods is not appreciable while handling large size texts [8]. The text detection methods range from using simple classifiers [9] to efficient multi-stage processes which employ many algorithms and layers [10] and [11]. Authors in [12] employed a method in which, after binarizing the input image, connected component analysis, using conditional random field (CRF) to detect the text lines. Authors in [13] introduced a new approach for segmenting the text from colour images. To locate the candidate text lines, the multi scale wavelet features and structural information are used. From the candidate text lines, the true text is identified using support vector machine (SVM) classifier. This approach has four stages. In the pre-processing step, input text blocks are rescaled with cubic interpolation, and for smoothing and removing noise, Gaussian filter is used. These rescaled image blocks are separated into text connected components and non-text connected components which are eliminated using component filtering procedure. K-Means clustering algorithm is then applied to group the remaining components into several text layers after which, a set of proper constraints are applied to detect the real text layers. An efficient document text extraction method which is computationally fast is proposed in [14]. Haar discrete wavelet transform is used to detect edges of the candidate text regions. Then thresholding technique is used to remove the remaining non-text edges from the image. To connect the isolated candidate text edges, morphological dilation operator is used. Then, based on the edge map, the line feature graph is generated. Further, improved Canny edge detector is utilized to detect text pixels and the spatial distribution of edge pixels helps to extract the stroke information. Finally, according to the line features, the image is filtered and exact text regions are isolated. Text embedded in complex coloured document image is detected by authors in [15]. They proposed edge based features for their work. The weighted sum of the R, G, and B components was used to convert the colour image to gray scale. Then Sobel masks are applied to detect the horizontal and vertical edges. This is followed by the elimination of weak edges. Then the edge image is split into non-overlapping blocks of $m \times m$ pixels, where m depends on the resolution of the image. Using pre-defined threshold, block classification is performed, and it differentiates the text from the image. Authors in [16] introduced a new method for text extraction from complex colour document images. Edge detection is performed using canny edge detector followed by the dilation morphological operator. This results in the creation of holes in most of the connected components that represent character strings, and the components without holes are removed, which corresponds to non-characters. The standard deviation of each connected component is computed for eliminating the remaining non-text components. Still, if noisy text region persists, it is replaced to improve the quality of the retrieved foreground.

The authors in [17] first performed colour space reduction after which segmentation and spatial regrouping was carried out for detecting text. The authors tried to tackle the problem of touching the text; however, the segmentation algorithm failed in the case of poor quality documents and specifically in case of video sequences. Authors in [18] proposed a method of text extraction which is done in three stages. Candidate text region is detected in the first stage by generating a feature map. Feature map is a binary image created using the edge characteristics like strength, density, orientation, and the pixel intensity of the feature map gives clues about the possibility of text regions.

Text region localization is the second stage that helps to detect the non-text regions. Two constraints are used to find and filter very small isolated blocks whose Text region localization is the second stage that helps to detect the non-text regions. Two constraints are used to find and filter very small isolated blocks whose width is very small compared to the height of the blocks, and in the third stage, character extraction is done using the existing optical character recognition (OCR) engines.

Authors in [19] introduced a method for detecting and extracting text regions from natural scene images whose resolution is low. For removing the constant background, discrete cosine transform (DCT) based high pass filter is used. The processed image is divided into 50×50 blocks. The texture feature matrix is computed for each of these blocks. With the help of the newly defined texture feature matrix, text and non-text blocks are classified. Then the text blocks are merged to obtain the new text regions. Then, post-processing is done to cover small portions of missed text in adjacent undetected blocks.

Authors in [20] proposed a new approach, which first isolates the letters and then groups them, after which words are restored. The segmentation process is based on toggle mapping morphological segmentation, and multiple SVM classifiers are used for classification.

The new scene text detection algorithm based on two classifiers is proposed in [21]. Candidate word regions are generated using the first classifier. The second classifier is used to filter the non-text regions. The maximally stable extremal region algorithm is used to extract the connected components from the image. The connected components are then clustered to generate candidate text regions which are normalized to ascertain whether the candidate region contains text or not.

The value of stroke width for each image pixel is found out using the novel image operator and this is used for text detection in natural images in [22]. This image operator is data dependent and local, thus it is fast and eliminates multi-scale computation. This simple algorithm detects the text irrespective of fonts and languages. By considering, the directions of the improved strokes, grouping of letters could be enhanced and curved text line can also be detected. Partition method for text detection using gradient and colour information of pixels is given by the authors in [23]. For locating the text, these features are used at the character level to study the regularity of text.

The authors in [24] first divided a video frame into 16 blocks and combined wavelet, median-moments, and K-Means clustering was applied to identify text candidates at the block level. Then all the blocks containing text candidates in the frame were integrated, and the text candidates were mapped on to a Sobel edge map of the original frame to obtain text representatives. Also, an iterative procedure was proposed by the authors called as angle projection boundary growing (APBG) to tackle multioriented text. The authors in [25] proposed a method based on the text pixel's context information to detect the text in natural scene images. The context is analyzed using the stroke properties and spatial distribution of the text line. SVM classifier is used to learn the context, and the performance of this method depends on the training samples used.

A method using Harris Corner detector proposed in [26] is based on structure modeling and character appearance. For detecting the text in natural scene images, the interest points detected are fed to the classifier and this method fully relies on the classifier, and on the training samples used. The text's horizontal intensity variations are used to analyze the images in transform domain. Spatial domain techniques also exploit this horizontal intensity variation to find out high probability regions that contain text.

Wavelet transform is used in [27] and DCT in [28] to detect the text. Spatial cohesion features like fill factor, size, aspect ratio, and horizontal alignment are employed to ascertain if the candidate text regions are stable with its neighborhood and false positives are discarded [29].

Authors in [30] discussed a superimposed text extraction method to detect text depicting player information and match score in sports videos. Colour histogram technique was used to extract key frames from the video to reduce the number of frames to be processed which were then converted to gray scale. Then detected text regions were cropped and edges were detected by applying Canny edge detection algorithm, and finally the identified text regions were fed to OCR system which generated index able keywords.

In summary, almost all of the previous works use techniques like connected component labelling, wavelet approximation and, template matching for detecting text automatically.

3. CONCLUSION AND FUTURE WORK

The works discussed in this paper concludes that the algorithms differ with different work modes. Algorithms differs with printed text and hand written text The pixel counting algorithm is simple to implement and we can conclude that it excels only for the printed text document. This algorithm can be used for a handwritten document only if it has factors like i) some useful parameters (ii) the document with uniform text size iii) and uniform interline spacing. But it fails to provide good results while working with handwritten text images. Also, skew correction module is required.

A histogram approach is flexible and provides better results outrivaled the previews for both printed and handwritten text documents. This approach is comparatively slower than pixel counting approach while dealing with printed document due the increase in the computation. On the other hand this algorithm gives better results for handwritten text document and provides results with high level of accuracy. The only disadvantage of the histogram algorithm as compared to the pixel counting approach is the increased computation and the resulting space complexity, thereby experiencing decrease in computational speed.

The future work is to implement the new algorithm and conduct a comprehensive comparison of the various techniques as discussed in the paper. This work will provide ease to researchers, scientists and engineers working with text image and provides them with application specific algorithm selection knowledge to comprehend the need of achieving faster and efficient methodologies for segmenting the text image.

References:

[1] R. Manmatha, J.L., Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents", *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, pp. 1212-1225.

[2] Laurent Joyeux, Samia Boukir, Bernard Besserer, " Film Line Scratch Removal using Kalmar Filtering and Bayesian Restoration", IEEE Workshop on the Application of Com-puter Vision, pp.1-6, 2000.

[3] Anil K Jain and Bin Yu, "Automatic text location in images and video frames," Pattern recognition, vol. 31, no. 12, pp. 2055–2076, 1998.

[4] Kwang In Kim, Keechul Jung, and Jin Hyung Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 12, pp. 1631–1639, 2003.

[5] VictorWu, Raghavan Manmatha, and Edward M Riseman, "Textfinder: An automatic system to detect and recognize text in images," IEEE Transactions on pattern analysis and machine intelligence, vol. 21, no. 11, pp. 1224–1229, 1999.

[6] Xiangrong Chen and Alan L Yuille, "Detecting and reading text in natural scenes," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, vol. 2, pp. II–366.

[7] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, "A robust system to detect and localize texts in natural scene images," in Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on. IEEE, 2008, pp. 35–42.

[8] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, “Text localization in natural scene images based on conditional random field,” in Document Analysis and Recognition, 2009. ICDAR’09. 10th International Conference on. IEEE, 2009, pp. 6–10.

[9] Yaowen Zhan, Weiqiang Wang, and Wen Gao, “A robust split-and-merge text segmentation approach for images,” in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. IEEE, 2006, vol. 2, pp. 1002–1005.

[10] S Audithan and RM Chandrasekaran, “Document text extraction from document images using haar discrete wavelet transform,” European Journal of Scientific Research, vol. 36, no. 4, pp. 502–512, 2009.

[11] Grover Sachin, Kushal Arora, and Suman K Mitra, “Text extraction from document images using edge information,” in IEEE India Council Conference, 2009.

[12] P Nagabhushan, Shivananda Nirmala, et al., “Text extraction in complex color document images for enhanced readability,” Intelligent Information Management, vol. 2, no. 02, pp. 120, 2010.

[13] Anil K Jain and Bin Yu, “Automatic text location in images and video frames,” Pattern recognition, vol. 31, no. 12, pp. 2055–2076, 1998.

[14] Kwang In Kim, Keechul Jung, and Jin Hyung Kim, “Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 12, pp. 1631–1639, 2003.

[15] VictorWu, Raghavan Manmatha, and Edward M Riseman, “Textfinder: An automatic system to detect and recognize text in images,” IEEE Transactions on pattern analysis and machine intelligence, vol. 21, no. 11, pp. 1224–1229, 1999.

[16] Xiangrong Chen and Alan L Yuille, “Detecting and reading text in natural scenes,” in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, vol. 2, pp. II–366.

[17] Anil K Jain and Bin Yu, “Automatic text location in images and video frames,” Pattern recognition, vol. 31, no. 12, pp. 2055–2076, 1998. 100

[18] Xiaoqing Liu and Jagath Samarabandu, “An edge-based text region extraction algorithm for indoor mobile robot navigation,” in Mechatronics and Automation, 2005 IEEE International Conference. IEEE, 2005, vol. 2, pp. 701–706.

[19] SA Angadi and MM Kodabagi, “A texture based methodology for text region extraction from low resolution natural scene images,” International Journal of Image Processing, vol. 3, no. 5, pp. 229–245, 2009.

[20] Jonathan Fabrizio, Matthieu Cord, and Beatriz Marcotegui, “Text extraction from street level images,” in CMRT09-CityModels, Roads and Traffic, 2009, vol. 38, pp. 199–204.

[21] Hyung Il Koo and Duck Hoon Kim, “Scene text detection via connected component clustering and nontext filtering,” Image Processing, IEEE Transactions on, vol. 22, no. 6, pp. 2296–2305, 2013.

[22] Boris Epshtein, Eyal Ofek, and Yonatan Wexler, “Detecting text in natural scenes with stroke width transform,” in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 2963–2970.

[23] Chucai Yi and YingLi Tian, “Text string detection from natural scenes by structurebased partition and grouping,” Image Processing, IEEE Transactions on, vol. 20, no. 9, pp. 2594–2605, 2011.

[24] Palaiahnakote Shivakumara, Anjan Dutta, Chew Lim Tan, and Umapada Pal, “Multioriented scene text detection in video based on wavelet and angle projection boundary

[25] growing,” Multimedia Tools and Applications, vol. 72, no. 1, pp. 515–539, 2014.

[26] Yuning Du, Genquan Duan, and Haizhou Ai, “Context-based text detection in natural scenes,” in Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, 2012, pp. 1857–1860.

[27] Chucai Yi and Yingli Tian, “Text extraction from scene images by character appearance and structure modeling,” Computer Vision and Image Understanding, vol. 117, no. 2, pp. 182–194, 2013.

[28] Huiping Li, David Doermann, and Omid Kia, “Automatic text detection and tracking in digital video,” Image Processing, IEEE Transactions on, vol. 9, no. 1, pp. 147–156, 2000.

[29] Yu Zhong, Hongjiang Zhang, and Anil K Jain, “Automatic caption localization in compressed video,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, no. 4, pp. 385–392, 2000.

[30] Thomas Retornaz and Beatriz Marcotegui, “Scene text localization based on the ultimate opening,” in International Symposium on Mathematical Morphology, 2007, vol. 1, pp. 177–188.