International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

An Efficient Method to Extract Digital Text From Scanned Image Text

Jenick Johnson
ECE Dept.,
Christ the King Engineering College
Coimbatore-641104, Tamil Nadu, India

K. Venkatesh
Asst. Professor
ECE Dept., Christ the King Engineering College
Coimbatore-641104, Tamil Nadu.

Suresh Babu. V
Asst. Professor
ECE Dept., Christ the King Engineering College
Coimbatore-641104, Tamil Nadu.

M. Varatharaj
Asst. Professor
ECE Dept., Christ the King Engineering college
Coimbatore-641104, Tamil Nadu.

***Abstract*—Extraction of text from an image document is one of the challenges faced by nowadays. The paper focuses on the problem of text detection from scanned image documents for the improvements in its conventional techniques. The paper uses the Optical Character Recognition (OCR) technique in order to extract actual text presented in the input image. Other techniques such as Maximally Stable Extremal Region (MSER) to estimate the scales and orientation CC extraction which is used as an algorithm in order to enhance the extraction and retrieval process.**

Key word-OCR, CC, MSER

I. INTRODUCTION

Text line detection and recognition has crossed lots of milestones in various fields such as unmanned drone, defense vehicle, visually impairment assistance, sign board detection, security surveillance, and even for unmanned driving. It is also seeking further developments in future to be more reliable and efficient. Therefore, there was a lot of researches and efforts on the account of creating a successful platform for text detection and recognition. Developing an ideal platform for extraction has led to many works like text-line detection in camera-captured document [1], text detection in natural scene with edge analysis [2]. In recent years many approaches have been proposed which actuated promising results [3-5]. In edge analysis method it is based on two steps: candidate edge combination and edge classification [2]. However, detecting and recognition text from a scanned image font and alignment [6,7]. And there were many researches and conventional ideas in order to achieve higher extracting mechanism and to develop an effective platform.

In future many technologies based on this concept such as sign board detection in traffic by unmanned vehicle and unmanned drone for defense purpose featuring image processing and face recognition, banks with face detection entry.

II. PROPOSED METHOD

1. MSER Algorithm

In this paper I have proposed an idea which combines the techniques of OCR and other robust algorithm such as extraction and maximally stable extremal region (MSER) [9]. The conventional binarization algorithm was replaced by MSER technique because this binarization algorithm was designed only for dark text on white background it also has complexity in processing in multiple scales.

2. *CC Extraction*

After MSER approach it is then proceeded to CC extraction by clustering to a small segment where each segment corresponds to a text-line candidate [10,11]. Hyung Il Koo paper [1] uses estimated the states (scale and orientations) of Connected Components and was able build text line candidates using these estimated states. The CC-based approach builds text-lines by clustering Connected Components and the region-based approach is implemented a classifier of local patches to find text-lines.

3.OCR

In Optical Character Recognition scanned document such as printed text, images which of typed format and that of sign board, billboard in photo or even text imposed on an image It is also now widely used as a form of information from computerized receipts, printed paper data records, printouts of static-data, passport documents, invoices, bank statements, business cards, mail, or any suitable documentation. It is a method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

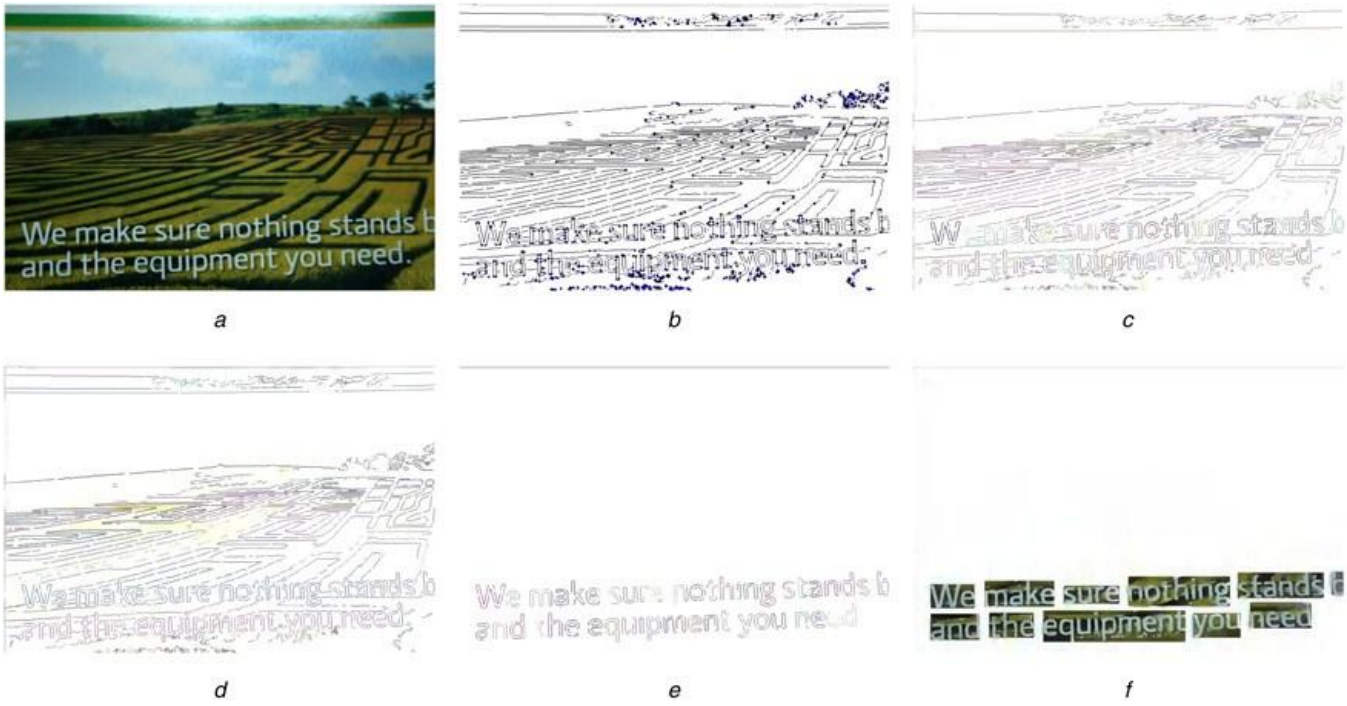


Image a, b, c, d, e, f is based on extraction using conventional based algorithm which is supported by only scanned image that does not contain any curved text-line. Therefore, in case of any, it will result in failure and generates false positives.

The Connected Components extraction is implemented by using, the MSER algorithm has been commonly used in recent scene text detection methods [12]. The MSER algorithm enables to have text component candidates by extracting both small segment and large structures at the same time. In [16], another MSER method which is proposed to resolve the overlaps between extracted CCs.

Consider a set of extracted CCs as

$$C = \{c_p\}$$

For each CC, it compute the center (x_p, y_p) and the covariance matrix Σ_p of pixel positions. The covariance matrix is then represented with

$$\Sigma_p = \sigma_1 v_1 v_1^\top + \sigma_2 v_2 v_2^\top \quad (2)$$

by implying the eigenvalue for decomposition [13], where σ_1 and σ_2 are eigenvalues ($\sigma_1 < \sigma_2$), and v_1 and v_2 are the corresponding eigenvectors so far. With this decomposition, it represent $\{c_p\}$ with ellipses be selected.

2a.

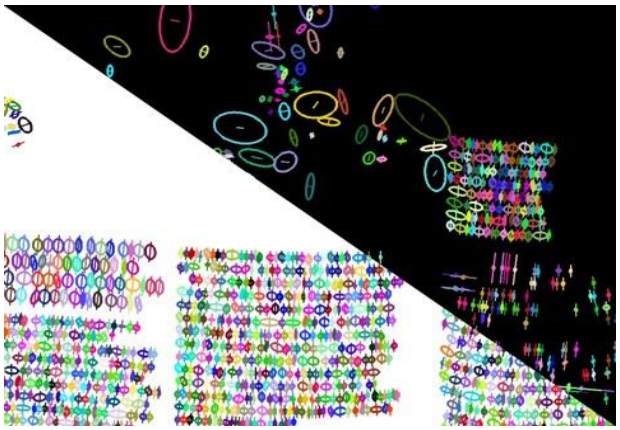
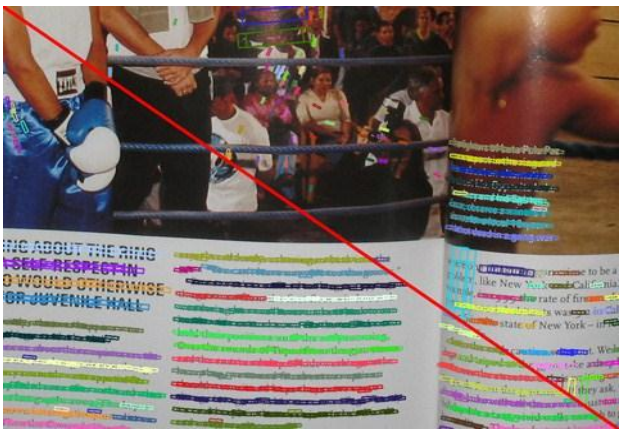


Image 2a. represents estimated states

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

2b.



Images, 2b represents bottom clustering process which is mainly performed in CC extraction.

B. definition of states for estimation

The state of c_p is defined as a pair of two values:

$$f_p = (\theta_p, s_p) \quad (3)$$

equation represents as θ_p is the orientation of a corresponding text-line, s_p is the interline spacing. For orientations, it quantize $[0,\pi]$ to N_D = 32 levels .Therefore the equation is given by:

$$\theta = \{\frac{k \times \pi}{N_D} \Big| k = 0, \dots N_{D-1}\} \quad (4)$$

the minor and major axes are v_1 and v_2, respectively. This super-pixel representation enables to generate a set of CCs memory-efficiently

In order to estimate the scales, the proposed cost function uses the Discrete Fourier Transform (DFT) of projection profiles and 10 discrete scales are chosen correspondingly to take the benefit the computational efficiency of Fast Fourier Transform (FFT) .

DISCRETE LEVELS OF INTERLINE SPACING ($S_p = \frac{N}{K}$). THE NUMBERS IN PARENTHESIS ARE REDUNDANT AND ARE NOT USED.

	k =5	k =4	k =3	k =2
N =64	12.8	16.0	21.3	(32.0)
N =128	25.6	32.0	42.7	(64.0)
N =256	51.2	64.0	85.3	128.0

We also define the distance between two states as

$$\|f_p - f_q\| = \|s_p - s_q\|s + \|\theta_p - \theta_q\|_\theta \quad (5)$$

III. TEXT-LINE EXTRACTION

After calculating the estimated states, the text-line candidates are generated and the non-text-lines in the candidate set are filtered out using a trained classifier. By using bottom-up clustering method, text-lines can be more effectively extracted to individual text-blocks, thereby text-block segmentation prior for the text-line candidate generation.

A. Text-block Segmentation

It is possible to extract text-blocks by removing the edges which are long when compared that of the estimated scales:

$$d_{pq} \geq \in \times \min(s_p, s_q) \quad (6)$$

then, in order to make clusters by partitioning C into

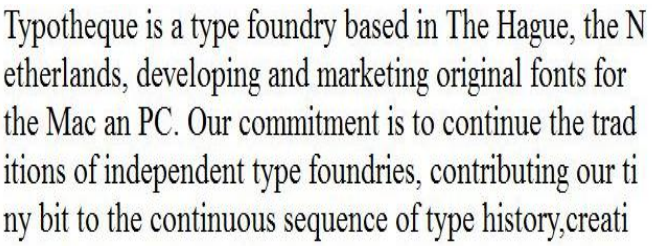
$$(C1, C2, \dots, CL) \quad (7)$$

This technique works well in many cases; however, it has difficulties in handling close text-blocks and unfolded book surfaces but it performs better in curved text lines.

To prevent from this problem, a projection profile-based method is also adopted [14]. That is, for each cluster C_i, first calculate the most frequent orientation of CCs in the cluster:

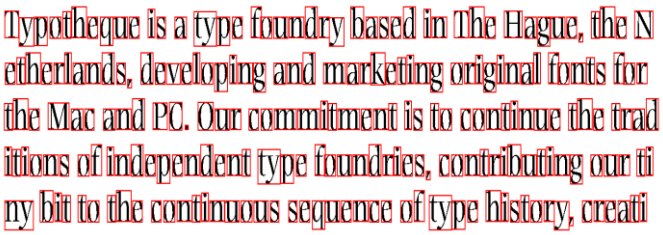
$$\theta_{Ci} = \text{argmax}_\theta \ |\{c_p \in C_i | \theta_p = \theta\}|. \quad (8)$$

That is the most frequent orientation θ_{Ci}, calculated is projection profiles in $[\theta_{ci} - \Delta, \theta_{ci} + \Delta]$


Typotheque is a type foundry based in The Hague, the N etherlands, developing and marketing original fonts for the Mac an PC. Our commitment is to continue the trad itions of independent type foundries, contributing our ti ny bit to the continuous sequence of type history,creati

3a. Input image before text-line segmentation

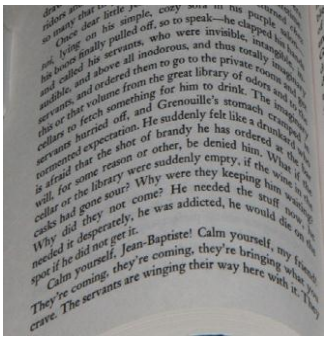
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017



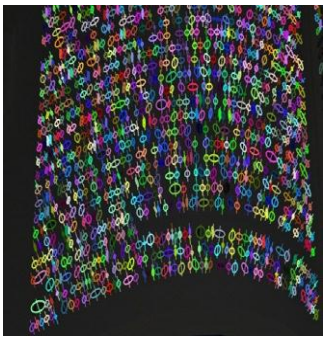
3b. Output image after text-line segmentation

B. Text-line candidate generation

Text-line candidates are generated by using a bottom-up grouping method for each text-block $C_i \subset C$ [36]. The technique works by drawing rectangle $ws_p \times hs_p$ for each CC, whose center is (x_p, y_p) and orientation is θ_p, and each connected region is considered to be a text-line candidate (to be precise, a candidate is a set of CCs). Single (w,h) alone doesn't work for all cases; a small scale variation results in over-segmentation and a large scale variation yields under segmentation. In order to address this problem, it adopt a sequence of w in the clustering and also propose a stopping criterion.



4a.



4b.

Image 4a, 4b denotes estimation of states that is by using this technique which relatively clusters all the text blocks into successively larger components until all regions are identified in the image.

IV. PROPOSED EQUATIONS AND CALCULATIONS

The alphabets error and error rates are determined by a proposed equation which is intended to calculate it accurately. By this approach it was able to estimate the extraction error in effectively.

Alphabets error rate[AER]=

$$\frac{\text{no.of alphabets extracted wrongly}}{\text{no.of actual alphabets in the image}} \quad (9)$$

Accuracy in text extraction = $\frac{1}{AER}$ (10)

CONCLUSION

In this paper, I proposed a text-line detection method from document images such that it would be able harvest text from image document specific text extraction mechanism. To develop this method, I developed CC-based approach for the scene text detection problem. The combined algorithm uses extracted CCs with the MSER algorithm and built text-line candidates by using the bottom up clustering. After evaluating this method with that of conventional dataset, this method compared favorably with conventional methods. And also used advantage of different extraction methods which concentrates on curved text, dark background and makes use of CC extraction mechanism for accurate text extraction based on clustering smaller components. OCR algorithm mainly used for text extraction from printed documents and also featured inter line spacing.

REFERENCES

[1] Hyung Il Koo, "Text line detection in camera-captured document images using the state estimation of connected components," IEEE Trans. Image Process., vol. 25, issue, pp. 5358–5368 , .11 , Nov 2016.

[2] Chong Yu , Yonghong Song, Quan Meng, Yualin Zhang ,Yang Liu "Text detection and recognition in natural scene with edge analysis,"IET comput.Vis.,2015,Vol.9,Iss.4,pp.603-613.

[3] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez I Bigorda, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazan Almazan, and L.-P. de las Heras, "ICDAR 2013 robust reading competition," in International Conference on Document Analysis and Recognition (ICDAR), Aug 2013, pp. 1484–1493.

[4] D. Karatzas, L. Gomez i Bigorda, A. Nicolaous, S. Ghosh, A. Bagdanov,

[5] M. Iwamura, J. Matas, L. Neumann, R. C. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading competition," in International Conference on Document Analysis and Recognition (ICDAR), Aug 2015, pp. 1156–1160.

[6] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Coupled snakelet model for curled textline segmentation of camera-captured document images," in International Conference on Document Analysis and Recognition, July 2009, pp. 61–65.

[7] M. Diem, F. Kleber, and R. Sablatnig, "Text line detection for heterogeneous documents," in International Conference on Document Analysis and Recognition (ICDAR), Aug 2013, pp. 743–747.

[8] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Ridges based curled textline region detection from grayscale camera-captured document images," in Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns. Springer-Verlag, 9 2009.

[9] X. C. Yin, Z. Y. Zuo, S. Tian, and C. L. Liu, "Text detection, tracking and recognition in video: A comprehensive survey," IEEE Trans. Image Process., vol. 25, no. 6, pp. 2752–2773, June 2016 S. S. Bukhari, F. Shafait, and T. M. Breuel, "Towards generic textline extraction," in International Conference on Document Analysis and Recognition (ICDAR), Aug 2013, pp. 748–752.

[10] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2012, pp. 1083–1090.

[11] F. Shafait and T. M. Breuel, "Document image dewarping contest," in Int. Workshop on Camera-Based Document Analysis and Recognition, 2007, pp. 181–188.

[12] J. Liang, D. DeMenthon, and D. Doermann, "Geometric rectification of camera-captured document images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, pp. 591–605, April 2008.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

[13] H. Cao, X. Ding, and C. Liu, "A cylindrical surface model to rectify the bound document," in International Conference on Computer Vision (ICCV), 2003.
[14] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: Reading text in uncontrolled conditions," in IEEE International Conference on Computer Vision (ICCV), Dec 2013, pp. 785–792.
[15] L. O'Gorman, "The document spectrum for page layout analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, Nov 1993.
[16] H. I. Koo and D. H. Kim, "Scene text detection via connected component clustering and nontext filtering," IEEE Trans. Image Process., vol. 22, no. 6, pp. 2296–2305, June 2013.
[17] J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in British Machine Vision Conference (BMVC), 2002, pp. 384–393.