International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 12, Issue 5 [May. 2023] PP: 17-23

Segmentation-Based Scene Text Recognition on Low Quality Natural Scene Images

Kiptanui Linus, Prabhakar C.J.
Department of Computer Science, Kuvempu University, Karnataka, India.
Corresponding Author:psajjan@yahoo.com

ABSTRACT: *Recognition of characters in natural scene text images remains a challenging task in low quality images due to various interfering factors such as blur, low resolution, and contrast. In this paper, we propose a novel segmentation based scene text recognition technique to handle the issue of low quality scene text images. We proposed to adopt a robust Graph Cut based segmentation technique that utilizes Markov Random Field (MRF), Maximally Stable Extremal Regions (MSER)], Canny edge detector and Stroke Width Transform to accurately generate the required seed pixels. We then applied graph cut technique to segment the text characters in the image. We performed feature extraction and classification using Co-occurrence of Histogram of Oriented Gradients (Co-HOG) and support vector machine (SVM) respectively. Finally, we did character recognition using the conventional Optical Character Recognition (OCR). Our method is more robust compared to the existing methods evident through experiments conducted on two popular datasets such as ICDAR 2003 and IIIT5K dataset.*

Keywords: *Scene Text, Text Recognition, OCR, Low quality images, Text Detection, Deep Learning*

Date of Submission: 20-04-2023 Date of acceptance: 03-05-2023

I. INTRODUCTION

The recognition of scene text has become a very active research area in computer vision recently. This comes as a result of the availability of vast amount of text images capture daily using hand held cameras and smart phones. These images containing texts carries high semantic information, which when properly utilized through proper recognition which give us useful information. However, the recognition of these scene texts is far more challenging compared to texts in scanned documents. This is due to the unconstrained conditions in natural environment where these images are taken. These challenges include; complex backgrounds, uneven illumination, shadows, occlusion, orientations, text fonts and styles, among others. All these factors make the process of scene text recognition a challenging one. Proper segmentation of scene texts improves largely the recognition accuracy. Therefore, there is a great need of introducing techniques that are able to segment scene texts more accurately, in order to improve the recognition rate. Recognition of scene text is of great importance in understanding the image and even its surroundings, this is important to both humans and computers [1].

Scene Text is a text that appears in an image taken from a natural environment using a camera. Whereas, recognition is the process of trying to identify various texts that are found in the captured image. Scene text recognition therefore, is a computer vision task that tries to make a computer to read or identify the various texts found in the image, According to a study, one trillion photos were captured in the year 2015, and this number keeps increasing at a very high rate. In the context of such large data collections that continue to grow, there are many challenging problems like recognizing and retrieving relevant content. Text in the scene images can play a crucial role in understanding images and its surrounding environment. Majority of people in most cases focus on the text in an image more than any other objects. Therefore, scene text recognition can be applied in various real time vision based applications in various instances of life including; language translation, navigation of domestic robots, number plate recognition, autonomous vehicles, domestic robot navigation, visually impaired reading aids and industrial automation and particularly in the field of logistics wherein images captured by cameras are mostly curved, distorted, and have low resolution. Given the rapid growth of camera-based applications readily available on mobile phones, understanding scene text is more important than ever, real time applications for scene text recognition are becoming increasingly popular. This is related to the problem of optical character recognition (OCR), which has a long history in the computer vision community. However, the success of (OCR) systems is largely restricted to text from scanned documents that in most cases have good image quality. Scene text exhibits a large variability in appearance and can prove to be challenging even for state-of-the-art (OCR) methods (Fig 1.). Many scene-understanding methods recognize objects and regions like roads, trees, sky in the image successfully, but tend to ignore the text on signboards. It is, therefore, important to recognize such kind of text in the natural environment for better

understanding of the world around us. During recognition, the text is assumed that it has been detected and its location has been known using various methods. In this paper, we focus on scene text recognition.

(a) (b) (c) (d)

Fig.1 Text in scene images from SVT Dataset (a) High contrast (b) Low contrast (c) Oriented text (d) presence of noise.

Most of the methods available for scene text segmentation and recognition do not produce desired results due to poor segmentation procedures that will also affect the recognition rate of the entire system. Mishra et al. [2] proposed a segmentation method using Gaussian Mixture Models (GMMs), the major drawback of this method is its inefficiency to correctly identify the character seeds when edges are broken Huge amount of segmentation techniques have been introduced for segmenting document text images [3][4][5][6]. The same techniques cannot be used to segment scene text images due to its various challenges and complexities. Therefore, there is an urgent need to come up with techniques and methods that can efficiently handle this challenging task. To handle scene text images, Shi et al. [7] develop a system that uses Conditional Random Field (CRF) to solve the issue of text recognition. Mishra et al. [8] used an energy minimization approach (CRF) to recognize words in the image. They utilize the unary and pair wise functions to get the optimal word. Wang et al. [9] proposed a technique for segmenting scene text using Markov Random Fields (MRF) and strokelets to recognize text in the image.

Pan et al. [10] proposed a supervised CRF that has to discriminate between the text and non-text components. Bissacco et al. [11] proposed a graph cut technique together with Markov random field (MRF) that was able to discriminate between the text and non-text components through minimizing MRF and energy terms. .

Neumann et al. [12] proposed a photo OCR that has been trained using synthetic data for text recognition. Lee et al. [13] introduced a photo OCR that combined both CNN and RNN. Neumann et al. [14] proposed a photoOCR that uses classifier to identify character regions and recognize the characters. Liu et al [15] proposed a confidence map of text generated by finding the average edge densities, orientations and variance by the use of local windows. Mancas et al. [16] proposed a technique where the color and spatial information are put together using selective clustering to identify regions with similar colors. After which Log Gabor is called for segmentation. Kumar el al. [17] used closely packed vertical edges to detect the text regions. Neumann et al. [18] proposed a technique of text segmentation using MSER to identify the text candidates and employed SVM for discrimination between text and non-text. .

Considering the above challenges in text segmentation, we proposed a robust Graph Cut[19][20] based segmentation technique that utilizes Markov Random Field (MRF) [21], Maximally Stable Extremal Regions (MSER) [22], canny edge detector and Stroke Width Transform (SWT) [23], to accurately generate the required seed pixels. We then applied graph cut technique to segment the text characters in the image. Our method is more robust compared to the existing methods in that, it combines the robustness of MRF, MSER, Canny edge detector to select the correct seed pixels before passing it to SWT for stroke detection and later segmentation and character recognition by graph cut and Optical Character recognition respectively.

.

II. PROPOSED METHOD

The flow diagram proposed framework is shown in the Fig 2. Our proposed framework is a scene text recognition system that takes a text image as input and performs character segmentation on the image using the combination of Markov Random Field (MRF) [21] and graph cut technique [19] [20] to segment the text characters in the image. In generating seed pixels, we adopted Maximally Stable Extremal Regions (MSER) [22], canny edge detector and Stroke Width Transform (SWT) [23] to accurately capture the required seed pixels which will boost the accuracy of segmentation by graph cut. We performed feature extraction and classification using Co-occurrence of Histogram of Oriented Gradients (Co-HOG) [24] and support vector machine (SVM) [25] respectively. Finally, we did character recognition using the conventional Optical Character Recognition (OCR) [26].

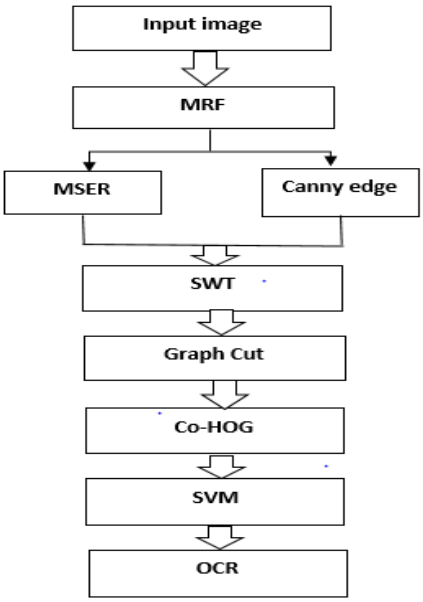

Fig 2. The flow diagram of our proposed framework

Scene character segmentation using graph cut follows the same procedure as the segmentation of images using graph cut model [19][20]. In every input image a graph G = (V, N). Where, V and N represent the set of nodes and edges respectively. To get good segmentation, Gibbs energy *E* is used which is guided by the color consistency of foreground, background and neighborhood color consistency. Gibbs energy *E* is defined as

$$E(X) = \sum_{i \in V} E_1(x_i) + \lambda \sum_{(i,j) \in N} E_2(x_i, x_j) \quad (1)$$

where $X = \{x_i\}$ is the set of binary labels in the image. $E_1(x_i)$ is the data energy representing unary cost and $E_2(x_i, x_j)$ is the smooth energy it encodes pairwise cost when the neighboring nodes i *and* j assumes the labels x_i and x_j respectively. λ *i*s a weighted term between data energy and smooth energy. Data energy measures the color difference of pixels on foreground and background seed pixels. It is defined as follows:

$$E_1(x_i) = \begin{cases} \frac{D_i^T}{D_i^T + D_i^B} & if \;\; x_i = 1 \\ \frac{D_i^B}{D_i^T + D_i^B} & if \;\; x_i = 0 \end{cases} \quad (2)$$

where, *T* and *B* represents text and background respectively. D_i^T and D_i^B represents the color difference between the text seeds and the background seeds. The data energy assigns a label "1" to a pixel if it has the same color as the seed pixels the opposite is also true. In the case of complex backgrounds, data energy alone cannot get good segmentation results. Therefore, smoothness energy is introduced.

Smoothness energy gives the neighboring pixels the same label if they have the same color. The pixels along the boundary, the smoothness energy is lower since the contrast along the text boundary is high and therefore, need to be assigned a different label. Based on the above facts, the smoothness energy between two nodes i and j can be defined as follows:

$$E_2(x_i, x_j) = \frac{|x_i - x_j|}{\|C(i) - C(j)\|^2 + \epsilon} \quad (3)$$

where, E_2 is large if the nodes i and j contains similar colors,$C(i)$ and $C(j)$ are given two different labels.($|x_i - x_j| = 1$) Ensures the denominator does not become zero. On the other hand, Gibbs energy *E* is obtained based on the Data and Smoothness energies and minimized effectively by the graph cut method. Seed pixels are required in order to compute the data energy. In this paper, we proposed a three-step seed pixel generation using Maximally Stable Extremal Regions (MSER). Canny edge detector and Stroke Width Transform (SWT) which have proved to be very effective in generating accurate seed pixels

The MSER is a technique that is used to detect blobs in an image. It extracts a number of covariant regions from an image, these regions remains almost the same through a number of thresholds. All pixels inside the MSER

region have either higher or lower intensity values as compared to other regions outside the MSER. After we obtained the Maximally Stable extremal Regions, we applied a canny edge detector to generate an edge map around the text candidates. This improves the accuracy of obtaining the seed pixels. The canny edge detector is able to robustly detect text edges by a single response, it gives one pixel wide ridges as the output during detection. The combination of MSER and canny edge detector gives edge-enhanced results, which improves the results of the SWT. The major task of the Stroke Width Transform(SWT) is to compute the width of the most likely stroke containing the pixel. The input into this phase, stroke width transform is the merged results of MSER and the canny edge detector. In the previous work by Shangxuan et al. [27], they used the Stroke Feature Transform, which did not capture adequate pixel seeds due to its inaccurate edge map. In this work, we propose a robust technique that combines MSER and canny edge detector to produce a more accurate edge map and in addition, the SWT ensures that the selected pixel seeds are indeed the strokes of the characters. The stroke width transform works as follows

1. The initial value of each element in SWT is set to ∞
2. Pixel direction follows the ray $r = p + n \cdot d_p, n > 0$
3. The gradient direction d_q at pixel q. if d_q is opposite to $d_p(d_q = -d_p \pm \pi/6)$ then SWT= |p-q| where $\theta = 0$, $\theta = 45$, $\theta = 90$, and $\theta = 135$ and d_p++ else discard the ray.

To find the candidate characters, any two neighboring pixels can be grouped together if they have the same stroke width and any two neighboring pixels can be grouped together if their SWT ratio is not greater than 3.0 this allows pixels with rather small variations to be grouped together [28]. To identify the background seeds, we adopted the technique in [29], where we utilize the Reverse Edge Direction (RED). Having the text edges from the previous methods, for each pixel p with the direction d_p the ray is set to the opposite direction $\pi - d_p$. The ray starts at pixel p and stops when it hits the edge pixel.

In graph cut segmentation, the unlabeled pixels are assigned a data energy to compute the color difference between the foreground and background pixels. During graph cut, smoothness energy is computed to facilitated accurate segmentation which is done only once since there is a clear discrimination between the text pixels and the non-text ones. We extracted features from the segmented characters using Co-occurrence of Histogram of Oriented Gradients (Co-HOG). Co-HOG descriptor able to capture the spatial information by considering the number of co-occurrence of oriented gradients among a group of pixels. It also stores their relative locations. Feature extraction can be achieved in three steps as follows:

Gradient Magnitude and Orientation Computation

A Sobel filter is used to calculate the horizontal and vertical gradient magnitude. In the case of color images, the gradient of each color channel is computed individually and the highest magnitude is selected. The gradient orientation ranges between $0^{\circ}\ to\ 180^{\circ}$.

Weighted Voting

The bi-linear interpolation is used to perform Weighted Voting between two neighboring orientation bins. However, combinations of pixel pairs with small and large gradient values are avoided as pixels with small values may have large weights if combined with pixels having large values.

Feature Vector Construction

To construct a feature vector, all the feature blocks are normalized, the normalized blocks can then be concatenated to form the feature vector.

The features we obtained from the previous step are classified using support vector machine (SVM). The main reason we used SVM is the ability to handle large amount of features extracted from the text with the help of overfitting protection. In our case, we used SVM to classify the text characters into two groups, characters and the non-characters. Lastly, we recognized the ones classified as characters using the conventional optical character recognition (OCR).

III. EXPERIMENTAL RESULTS

We performed our experiments on two public available datasets ICDAR 2003 [30] and IIIT5K [31] datasets.
ICDAR 2003 dataset contains 509 cropped scene text images. It contains images with variation in illumination, font and styles and complex background. Fig 3 shows sample images of ICDAR 2003 dataset.

Fig 3. *Natural scene text image sampleles from IIIT 5K Dataset*

IIIT 5K dataset contains 3000 cropped word images with text both in natural scenes and born-digital images. It is the most challenging dataset as it contains images with variation in illumination, font and styles and complex background. Fig. 4 shows sample images of IIIT 5Kdataset.

Fig 4. *Natural scene text image samples from IIIT 5K Dataset*

There are two metrics for performance assessment: word level recognition rate and normalized edit distance. The former is quite strict, as it requires that each character is correctly recognized. The latter is relatively looser, since it can tolerate partial local errors for each word. The performances of the algorithms are measured by word level recognition rate, which is more often used for quantitative comparison. All the recognition accuracies on the above two public datasets, obtained by the proposed model and the recent state-of-the-arts techniques including the approaches based on handcrafted features and deep models are shown in Table (1). These results prove that the proposed framework significantly improves the performance of text recognition even in the presence of the blurred, uneven illumination and different orientations of text in real scenes. From the results demonstrated in Table 1 on the ICDAR 2003, and IIIT5k dataset, it can be seen that the proposed framework outperforms existing methods with respect to character recognition accuracy.

Table (1): Comparison of the text recognition results of the proposed method with the state-of-the art methods on the ICDAR 2003, and IIIT5K datasets.

Methods	ICDAR2003	IIIT5K
Deep Features	**91.50**	**90.00**
Strokelets	**80.33**	**88.48**
Photo OCR	**90.39**	**84.00**
Discriminative Feature Pooling	**76.00**	**88.00**

The proposed method	**92.48**	**93.81**

IV. CONCLUSION

In this study, scene text recognition framework is proposed in order to overcome the challenges of scene text such contrast variations, uneven illumination blur and font size variations based on segmentation of scene text using graph cut method. Our method is more robust compared to the existing methods in that, it combines the robustness of MRF, MSER, Canny edge detector to select the correct seed pixels before passing it to SWT for stroke detection and in further stage segmentation and character recognition is done by graph cut and Optical Character recognition respectively. We presented recognition results of the proposed algorithm for scene text using two public datasets based on character recognition metric. We done segmentation of scene text from complex background of the scene text images using graph cut and other techniques which depicted outstanding results in segmentation. On the other hand, the feature extraction using CO-HoG and classification using SVM performed extremely well on the segmented scene texts. The experimental results demonstrated a great improvement in terms of character recognition and therefore, our proposed method shows an outstanding performance compared to the existing methods.

REFRENCES

[1]. T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In IEEE 12th International Conference on Computer Vision (ICCV), pages 2106–2113. IEEE, 2009.

[2]. Mishra, K. Alahari, and C. Jawahar, "An RNRF model for binarization of natural scene text," in Document Analysis and Recognition (ICDAR), International Conference on, 2011, pp. 11-16.

[3]. S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," International Journal on Document Analysis and Recognition (lJDAR), vol. 13, no. 4, pp. 303314, 2010.

[4]. W. Niblack, An introduction to digital image processing. Strandberg Publishing Company, 1985.

[5]. N. R. Howe, "A Laplacian energy for document binarization," In Document Analysis and Recognition (ICDAR), International Conference on, 2011, pp. 6-10.

[6]. B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum." in Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 2010, pp. 159-166.

[7]. Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S., Zhang, Z.: Scene text recognition using part-based tree-structured character detection. In: CVPR, pp. 2961–2968. IEEE (2013)

[8]. Mishra, A., Alahari, K., Jawahar, C.: Top-down and bottom-up cues for scene text recognition. In: CVPR, pp. 2687–2694. IEEE (2012).

[9]. Wang,Y.,Shi,C.,Xiao,B.,Wang,C.:Mrf based text binarization in complex images using stroke feature. In: ICDAR'15, pp. 821–825. IEEE (2015).

[10]. Pan, Y.F., Hou, X., Liu, C.L.: A hybrid approach to detect and localize texts in natural scene images. IEEE Trans .Image Process. 20(3), 800–813 (2011).

[11]. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: Photo OCR: Reading text in uncontrolled conditions. In: ICCV, pp. 785–792. IEEE (2013)

[12]. Neumann, L., Matas, J.: Real-time scene text localization and recognition.In:2012 IEEE Conference on CVPR, pp.3538–3545. IEEE (2012).

[13]. Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for OCR in the wild. In: CVPR, pp. 2231–2239 (2016).

[14]. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. IEEE Trans. PAMI 38(9), 1872–1885 (2016)

[15]. X. Liu and J. Samarabandu, "Multis cale edge-based text extraction from complex images," in Multimedia and Expo, International Conference on, 2006, pp. 1721-1724.

[16]. C. Mancas-Thillou and B. Gosselin, "Color text extraction with selective metric-based clustering," Computer Vision and Image Understanding, vol. 107, no. 1, pp. 97-107,2 007.

[17]. Kumar M, Lee G(2010) Automatictext locationfrom complexnatural sceneimages.In: Proceedingsof international conference on computer and automation engineering. Singapore, pp 594–597

[18]. Neumann L, Matas J (2010) A method for text localization and recognition in real-world images. In: Proceedings of the 10th Asian conference on computer vision (ACCV). New Zealand, pp 30– 35

[19]. Shangxuan Tian, Shijian Lu, Bolan Su, Chew Lim Tan: Robust Text Segmentation using Graph Cut. In 13th International Conference on Document Analysis and Recognition (ICDAR) IEEE, 2015.

[20]. C. Rother, Y. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," in ACM Transactions on Graphics (TOG), vol. 23, no. 3,2 004, pp. 309-314.

[21]. A. H. S. Solberg, T. Taxt and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 34, no. 1, pp. 100-113, Jan. 1996, doi: 10.1109/36.481897.

[22]. H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in Image Processing (ICIP), International Conference on, 2011, pp. 2609-2612.

[23]. Epshtein, B., E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010.

[24]. Shangxuan Tian , Shijian Lu , Bolan Su and Chew Lim Tan; Text Recognition using Co-occurrence of Histogram of Oriented Gradients. On 12th International Conference on Document Analysis and Recognition, 2013 IEEE.

[25]. Joachims T. (1998) Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer.

[26]. Khaoula Elagouni, Christophe Garcia, Franck Mamalet, Pascale Sébillot. Combining Multi-Scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR. 10th IAPR International Workshop on Document Analysis Systems, DAS, Mar 2012.

[27]. Shangxuan Tian, Shijian Lu, Bolan Su, Chew Lim Tan: Robust Text Segmentation using Graph Cut. In 13th International Conference on Document Analysis and Recognition (ICDAR) IEEE, 2015.

[28]. Boris Epshtein, Eyal Ofek, Yonatan Wexler: Detecting Text in Natural Scenes with Stroke Width Transform, 978-1-4244-6985-7/10/$26.00 ©2010 IEEE.

[29]. H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in Image Processing (ICIP), International Conference on, 2011, pp. 2609-2612.

[30]. Lucas,S.M., Panaretos,A., Sosa,L., Tang, A., Wong,S., and Young,R.(2003).robust reading competitions, in: International Conference on Document Analysis and Recognition,, (pp. 682–687).

[31]. Mishra, A., Alahari, K.,and Jawahar, C.(2012). Scene text recognition using higher order language priors. In: Proceedings of British Machine Vision Conference.