New Moments based Fuzzy Similarity Measure for Text Detection in Distorted Social Media Images

[1]Soumyadip Roy, [2]Palaiahnakote Shivakumara, [1]Umapada Pal, [3]Tong Lu and [4]Michael Blumenstein

[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. Email: soumyadiproy58@gmail.com, umapada@isical.ac.in.
[2]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Email: shiva@um.edu.my
[3]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China. Email: lutong@nju.edu.cn
[4]Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia. Email: Michael.Blumenstein@uts.edu.au.

Abstract. A trend towards capturing or filming images using cellphone and sharing images on social media is a part and parcel of day to day activities of humans. When an image is forwarded several times in social media it may be distorted a lot due to several different devices. This work deals with text detection from such distorted images. In this work, we consider images pass through three mobile devices on WhatsApp social media, which results in four images (including the original image) Unlike the existing methods that aim at developing new ways, we utilize the results detected by the existing ones to improve performances. The proposed method extracts Hu moments and fuzzy logic from detected texts of images. The similarity between text detection results given by three existing text detection methods is studied for determining the best pair of texts. The same similarity estimation is then used in a novel way to remove extra background or non-texts and restoring missing text information. Experimental results on own dataset and benchmark datasets of natural scene images, namely, MSRA-TD500, ICDAR2017-MLT, Total-Text, CTW1500 dataset and COCO datasets, show that the proposed method outperforms the existing methods.

Keywords: Social media images, Text detection, Moments, Correlation coefficient.

1 Introduction

Social media is a huge platform for sharing and communicating data like images and videos [1]. For example, use of cellphone cameras for capturing selfie photos and its passing on social media mobile applications, such as YouTube, Snapchat, Instagram and Facebook is very popular now. It is common that most of the time images pass through several processing stages (devices or cellphones of different configuration) before reaching its destination. This process involves downloading and uploading the images but not forwarding the image to another person. During this process, one can expect variations in the quality of images due to different configurations of cellphone

cameras and the use of social media. Besides, the process of rendering and transmission through the network creates more causes to image quality [2]. This work targets common cells or devices that are used widely. It is also true that one cannot decide exact configuration including resolution of camera and other parts of the devices as input image changes. Therefore, it is certain that the content of the image degrades. Since the process involves degradations due to different rotations, scaling and loss of quality due to other internal operations, we consider such effect as distortion rather than degradation, which covers the effect of both internal and external factors. As a result, the methods developed in the past for text detection in images may not work well for such images due to unpredictable variations in quality. It is expected inconsistent results for images of different quality. It is evident from the illustration shown in Fig. 1, where the text detection methods [3-5], namely, CTPN [3], EAST [4] and SegLink [5], which explore deep learning for text detection in natural scene images, do not detect texts accurately for all the images as shown in Fig. 1(a) -Fig. 1(c), respectively. However, the proposed method detects texts well for all the images shown in Fig. 1. Fig. 1 also shows that each text detection method reports different results for different images. Therefore, there is a gap for addressing this issue to improve text detection performance irrespective of a number of devices and social media.

(a) Text detection results of the CTPN method.

(b) Text detection results of the EAST method.

(c) Text detection results of the SegLink method.

(d) Text detection results of the proposed method.

Fig. 1. Text detection of the proposed and different existing methods for the transmitted social media images on different mobile devices. F1, F2, F3 denote images passes through respective mobile devices.

In Fig. 1, the first is the original image (input image), F1 denotes the image which passes through cellphone camera with the configuration of 1=Lenovo A7000 (Display-5.5, inch (720x1280), Processor-MediaTek MT6752m, rear camera-8MP, RAM-2GB, Storage-8GB, Battery Capacity-2900mAh,OSAndroid- 5.0). Similarly, for F2, we use Lenovo K8 Note (Display-5.50-inch (1080x1920), Processor-MediaTek Helio X23, rear camera-13MP, RAM-3GB, Storage-32GB, Battery Capacity4000mAh, OSAndroid- 7.1.1). For F3, we use cellphone camera of 3=MI Redmi 5 (Camera- 12 MP Rear Display- 14.4 centimeters (5.7-inch) HD+ Full screen display with 720x1440 pixels , Memory-4GB, Storage-64GB, Operating System and Processor- Qualcomm Snapdragon 450 octa-core processor, Battery- 3300 mA). For creating all the above three images, we use WhatsApp social media for sharing. In this work, we consider four images including the original one as the input and case study. For each image, the proposed method considers text detection results given by the respective three text detection methods. We believe that since deep learning models are powerful, one or another deep learning model should give good results for images of different qualities. Instead of relying on the results of one text detection method, we choose the results of three text detection methods. In this work, the number of devices as well as the number of text detection methods is limited to three. This is determined based on our preliminary experiments because as the number of devices and text detection methods increases, the complexity of the problem increases, which is beyond the scope of the work. Our result confirms that 3 devices and 3 text detection methods are optimal for achieving better results. This means when the images are affected by unpredictable and multiple adverse factors, one cannot expect better results by the single method. As a result, the challenge requires integrating the strengths of more than one methods.

2 Related Work

Since there are no existing methods that are developed for text detection in distorted social media images, we review the methods developed for text detection in natural scene images and low-quality images caused by poor light and affected by distortions.

For example, Tian et al. [6] proposed scene text detection under weak supervision, which aims at addressing the challenge of multi-orientation, Deng et al. [7] proposed detecting scene texts via instance segmentation, which focuses on the problem of text and non-text pixels separation, Zhou et al. [4] proposed an efficient and accurate scene text detector, which targets on arbitrary orientation text detection, Shi et al. [5] proposed detecting oriented texts in natural images by linking segments, which addresses multi-oriented text detection, and Liu et al. [8] proposed detecting curved texts in the wild, which finds a solution to complex curved text detection. He et al. [9] proposed multi-oriented and multi-lingual scene text detection with direct regression, which considers multi-lingual images for text detection. Xu et al. [10] proposed learning a deep direction field for irregular scene text detection, which considers irregularly shaped characters for detection in natural scene images. Raghunandan et al. [11] proposed multi-script-oriented text detection and recognition in video/scene/Born digital images, which focus on the causes of multi-type texts in images along with multi-oriented and multi-script.

Tang et al. [12] proposed detecting dense and arbitrarily-shaped scene texts by instance-aware component grouping, which considers the challenges of thick texts written on bottles, tin and other objects, where we can expect irregularly shaped characters without spacing between text lines much. Most of the above methods explore deep learning for text detection in natural scene images and find solutions to complex issues of orientation, multi-script, multi-type, complex background and irregularly shaped character texts. However, the methods ignore low quality images obtained by low light and distortion. To overcome this limitation, the following methods are proposed.

Huang et al. [13] proposed an end to end vessel plate number detection and recognition using deep convolutional neural networks and LSTMs. Deep learning has been used for achieving results irrespective of light conditions. Panahi et al. [14] proposed accurate detection and recognition of dirty vehicle plate numbers of high-speed applications. The method is proposed to addresses challenges of images affected by weather and lighting conditions. Wahyono and Jo [15] proposed LED dot matrix text recognition in natural scenes. The method uses Canny edge operator for obtaining edge components. Then the method explores characteristics of connected component analysis for text LED dot matrix type text detection. Shemarry et al. [16] proposed an ensemble of AdaBoost cascades of 3L-LBPs classifiers for license plate detection with low quality images. The method explores Local Binary Patten features for detecting license plates including low light images. Lin et al. [17] proposed an efficient license plate recognition system using convolution neural networks. The method finds a solution to the challenge of low light or limited light images. Mohanty et al. [18] proposed an efficient system for hazy scene text detection using deep CNN and patch NMS. In this work, the method considers images affected by haziness as poor quality images. The method considers hazy scene text detection as a classification problem, and hence it classifies hazy scene images into one class. Then it proposes a method for text detection in hazy scene images. It is noted from the above review that none of the methods use social media images or images pass through devices of different configurations for text detection. In addition, the methods consider license plate images for text detection but not images considered in this work. Therefore, it is not sure the above text detection methods work for distorted social media images.

Hence this work aims at developing a new method for text detection in distorted social media images. It is noted that [2] as the number of passes increases through different devices, image quality changes. This is due to cellphone camera of different configurations, social media and limitation of communication network systems. It is also true from the above review that deep learning is a powerful model to solve complex issues. Therefore, the same deep learning models can be used in a different way for detecting texts in distorted social media images. In addition, since text detection is preprocessing steps for text recognition and understanding, the methods developed in the past focus in addressing different challenges of text detection. This motivates us to use text detection results given by three existing text detection methods to choose the best results rather than proposing a new method. For each input image, the proposed method compares the combination of results of three text detection methods for detecting texts in distorted social media images. In order to compare the results of combination, motivated by the method [19, 20] where it is shown that Hu moments are independent of

character position, size or orientation and insensitive to variations in shapes, we explore the same Hu moments to tackle the challenges posed by images of different quality. It is true that due to variations in quality, it is not hard to find a match between the same pixels in text detection results. To handle this uncertainty, inspired by the method [21] where fuzzy logic based similarity measure is proposed for multimedia content recommendations, we explore the same with Hu moments for finding similarity between the results of the three combinations. The main contribution is exploring the combination of Hu moments and fuzzy logic based similarity for text detection to address challenges of distorted social media images. As per our knowledge, this is the first work for addressing such issues.

3 Proposed Method

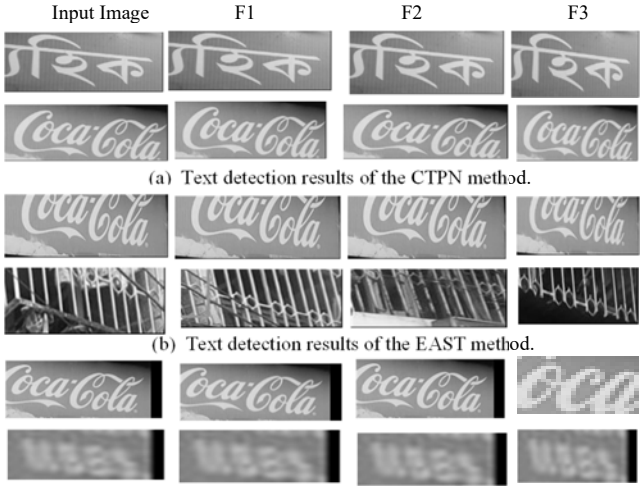
For each input image, the proposed method obtains three combinations of text detection results by the three text detection methods, namely, Method-1 & Method-2, Method-1 & Method-3, Method-2 & Method-3. Since the same image is passing through different devices and social media, there is no need to choose all the possible combinations of the results. The pairs are decided based on the correspondence of text location in text detection result images. For each pair, the proposed method extracts Hu moment based features in column and row wise. Then the proposed method performs fuzzy logic similarity estimation through person coefficient calculation for each pair using Hu moments features. Based on similarity with a certain threshold, the proposed method finds the best pair out of three pairs, which is called a candidate pair. It is true that due to variations in the quality of images, text detection methods may not fix closed bounding boxes, resulting in extra background information, chances of missing text information and detecting non-texts as texts. To overcome these challenges, the proposed method explores the same similarity measure estimation for solving the above-mentioned issues. This is valid because if a candidate pair is the same, the similarity measure estimated for the columns from left to right and right to left converges to almost zero. It is verified by the similarity measure estimation for rows from top to bottom and bottom to up. The same idea is used for solving the other three issues, which will be discussed in the subsequent sections. The above process results in a text with a closed bounding box irrespective of orientation, script, font size, shape of the characters in text line.

3.1 Text Detection for Distorted Images

As mentioned in the previous section, for text detection in input images, we prefer to choose CTPN [3], EAST [4] and SegLink [5] methods. The reason to choose the above three methods is as follows. The method called CTPN proposes connectionist text proposal networks for text detection in natural images. The connectionist text proposal network explores rich context information of images, making it powerful to detect extremely ambiguous texts. In addition, CTPN is invariant to multi-scale and multi-language without further post processing. The method EAST proposes an Efficient and Accurate Scene Text Detector (EAST) for text detection in natural scene images. The focus of the method is to detect text lines of arbitrary oriented text lines in images. For

this, EAST proposes a single neural network. The method Selina proposes Segment Linking (SegLink) for oriented text detection in natural scene images. The main idea of SegLink is to divide each text into segments and links. A segment is an oriented box covering a part of a word or text line, and a link connects two adjacent segments after confirming two segments are belonging to the same line.

The above three methods are popular and the codes are available publicly. Besides, the scope of the above three methods covers challenges of the proposed work. Therefore, we prefer to choose the above three methods for text detection. Sample text detection results given by the three text detection methods are shown in Fig. 1, Fig. 2(a) and Fig. 2(c), respectively. It is observed from Fig. 3 that CTPN gives better results for the original image, F1, F2 and F3, while EAST misses text information as well as to detect non-texts as texts. In the same way, SegLink misses low quality texts. This shows that out of the three methods, there are chances of expecting good results at least by one method. At the same time, we can also note that the methods report inconsistent results for images of different quality.



(a) Text detection results of the CTPN method.

(b) Text detection results of the EAST method.

(c) Text detection results of the SegLink method.

Fig. 2. Sample text detection results of different existing methods

3.2 Moments based Fuzzy Logic Similarity Measure for Text Detection

For the text detection results obtained by the three text detection methods, to choose the best match, we explore Hu moments and Fuzzy logic similarity measure estimation. Hu moments as defined in Equation (1) to Equation (5), give 7 values for each column and row of input images. For Hu moments, the proposed method uses triangular rule of fuzzy to estimate the membership function defined in Equation (6), where a denotes lower most value, b denotes upper most value, and m denotes midpoint. This results in a vector containing 7 fuzzy values. The vector containing fuzzy values are passed to Person

coefficient as defined in Equation (7), which estimates the similarity between two vectors of columns or rows of the pair results given by the three text detection methods.

$$M_{i,j} = \sum_x \sum_y x^i y^j I(x,y) \tag{1}$$

where *i* and *j* are integers (e.g., 0, 1, 2 ….). These moments ($m_{i,j}$) are often referred to as raw moments to distinguish them from the central moments mentioned later. Here x and y are image co-ordinates, $\bar{x}$ and $\bar{y}$ are the centroids of the group of connected pixels.

$$\bar{x} = \frac{M_{10}}{M_{00}} \quad \bar{y} = \frac{M_{01}}{M_{00}} \tag{2}$$

Central moments ($u_{i,j}$) is calculated by

$$u_{i,j} = \sum_x \sum_y (x - \bar{x})^i (y - \bar{y})^j I(x,y) \tag{3}$$

$$\text{Normalized Central Moments } (ni,j) = n_{i,j} \frac{u_{i,j}}{u_{00}^{\frac{i+j}{2}+1}} \tag{4}$$

From here we calculate the 7 vector hu moments-

$$\begin{aligned}
h_0 &= n_{20} + n_{02} \\
h_1 &= (n_{20} - n_{02})^2 + 4n_{11}^2 \\
h_2 &= (n_{30} - 3n_{12})^2 + (3n_{21} - n_{03})^2 \\
h_3 &= (n_{30} + n_{12})^2 + (n_{21} + n_{03})^2 \\
h_4 &= (n_{30} - 3n_{12})(n_{30} + n_{12})[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2] + (3n_{21} \\
&\quad - n_{03})[3(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2] \\
h_5 &= (n_{20} - n_{02})[(n_{30} + n_{12})^2 - (n_{03} + n_{21})^2] + 4n_{11}(n_{30} + n_{12})^2(n_{21} \\
&\quad + n_{03}) \\
h_6 &= (3n_{21} - n_{03})(n_{30} + n_{12})[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2] + (n_{30} - 3n_{12})(n_{21} \\
&\quad + n_{03})[3(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2]
\end{aligned} \tag{5}$$

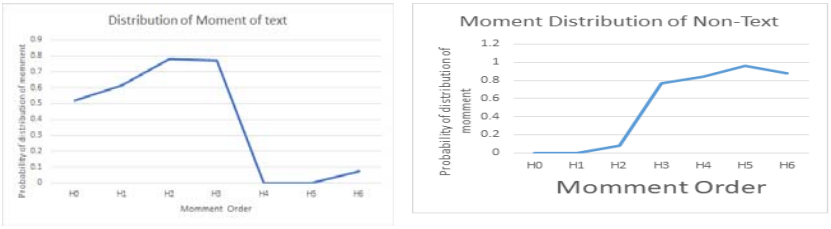
$$\begin{aligned}
u(X) = (X - a)/(m - a) \;\; if\; X < mid \\
(b - X)/(b - m)\; if\; X >= mid
\end{aligned} \tag{6}$$

where, X is the Hu Momment , a is the lower most value of Hu Momment vector, and b is the upper most value of Hu Momment vector and $mid = (a + b)/2$.

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}})(\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}})} \tag{7}$$

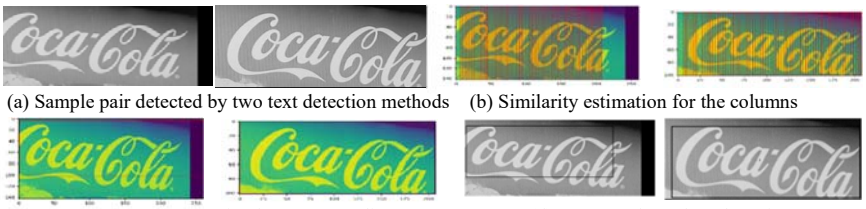
Where, *x* and *y* are two vectors containing *n* fuzzy values.

The effects of Hu moment for text and non-text results are shown in Fig. 3(a) and Fig. 3(b), where it is noted that the distribution of Hu moments for texts is smooth compared to the distribution of Hu moments of non-texts. This is the advantage of exploring Hu moments for text detection in this work.



(a) Hu moments for text (b) Hu moments for non-text

Fig. 3. The effect of Hu moments for the text and non-text



(a) Sample pair detected by two text detection methods (b) Similarity estimation for the columns

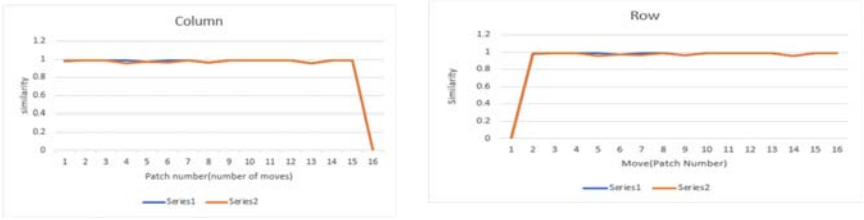
(c) Similarity estimation for the corresponding rows (d) Common boundary detection

Fig. 4. Candidate pair detection

For a sample pair given by text detection methods as shown in Fig. 4(a), the proposed method estimates the similarity measure for the columns corresponding to the first and the second images as shown in Fig. 4(b). Similarly, similarity estimation for the rows corresponding to the first and the second images is estimated as shown in Fig. 4(c). By analyzing the similarity values of columns and rows, the proposed method finds a common boundary for both the first and the second images as shown in Fig. 4(d). It is noted from Fig. 4(b) and Fig. 4(c) that the proposed method finds the columns and rows which indicate starting and ending texts in images respectively. This is valid because the similarity matches when both columns/rows have the same text information, else it does not match due to different information. This step outputs correct boundaries for texts, which is called a candidate pair out of three pairs. This step helps us to remove extra background in the pair of images. However, this step does not help us to detect missing text pairs and non-text pairs.

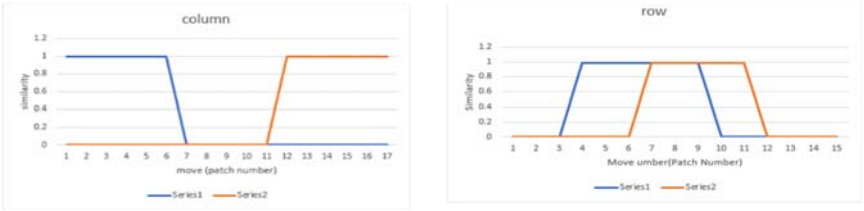
In each pair, if one contains the full text and another one misses a few characters as shown in Fig. 5(a), and if one is correct text and another result is non-text as shown in Fig. 5(c). For the above cases, the proposed method estimates similarity for columns and rows, namely, left-right, right-left for columns, and top-bottom, bottom-up for rows as shown in Fig. 5(b) for missing pair. It is observed from Fig. 5(b) that both the similarity score calculated for left-right (Series-1) and right-left (Series-2) is almost the same up to a certain point, and then there is a sudden drop at the similarity score of both. At the same time, the similarity scores calculated for rows from top-bottom (Series-1) and bottom-up (Series-2) in the same way as shown in Fig. 5(b), both the lines give almost the same values without dropping in contrast to column similarity. This indicates the missing pair.

(a) Missing pair text detection



(b) Column and row comparison

(c) Text-non-text pair detection



(d) Column and row comparison

Fig. 5. Text restoration using column and row comparison. Series-1 represent the line for left to right and Seriies-2 represents the line from right to left in case of column graphs. In row graphs, Series-1 for top to bottom and Series-2 represents bottom-up operation.

For the case of text and non-text pair shown in Fig. 5(c), the similarity graphs obtained for columns and rows do not give any correlation as shown in Fig. 5(d), where we can see Series-1 and Series-2 of both graphs behave differently. This indicates that it is a non-text pair. To restore the missing text, the proposed method finds the image which has a shorter boundary compared to the other images of the pair. The proposed method expands columns and rows of short boundary images, and then similarity scores are estimated for corresponding columns and rows. As long as texts exist while expanding, the expanding process continues and terminates when there is a big difference, which indicates the end of a text. For a text-non-text pair, we use recognition results through edit distance to remove non-text images from the pair.

4 Experimental Results

As a new work, there are no standard datasets for evaluating our proposed method. Therefore, we create our own dataset collected by different natural scene images of

arbitrarily oriented texts, multi-script and irregular shaped texts. This includes 2000 images for experimentation. All these images are used for testing the proposed and existing methods. However, for determining parameters and converging criteria, we choose 500 images randomly across databases, including a standard database of natural scene images. We divide 2000 images into four classes, namely, input image, F1 which passes through device-1, F2 which passes through device-2, and F3 which passes through device-3. In other words, each set contains 4 images. Details of the devices and configurations are described in the Introduction Section. As a result, our dataset consists of 500 sets of four classes. For this work, In order to evaluate the effectiveness and usefulness of the proposed method, we also consider the following standard datasets.

MSRA-TD500: This dataset provides 300 images for training and 200 images for testing. It includes multi-oriented texts of English and Chinese languages. **CTW1500**: This dataset provides 1000 images for training and 500 images for testing. It is created basically for evaluating curved text detection in scene images. **Total-Text**: This dataset provides 1255 images for training and 300 for testing. This is also the same as CTW1500 dataset but with more variations, which includes low resolution, low contrast and complex backgrounds. **ICDAR 2017 MLT**: This dataset provides multi-lingual text images, including 7200 training images, 1800 validation images and 9000 testing images. The dataset consists of images of 9 languages in arbitrary orientations. **COCO-Text**: This dataset is created not with the intention of text detection, hence texts in images are more realistic and there are a large number of variations compared to all the other datasets. As a result, this dataset is much more complex than the other datasets for text detection. It provides 43686 images for training, 20000 images for validation and 900 images for testing.

It is noted that in case of all the five benchmark datasets, there are no images which represent F1, F2 and F3. Therefore, we consider each image as the original one and then use the same image to create F1, F2, F3 images with the same above-mentioned devices. In this way, we create datasets for experimentation in this work. The size of testing samples for all the respective datasets is actually the number of testing samples multiplied by four. For evaluating the performance of the proposed and existing methods, we follow the standard instructions and evaluations discussed in [9]. More information about Recall (R), Precision (P) and F-Measures (F) can be found from [9]. The proposed method calculates the above measures for four classes, namely, Input image, F1, F2, F3 and Average. For Average, the proposed method considers the results of all the four type images as one image results for calculating Precision, Recall and F-measure. However, for our dataset, since there is no ground truth, we count manually for calculating measures. To show the objectiveness of the proposed method, we compare it with the sate-of-the art methods, namely, Tian et al. [3] which proposes Connectionist text Proposal Network (CTPN) for text detection in natural scene images, Zhou et al. [4] which proposes An Efficient and Accurate Scene Text Detector (EAST) for text detection in natural scene images, and Shi et al. [5] propose a Segment Linking (SegLink) based method for text detection in natural scene images. The codes of the above-methods are available publicly and popular for text detection in natural scene images because they address almost all the challenges of text detection in natural scene images.

4.1 Experiments on Distorted Social Media Images

As discussed in the previous section, quantitative results of the proposed and existing methods for Input, F1, F2 and F3 are reported in Table 1, where it is noted that the proposed method outperforms the existing methods in terms of Recall, Precision and F-measure for all the four type experiments. It is also noted from Table 1 that the existing methods are not consistent. This shows that the existing methods are not good enough to handle the distortion created by different devices, social media and networking. On the other hand, since the proposed method chooses the best results from the results of the three existing text detection methods through Hu moments based fuzzy logic similarity measures, the proposed method is capable of handling such issues. It is observed from Table 1 that as number of passes increases, the results of the existing methods change compared to the results of the input images. Some methods score better results for F2 and F3 compared to F1, while some methods do not. However, the proposed method is almost the same for all the four types of experiments. This is the advantage of the proposed method over the existing methods. In this work, we use the system with the following configuration, 8 gb Ram, Nvidia Ge force 940m Graphics card and on Ubuntu 18.03 and Tensorflow 1.3 for all the experiments. Our experiments show that the average processing time of each image taken by the proposed method for the datasets, namely, MSRA TD500, CTW 1500, Total Text, ICDAR 2017 MLT and MS COCO dataset is between 3.3 to3.5 seconds.

Table 1. Performances of the proposed and existing methods on our dataset

Methods	Input			F1			F2			F3			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EAST[4]	65.0	60.0	62.4	63.0	65.0	63.9	62.0	66.0	63.9	64.5	66.5	65.4	60.0	64.5	62.1
Seglink[5]	81.0	75.0	77.8	83.0	74.0	78.2	85.0	76.0	80.2	86.0	76.5	80.9	80.0	74.2	77.0
CTPN[3]	83.0	85.0	83.9	83.0	82.0	82.5	83.5	81.0	82.2	81.0	84.0	82.4	80.0	83.0	81.4
Proposed method	**85.0**	**89.0**	**86.9**	**87.0**	**89.0**	**87.9**	**87.0**	**89.0**	**87.9**	**87.0**	**89.0**	**87.9**	**87.0**	**89.0**	**87.9**

4.2 Experiments on Benchmark Dataset of Natural Scene Images

Sample qualitative results of the proposed and existing methods for images of MSRA-TD500 dataset are shown in Fig. 6(a)-Fig. 6(d), respectively. It is observed from Fig. 6 that as discussed above, the existing method does not report consistent results for F1, F2 and F3 compared to the input, while the proposed method detects all the texts correctly. Quantitative results of the proposed and existing methods for natural scene text datasets, namely, MSRA-TD500, CTW1500, Total-Text, IDAR2017-MLT and COCO, are reported in Table 2-Table 6, respectively. Table 2-Table 6 show that the proposed method is better than the existing methods for all the four type experiments in terms of recall, precision and F-measure for all the datasets. This shows that the proposed idea can be used for solving still more complex issues by utilizing the existing methods results. However, it is noted from Table 5 and Table 6, the existing methods report poor results compared to the other datasets. This is because the datasets, namely8, ICDAR2017-MLT and MS-COCO are much more complex than other datasets as pointed out earlier. This leads to obtaining poor results for the proposed method because the performance of the proposed method depends on the success of the existing methods. If all the three existing methods report poor results, obviously, the proposed method

reports poor results. This indicates that choosing the existing methods is also important for achieving better results for complex issues. Overall, the proposed work gives a message that the combination of text detection results can cope with the distortion/challenges caused by different devices, social media and networks.

Table 2. Performance of the proposed and existing methods on MSRATD-500 dataset

Methods	Input			F1			F2			F3			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EAST[4]	87.3	67.4	76.0	90.0	60.0	72.0	85.0	62.0	71.7	89.0	61.0	72.3	88	61.4	72.3
Seglink[5]	86.0	70.0	77.0	86.5	74.0	79.7	83.0	76.0	79.3	85.0	75.0	79.6	86.5	73.0	79.1
CTPN[3]	83.0	73.2	77.8	80.0	79.0	79.4	83.5	80.0	81.7	82.0	81.0	81.5	80.5	77.5	78.9
Proposed method	**89**.0	**80**.0	**84.26**	**84**.0	**82**.0	**82.9**	**85**.0	**84**.0	84.4	**86**.0	**83**.0	84.4	**89**.0	**82**.0	85.3

(a) Text detection of the CTPN method

(b) Text detection of the EAST method

(c) Text detection of the SegLink method

(d). Text detection of the proposed method

Fig. 6. Text detection for benchmark dataset of MSRATD-500 dataset

Table 3. Performance of the proposed and existing methods on CTW1500 dataset

Methods	Input			F1			F2			F3			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EAST[4]	78.7	49.1	60.4	79.0	55.0	64.8	77.0	53.5	63.1	79.5	55.5	65.3	78.0	53.5	63.4
Seglink[5]	42.3	40.0	40.8	50.0	50.0	50.0	49.0	53.0	50.9	49.0	51.0	49.8	49.0	50.0	49.4
CTPN[3]	70.0	**88**.0	77.9	75.0	88.0	80.9	75.5	89.0	81.7	76.0	86.0	80.6	73.0	88.0	79.8
Proposed method	**81**.0	**88**.0	**84.3**	**83**.0	**88**.0	**85.4**	**83.5**	**90**.0	86.6	**82**.0	**89**.0	85.3	**83**.0	**88**.0	85.4

Table 4. Performance of the proposed and existing methods on Total Text dataset.

Methods	Input	F1	F2	F3	Average

	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EAST[4]	50.0	36.2	42.0	55	32.0	40.4	55.5	31	39.7	53.0	35.0	42.1	53.7	33.0	40.8
Seglink[5]	30.3	23.8	26.7	30.5	25.0	27.4	30.0	25.5	27.5	32.0	26.0	28.6	30.0	25.0	27.2
CTPN[3]	82.7	74.5	78.4	82.7	76.0	79.2	84.0	76.0	79.8	83.0	77.0	79.8	81.0	75.0	77.8
Proposed method	**86.0**	**77.0**	**81.2**	**86**	**80**.0	**82.8**	**87**.0	**79**.0	82.8	**85**.0	**80.5**	82.6	**86**.0	**80**.0	82.8

Table 5. Performance of the proposed and existing methods on ICDAR 2017 MLT dataset.

Methods	Input			F1			F2			F3			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EAST[4]	44.4	25.5	32.4	46.0	27.0	34.0	45.0	26.0	32.9	45.0	26.0	33.0	45.5	26.5	33.4
Seglink[5]	67.7	34.7	45.9	66.0	36.0	46.5	66.0	35.5	46.1	66.0	37.0	47.1	66.0	35.5	46.1
CTPN[3]	71.1	55.5	62.3	73.0	56.0	63.3	71.0	55.5	62.3	71.0	55.5	63.7	72.5	55.0	62.5
Proposed method	**74**.0	**56**.0	**63.7**	**76**.0	**59**.0	**66.4**	**76.5**	**58**.0	65.9	**76.5**	**60**.0	67.6	**76**.0	**59**.0	66.4

Table 6. Performance of the proposed and existing methods on MS-COCO dataset

Methods	Input			F1			F2			F3			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EAST[4]	30.0	30.0	30.0	32.0	26.0	28.6	32.5	26.5	29.1	33.0	25.0	28.4	31.0	27.0	28.8
Seglink[5]	60.0	55.0	51.4	62.0	60.0	60.9	61.0	59.0	59.9	60.0	60.0	60.0	59.0	58.0	58.5
CTPN[3]	55.0	55.0	59.5	59.0	60.0	59.4	57.0	59.0	57.9	58.0	61.0	59.4	58.0	58.5	58.2
Proposed method	**62**.0	**60**.0	**63.7**	**61**.0	**61**.0	**61**.0	**62**.0	**61.5**	61.7	**63**.0	**62**.0	62.5	**62**.0	**61**.0	61.5

5 Conclusion and Future Work

In this work, we have proposed a new idea of exploring the combination of text detection results of the existing methods for overcoming the problems of distortion created by different devices, social media and networks. To choose the best pair from the results of the existing text detection methods, we have combined Hu moments and fuzzy logic similarity measure in a new way for solving the issues. Similarity measure estimation is used for removing excessive background, restoring missing text information and eliminating false positives. Experimental results on our datasets and different datasets of natural scene images show that the proposed method is better than the existing methods, and it is consistent for four type experiments in contrast to the existing methods. However, it is noted from the experimental results that choosing the number of devices, social media, and existing text detection methods are important in achieving better results. Therefore, our future work shall be providing a roadmap for choosing an optimal number for the above.

References

1. D. Ghadiyaram, J.Pan, A. C. Bovik, A. K. Moorthy, P. Panda and K. C. Yang, “In-Capture mobile video distortions: A study of subjective behaviour and objective algorithms”, IEEE Trans. CSVT, 28, 2018, pp 2061-2077.
2. D. Ghadiyaram, J.Pan, A. C. Bovik, “A subjective and objective study stalling events in mobile streaming videos”, IEEE Trans. CSVT, 29, 2019, pp 183-197.
3. Z. Tian, W. Huang, T. He, P. He and Y. Qiao ”Detecting Text in Natural Image with Connectionist Text Proposal Network”. In Proc. ECCV, 2016, pp 56-72.

4. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “East: an efficient and accurate scene text detector,” in Proc. CVPR, 2017, pp. 2642–2651.
5. B. Shi, X. Bai, and S. Belongie, “Detecting oriented text in natural images by linking segments”. In Proc. CVPR, 2017, pp. 3482–3490.
6. S. Tian, S. Lu and C. Li, “WeText: Scene text detection under weak supervision”, In Proc. CVPR, 2017, pp 1501-1509.
7. D. Deng, H. Liu, X. Li and D. Cai, “PixelLink: Detecting scene text via instance segmentation”, In Proc. AAAI, 2018.
8. Y. Liu, L. Jin, S. Zhang and S. Zhang, “Detecting curve text in the wild: New dataset and new solution”, arXiv:1712.02170, 2017.
9. W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, “Multi-Oriented and multi-lingual scene text detection with direct regression”, IEEE Trans. IP, 2018, pp 5406-5419.
10. Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang and X. Bai, “TextField: Learning a deep direction field for irregular scene text detection”, IEEE Trans. IP, 2019 (to appear).
11. K. S. Raghunandan, P. Shivakumara, S. Roy, G. Hemantha Kumar, U. Pal and T. Lu, “Multi-script-oriented text detection and recognition in video/scene/born digital images”, IEEE Trans. CSVT, 29, 2019, pp 1145-1162.
12. J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu and X. Bai, “Detecting dense and arbitrary-shaped scene text by instance-aware component grouping”, Pattern Recognition, 2019 (to appear).
13. S. Huang, H. Xu, X. Xia and Y. Zhang, “End-to-end vessel plate number detection and recognition using deep convolutional neural networks and LSTMs”, In Proc. ISCID, 2018, pp 195-199.
14. R. Panahi and I. Gholampour, “Accurate detection and recognition of dirty vehicle plate numbers for high speed applications”, IEEE Trans. ITS, 18, pp 767-779, 2017.
15. Wahyono and K. Jo, “LED dot matrix text recognition method in natural scene”, Neurocomputing, 151, 2015, pp 1033-1041.
16. M. S. A. Shemarry, Y. Li and S. Abdulla, “Ensemble of adaboost cascades of 3L-LBPs classifiers for license plated detection with low quality images, ESWA, 92, pp 216-235, 2018.
17. C. H. Lin, Y. S. Lin and W. C. Liu, “An efficient license plate recognition system using convolutional neural networks”, In Proc. ICASI, pp 224-227, 2018.
18. S. Mohanty, T. Dutta and H. P. Gupta, “An efficient system for hazy scene text detection using a deep CNN and patch-NMS”, In Proc. ICPR, 2018, pp 2588-2593.
19. M. Nawali and S. Liao, “A new fast algorithm to compute continuous moments defined in a rectangular region”, Pattern Recognition, 89, 2019, pp 151-160.
20. M- K. Hu,” Visual pattern recognition by moment invariants”, IRE Transactions on Information Theory, 1962, pp 179-187.
21. S. Kant, T. Mahara, V. K. Jain and D. K. Jain, “Fuzzy logic similarity measure for multimedia contents recommendation”, Multimedia Tools and Applications, 78, 2019, pp 4107-4130.