Journal of Scientific & Industrial Research
Vol. 80, September 2021, pp. 817-827

Scene Text Extraction using Convolutional Neural Network with Amended MSER

Aparna Yegnaraman* and Valli S

Department of Computer Science and Engineering, College of Engineering, Guindy, Anna University,
Chennai 600 025, TN, India

Received 06 February 2021; revised 15 August 2021; accepted 29 August 2021

Content in the text format helps to communicate the relevant and specific information to users meticulously. A beneficial approach for extracting text from natural scene images is introduced which employs amended Maximally Stable Extremal Region (a-MSER) together with deep learning framework, You Only Look Once YOLOv2 network. The proposed system, a-MSER with Scene Text Extraction using Modified YOLOv2 Network (STEMYN), performs remarkably well by evaluating three publicly available datasets. The method a-MSER is used to identify the region of interest based on the variation of MSER. This algorithm considers intensity changes between text and background very effectively. The drawback of original YOLOv2, the poor detection rate for small-sized objects, is overcome by employing 1 × 1 layer with image size enhanced from 13 × 13 to 26 × 26. Focal loss is applied to improve upon the existing cross entropy classification loss of YOLOv2. The repeated convolution layer in the steep layer of the original YOLOv2 is removed to reduce the network complexity as it does not improve the system performance. Experimental results demonstrate that the proposed method is productive in identifying text from natural scene images.

Keywords: Convolution layer, Deep learning framework, Focal loss, Maximally stable extremal regions, YOLOv2

Introduction

Text has been regarded as a symbolic system of communication for more than thousands of years. It is an invention of mankind that reveals human thoughts and emotions and carries accurate and valuable high-level semantics. Text ingrained in image and video encapsulate an abundant information source for different applications like image-based geo-location, mobile visual searches, content-based image retrieval, and automatic sign translation. Scene text extraction is still a major challenge due to the discrepancy in text size and color, complicated background, and unrestrained illumination, etc. At present, scene text detection has become a compelling aspect of computer vision and pattern recognition techniques, as well as an active research hotspot in the field of document analysis and recognition. In this work, we propose a novel text detection algorithm that employs MSERs[1] to detect region of interest (ROI). Here a new method a-MSER is proposed which takes into account the intensity changes between text and background very effectively. Once the ROIs are identified based on a-MSER, then they are used as input to the fully Convolutional Neural Network (CNN) which we call it as STEMYN. The typical text detection methods include many segments functioning simultaneously with numerous processing steps. Heuristic guidelines and criterion are to be construed and tuned. The speed of detection is significantly lower and very hard to obtain satisfactory outcomes. With the advent of convolutional neural network (CNN)[2–6] based object detection frameworks, there is a considerable effect in scene text extraction which apparently exhibited excellent enhancement in accuracy and fast detection. R-CNN[7] involved CNN to extract features, setting a trend in the object-detection framework and achieved exemplary outcomes than the state-of-the-art methods back then. But R-CNN detection speed was not fast and this contributed to the launch of its derivatives such as Fast RCNN[8], Faster RCNN and MaskRCNN.[9] These approaches facilitate the generation of region proposal and hence speed of object detection is appreciably enhanced. Nevertheless, all these methods have two-stage framework that made them more complicated and slower than the regression-based methods.

Redmon *et al.* introduced YOLO[6], a regression-based approach that employs single CNN adept at predicting bounding boxes along with class probabilities. But YOLO has a low recall and higher

*Author for Correspondence
E-mail: aparna.yegnaraman@gmail.com

localization errors than the region-based methods. The second version of YOLO is YOLOv2[(4)] which is created with the aim of improving the accuracy and simultaneously its speed. Although comparatively faster, YOLOv2 has some limitations in detecting small-sized objects. In this work, these limitations are overcome and a modified model STEMYN is presented. Non-Maximal Suppression (NMS) is applied to remove the redundant bounding boxes for a particular text occurrence. Then, the required text portions present in natural scene images are localized by bounding box and obtained as output. In a nutshell, we can say that STEMYN with a-MSER could produce relatively satisfactory results on various text detection ICDAR benchmark datasets, including ICDAR 2013[(10)], ICDAR 2015[(11)] and MSRA-TD500.[(12)] The contributions of this work are:

- a-MSER is applied to identify the Region of Interest (ROI) taking into account intensity variations between text and background effectively.
- The original YOLOv2 could not detect small-sized objects. This has been overcome by introducing a 1×1 layer to the existing network with image size enhanced from13×13 to 26×26.
- The complexity of the network is reduced by removing the repeated convolution layerin steep layers as it does not improve the system performance.
- Focal loss is introduced instead of the cross-entropy classification loss to improve the system performance.

Related Works

Detecting text from natural scene images is a prominent research area being worked upon for decades together. Many authors have come up with different ideas accompanied with lot of comprehensive surveys.[13,14]

Classic Methods

The classic methods for scene text extraction are sliding window based and Connected Component (CC) based methods.

Sliding window based methods[15,16] are most favoured. These methods scan every image patch with the help of sliding window exploiting the texture property of text. Machine learning algorithms are employed for categorization of text and non-text. Generally, the framework seems to be simple here but the classification is quite complex due to lot of calculations involved in segregating the various windows.

The major approaches employed in CC based methods[17,18] are Maximally Stable Extremal Regions (MSER) and Stroke Width Transform (SWT). The basic assumption behind these methods is, characters consist of one or more connected components and this property is exploited to explore individual character or stroke. The features employed in detecting text are colour, edge, gradients or combination of these and carry out supplementary substantiation for false positives removal.

MSER based CC methods[1] have performed very well when applied to ICDAR datasets. There are many works based on this. But still, there are many issues to be addressed. The review of various works on text detection based on the MSER is displayed in Table 1. All these existent methods for MSERs pruning still have room for improvement in terms of accuracy and speed. Our method a-MSER is definitely better than these approaches as it could handle texts with excessive blurs, no contrast against the background, non-uniform illumination (reflecting surfaces) and unusual fonts.

Deep Learning Based Methods

Text detection has taken a new direction after the arrival of deep learning-based approaches for object detection and semantic segmentation. They can be primarily classified into three methods: segmentation based, end-to-end based and regression based.

End-to-end Methods perform not only localization but also recognition of text from images. The prominent detection and recognition methods were linked together by He *et al.*[19] and Liu *et al.*[20] and trained them in an end-to-end manner. Lyu *et al.*[21], based on the inspiration from the model created by He *et al.*[9], proposed an end-to-end trainable network through semantic segmentation. Wei *et al.*[22] proposed end-to-end text spotting with text detector and recognizer based on spatial attention bidirectional Long Short-Term Memory (SA-BiLSTM) decoder. In these end-to-end methods, the accuracy in detection is improved on the basis of the recognition results, as the technique would be on to train detection and recognition modules collectively.

Segmentation Based method is one of the prevailing methods in text detection. The texts that are close to each other were differentiated by Pixel link[23] by discriminating pixel connections between various instances of text. Lyu *et al.*[24] employed corner point localization of text bounding boxes and segmentation of text region in relative positions for extracting scene text. A kernel-based framework, namely, Progressive Scale Expansion Network (PSENet) was suggested by Wang *et al.*[25] to locate arbitrarily shaped text instances as it performed pixel level segmentation. Xie *et al.*[26] came up with a Supervised Pyramid Context Network that employed instance segmentation framework and context information to detect arbitrarily shaped text. Dai *et al.*[27] amassed the improved text features and input them into box refinement network and box-aware context-based text segmentation module to acquire more accurate text boundaries.

The methods described here are established on proposal-free semantic and instance segmentation technique. The performances of these approaches are highly influenced by the robustness of segmentation results.

Regression Based Method is comprehensible, simple and inspired from current developments on object detection frameworks. Regression is applied to get the bounding boxes on the proposals generated. Gupta *et al.*[33] evolved Fully Convolutional Regression Network (FCRN) that is adept at predicting bounding boxes in an image, but the drawback was dependency on classifier and regression steps to remove false positives. Textboxes[34] and its next variant Textboxes++[35]changed the anchor scales and shape of convolution kernels to accommodate to the various text aspect ratios, but failed in case of dense and large angle texts. Liao *et al.*[36] proposed rotation sensitive regression detector with rotation invariant features, but failed to deal with vertical text lines and text line with large character spacing. The multi oriented scene text was detected by Ma *et al.*[37] with the help of rotation region proposals but failed while detecting extremely small text instances and long text lines. Shi *et al.*[38] adopted SSD[3] to detect oriented text and predicted text segments which are linked into complete instances using the linkage prediction. This method also faced the same failure as the previous one. Our method stands higher in the context of detecting texts with different font sizes, large spacing between characters and long text lines when compared to all these approaches.

Methodology

The framework of the proposed method is shown in Fig. 1 wherein the text regions of the input image are identified using a-MSER and sent to STEMYN model which is discussed in the later part of this section.

a-MSER

Mixed pixels are those pixels that lie between bright background and dark regions, and vice-versa. The proposed method a-MSER handles these pixels effectively by finding out the stability of an extremal region properly. MSERs are controlled by a parameter delta (Δ), which controls how the stability is calculated. The value for this parameter delta (Δ) is chosen from the intensity profile of the given image. For some images, regions might be detected with a lower Δ and for some other images with a higher Δ. A region is stable if it has small variation. This MSER

Table 1 — Summary of different works pertaining to text detection based on MSER

Author	Method Used	Drawbacks
Chen *et al.*[28]	Produced edge enhanced MSERs based on the complementary properties of Canny edge detector combined together with MSERs	Failed due to excessive blurs
Neumann *et al.*[29]	Suitable Extremal Regions (ERs) are selected in real-time by a sequential classifier Adaboost based on features specific to text and this ER method is preferred over MSER due to reduced memory footprint	Failed when there are characters with no contrast, single character and multiple characters joined together
Yin *et al.*[30]	Proposed pruning of MSER trees by applying two algorithms based on parent-children elimination operation, namely linear reduction and tree accumulation algorithm	Failed to deal with very complex background, non uniform illumination (with reflective surfaces), highly blurred text and unusual fonts
Huang *et al.*[31]	Employed a deep CNN model to learn high level features from the MSER detector	Failed when there are strong masks covering texts and also when there is no strong text information and easily confused with the background
He *et al.*[32]	Proposed an improved version of Huang *et al.*[31], a Text-CNN model which provides additional supervised information that would aid the model with more specific text features, from low-level region segmentation to high-level binary classification	Failed to deal with extremely ambiguous text information and easily confused with its background

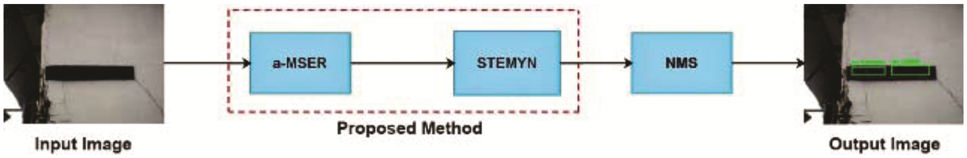

Fig. 1 — Framework of the Proposed Method for text detection

algorithm detects "maximally stable" regions that have a lower deviation than the regions one level above or below. The inverse of the relative area variation of the region G when the intensity level is increased by Δ is the stability of the extremal region G.[39] The variation is given in Eq. (1).

$$\frac{|G(+\Delta)-G|}{|G|} \qquad \ldots (1)$$

In Eq. (1), $|G|$ denotes the area of the extremal region G, G(+Δ) is the extremal region +Δ levels up which contains G and $|G(+\Delta)-G|$ is the difference in the area of the two regions.

For illustration purposes, we will consider an input image as shown in Fig. 2(a). The intensity profile of this image is given in Fig. 2(b). The delta (Δ) value for this image is chosen as the mean of the intensity divided by 10 which results in Δ = 11. The effect of delta (Δ) on this image is shown in Fig. 3 by gradually increasing it from 1 to 40. We see from Fig. 3 that at Δ = 11 the text regions are clear against the background. We also see that as we increase Δ, fewer and fewer regions are detected until finally at Δ = 40, there is no region G which is stable at G (+Δ).

The input image is converted to grayscale from RGB. MSERs are extracted for both dark-on-bright and bright-on-dark regions of the input grayscale image. The resulting MSER of the combined regions is achieved by summing up the bright-on-dark region with the complement of the dark-on-bright region. The result of combining the two regions of Fig. 4(a) as the input image is shown in Fig. 4(d).

The a-MSER algorithm to detect ROI is given below.

Algorithm 1: ROI detection based on a-MSER algorithm

```
Input: Scene Image $I_s$
Output: Detected ROI $I_{roi}$
for each $I_s$ do
    Convert the image to grayscale $I_g$ = rgb2gray($I_s$)
    Compute "maximally stable" regions for the image $I_g$ that have a lower variation than the regions one level above or below where variation is given by Eq. (1)
    Derive the list of pixels belonging to that region with the region seeds and image
    Extract MSER for both bright-on-dark $I_{bod}$ and dark-on-bright $I_{dob}$ regions by choosing the delta value based on intensity profile
    Take complement of the dark-on-bright region and combine it with bright-on-dark region resulting in the required ROI $I_{roi} = I_{bod}$ +imcomplement($I_{dob}$)
end
```

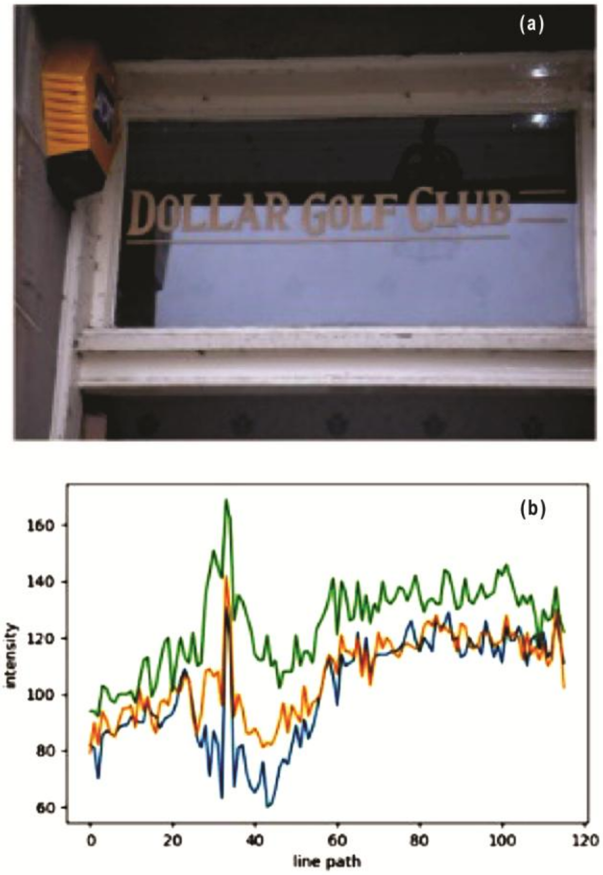

Fig. 2 — (a) Input Image, (b) Intensity Profile

The comparison of our a-MSER method on other existing MSER methods is shown in Fig. 5 and we found that all the text characters could be detected well with our method as clearly illustrated in Fig. 5(c). The result of Chen *et al.*[31] is Fig. 5(a) and that of Li *et al.*[40] is Fig. 5(b). To prove the importance of a-MSER in our proposed method, an input image was

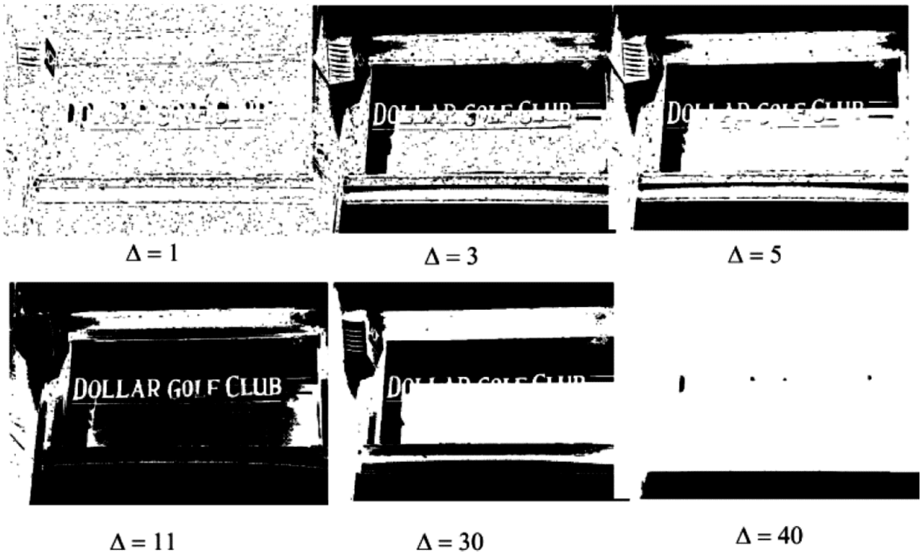

Fig. 3 — Effect of Δ

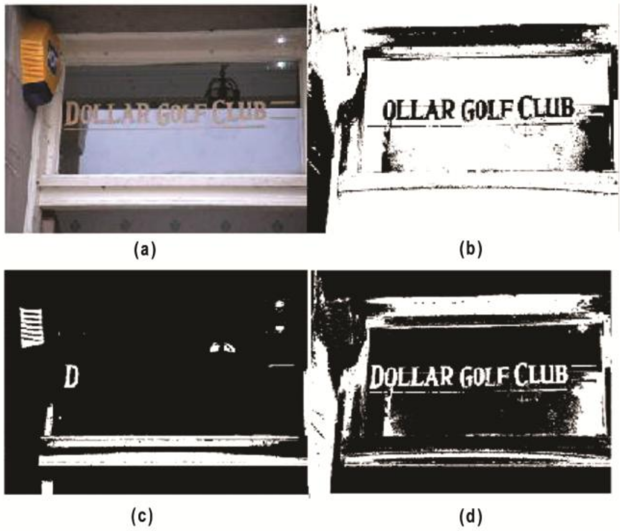

Fig. 4 — (a) Input Image, Detected MSER Regions (b) Dark-on-bright regions, (c) Bright-on-dark regions, (d)Combined MSER regions

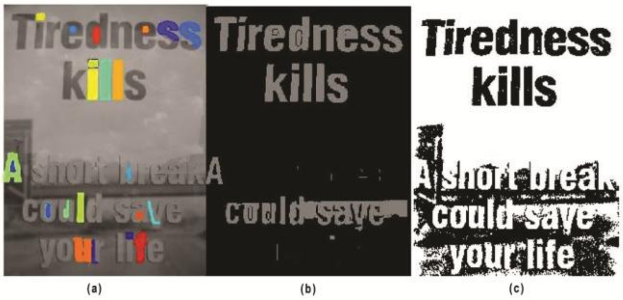

Fig. 5 — Detected MSER Regions (a) result of Chen *et al.*[31], (b) result of Li *et al.*[40], (c) result of a-MSER

directly given to the STEMYN model skipping a-MSER step. This has a significant impact on text detection as shown in Fig. 6. The text detection result

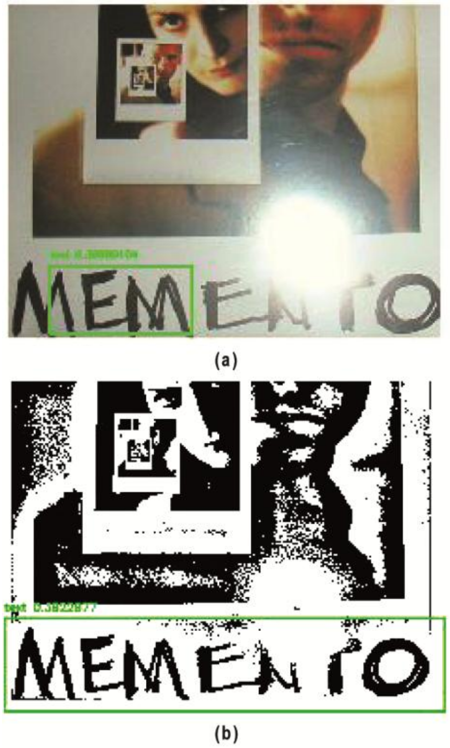

Fig. 6 — Text Detection Result (a) Without a-MSER (b) With a-MSER

without a-MSER is shown in Fig. 6(a) and with a-MSER is shown in Fig. 6(b). When a-MSER is present, the difference between text and background regions is clearly demarcated which aids in better text detection results.

STEMYN Model

The proposed model STEMYN uses modified YOLOv2[4] (an improved version of YOLO[6]), an object detection system targeted for real-time processing. It is a full-fledged CNN capable of doing classification, localization and detection at a single shot. The model YOLOv2 is trained and tuned to find the location of the text characters in a natural scene image by not only classifying the image (e.g., a binary classification problem: whether there is text present in an image or not) but also locating a bounding box around the text, if present. Detection goes a level ahead by focusing to locate multiple instances of text occurrences, by marking their locations. For real-time processing, SSD[3] gave strong competition to YOLO demonstrating a higher accuracy. Further, YOLO has relatively low recall to region proposal-based approaches and also it makes more localization errors. These errors are fixed by YOLOv2 with the focus in accuracy improvement and speedy detection. There are more recent versions of YOLO namely YOLOv3[(5)] and YOLOv4[(2)]. When compared to YOLOv2, they are slower due to very deep networks which cannot be trained using CPU alone. Even, VGG16 is slower than YOLOv2 as it uses only 8.52 billion operations for a forward pass whereas the former requires 30.69 billion floating-point operations.[4] Many of the text detection frameworks use Visual Geometry Group (VGG) models as the base feature extractor.[41] Though it is powerful and accurate, it is highly complex. On the other hand, YOLO framework is a customized version of GoogleNet architecture.

Many research works on text detection that are based on SSD and R-CNN are prevalent. This is a different approach based on modified YOLOv2 which has already proved its mark in the field of object detection. Definitely, YOLOv2 is a cynosure in the field of scene text detection.

Architecture

The STEMYN model has 22 convolution layers as shown in Fig. 7 with 6 max pooling layers compared to the original YOLOv2 model wherein there are 22 convolution layers with 5 max pooling layers. This model is different from the original YOLOv2 in two aspects.

(i) 1 × 1 convolution layer is added and the image size is enhanced from 13 × 13 to 26 × 26. For small objects, YOLOv2 has a poor detection rate. What happens here is after running through different convolution and max pooling layers, the image becomes very small which in-turn reflects in the inability to extract features that aid in better detection accuracy. So, the enhancement in this network comes after the 13th convolution layer. The output of this layer 26 × 26 is used as input to a separate path of 5 convolution layers with 1 max pooling layer. We have already seen that stacking more layers in a neural network could lead to better results from GoogleNet, ResNet and VGGNet. ResNet at the same time proves that accuracy cannot be improved if we exceedingly deepen the network. Based on this, a 1 × 1 convolution layer is added to the network as shown in the yellow-coloured box in Fig. 7. This added convolution layer is equal to a non-linear transformation which can also improve the articulateness of the network. Through these enhancements, 15th–19th layers which were all 13 × 13 images size in the original network has now been changed to 26 × 26. This allows the network to learn more effective features with considerable size images.

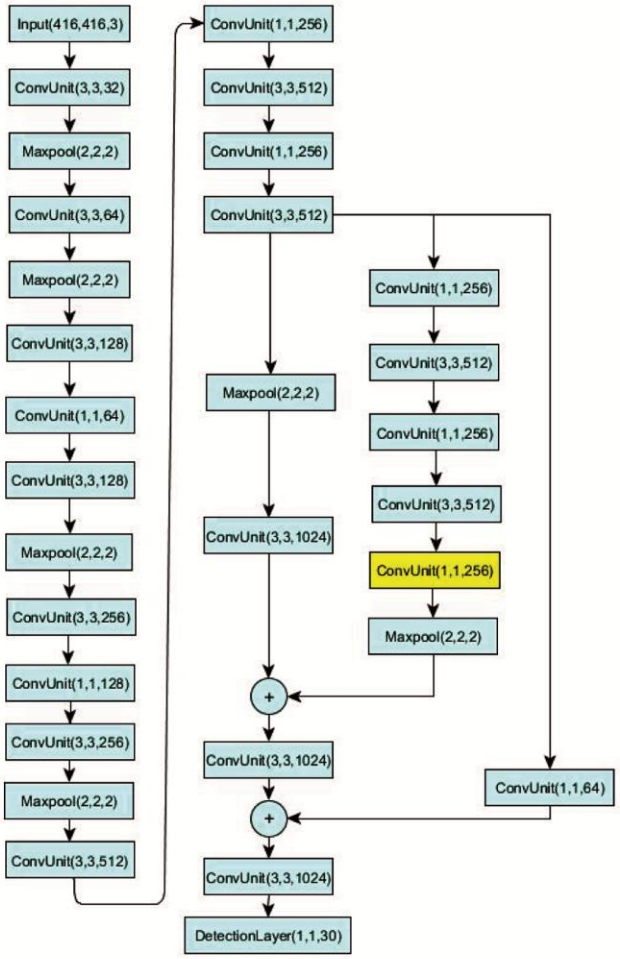

Fig. 7 — STEMYN Architecture

(ii) The repeated convolution layer in steep layers is removed. Basically designed as an object detection model YOLOv2 detect the different sets of classes such as people, cars, houses, aeroplanes, etc. The original YOLOv2 has got three continuous repeated 3×3×1024 convolution layers in steep layers. Generally, this is for dealing with multiple classes with vast differences. But in our case of detection, we are dealing with only one class namely text. Here repeated convolution layers may not improve performance rather make the model more complex. Therefore, we removed one $3 \times 3 \times 1024$ convolution layer from steep layers as shown in Fig. 7. The other existing aspects of the YOLOv2 network remain intact.

The original YOLOv2 predicts detection on a 13×13 feature map given an input of dimension 416×416 that may not be enough to extract smaller objects. To overcome this, a pass through layer is added that brings features from an earlier layer at 26×26 resolution as shown in Fig. 7. The pass-through layer integrates the higher resolution features with the lower resolution features by piling adjoining features into various channels. 1×1 convolutions are used to shrink the feature representations between 3×3 convolutions. It predicts detection on a 13×13 feature map which is the size of the grid as well. Each grid predicts 5 bounding boxes with each bounding box represented by 6 elements (x, y, width, height, class and confidence score) hence giving output tensor of $13 \times 3 \times 30$. This is accomplished by the final detection layer as shown in Fig. 7.

Training

The training process of the proposed text detection system is shown in Fig. 8. The network is trained using ADAM[42] optimizer. The training happens for 25 epochs with a starting learning rate of 10^{-4} with a batch size of 8 for ICDAR 2013, MSRA TD500 and ICDAR 2015 datasets. ICDAR 2013 dataset contains 229 training images and 233 testing images, ICDAR 2015 has 1000 and 500, MSRA TD500 has 300 and 200 respectively. Multistage training as shown in Fig. 8 is employed to train our model exclusively with these three small datasets employing CPU solely for cost efficiency. At every stage of training, one-fourth of the training images from each of the datasets are taken as input for training. The model weights file created from the first group of training images is given as pre-trained weights for the next stage of training. This way it proceeds for all the groups and results are better at each stage output as shown in Fig. 8. This shows that even with a small dataset we can achieve good results. The clustering method namely k-means is used to determine bounding box anchors with value k set to 5. The anchors generated for the ICDAR 2013 dataset are: (0.45, 0.34), (0.70, 1.06), (1.30, 0.51), (2.22, 1.48) and (4.61, 3.47). The

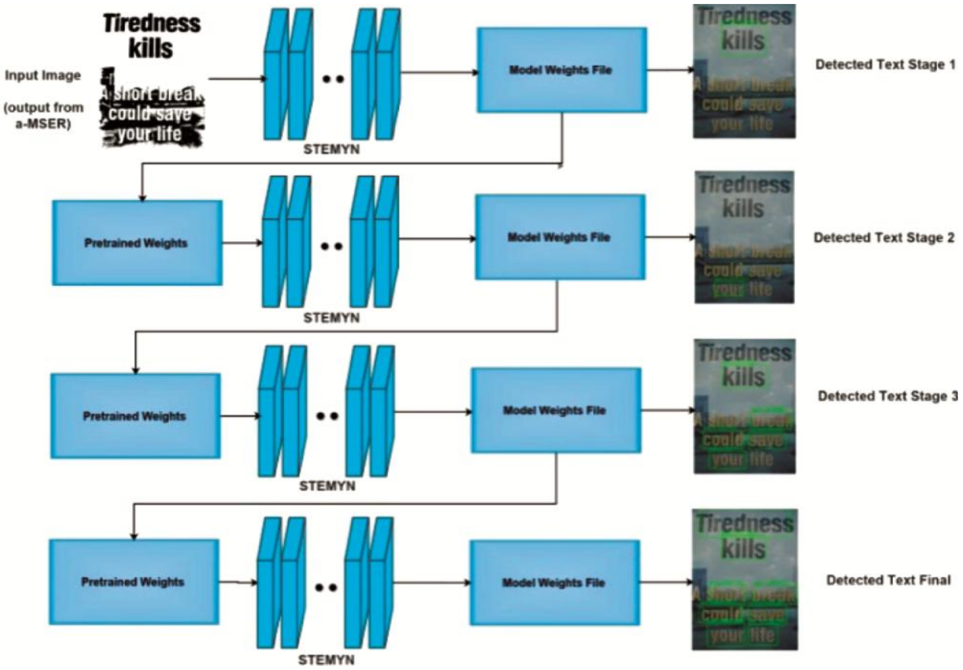

Fig. 8 — Training Process of Proposed Text Detection System

training set is automatically divided into a training set and validation set with 80:20 split. The training process stops when the loss on the validation set has not improved in ten consecutive epochs.

Loss Function

The model's loss function is same as YOLOv2's loss function as shown in Eq. (2) but for the modification in the classification loss (instead of cross entropy (CE), focal loss (FL)[43] is used).

$$\lambda_{\text{coord}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{txt}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]+$$
$$\lambda_{\text{coord}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{txt}\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right]+$$
$$i=0S^2{}_{j=0B}\mathbb{1}ijtxtC_i-C_i2+\lambda_{\text{notxt}}i=0S^2{}_{j=0B}\mathbb{1}ijnotxtC_i-C_i2+$$
$$i=0S^2\mathbb{1}itxt_{c\in classes}p_i(c)-p_i(c)2 \quad \ldots (2)$$

where, x_i, y_i, is the location of the centroid of the anchor box, w_i, h_i, is the width and height of the anchor box, C_i is the *Objectness*, i.e. confidence score of whether there is any text or not, and $p_i(c)$ is the classification loss. We see that almost all losses are mean squared errors, except classification loss for which FL[43] is used. $\sum^{B}{}_{j=0}$ in Eq. (2) denotes loss computation for each anchor box (5 in total), $\sum^{S^2}{}_{j=0}$ in Eq. (2) denoted loss computation for each of the 13×13 cells where S=13. When there is text present in the cell i, $\mathbb{1}_{ij}^{txt}$ is 1, else 0. When there is no text present in the cell i, $\mathbb{1}_{ij}^{notxt}$ is 1, else 0. When a particular class i.e., text is predicted, $\mathbb{1}_{i}^{txt}$ is 1, else 0. λs are constants. They are used to independently weigh parts of the loss functions to increase model stability. To focus more on detection, λ_{coord} is kept highest for coordinates with value 5 and λ_{notxt} is lowest for confidence predictions with value 0.5 when there is no text.

FL is an extended version of CE loss. FL is given in Eq. (3)

$$FL = -\alpha(1-p)\gamma log\gamma \quad \ldots (3)$$

where, p[0;1] is the model's estimated probability, α is offset class imbalance of number of examples and γ focuses more on hard examples. When tried for various values of γ (α being scalar factor for this criterion generated through torch.rand function), FL proved to be less while compared with CE which is tabulated in Table 2.

Non-Maximal Suppression (NMS)

The grid design in the YOLOv2 network imparts spatial diversity in bounding box predictions. Most of the time there is clarity on which grid cell a particular word falls into and hence the network correctly predicts one box for each word. But sometimes words in big font or those near the border of the multiple cells can be well localized by multiple cells. In these scenarios, NMS helps in removing multiple detections with nms threshold as 0.3.

Table 2 — Comparison between losses - FL and CE

Gamma	CE	FL
0	1.6223	1.6223
1	1.6544	1.3424
2	1.4402	0.8469
3	1.4933	0.7054
4	1.5807	0.6561

Prediction

The successful results of the text detection model (a-MSER + STEMYN) on ICDAR 2013 dataset are shown in Fig. 9(a). The first image in Fig. 9(a) has not been detected so far by any of the detection networks as it has very little difference between the foreground and background. But this model could detect that with the a-MSER output of that image taken as input for prediction. In a similar manner, it could detect text from images containing letters with unusual font size and very small fonts as shown in second and third images respectively and non-uniform illumination as shown in fourth image which were left undetected by many of the networks especially Tang & Wu[44] which uses three separate CNN for detection, segmentation and classification. Similarly, the successful results of this detection model on Incidental Scene Text ICDAR 2015 and MSRA TD500 are shown in Figs. 9(b) and 9(c) respectively. We could see that this model could detect text from images with varied font sizes and present in multilingual environment.

Results and Discussion

Datasets

The evaluation of the proposed system was performed on benchmark datasets: ICDAR 2013,2015 and MSRA TD500, i.e., the words are from ICDAR Robust Reading Competitions focused scene text dataset[10], incidental scene text dataset[11] and MSRA dataset.[12]

Metrics

The proposed system is assessed based on criterion such as Precision, Recall and F-measure. The PASCAL VOC style intersection-over-union (IoU) overlap method is used for finding out the

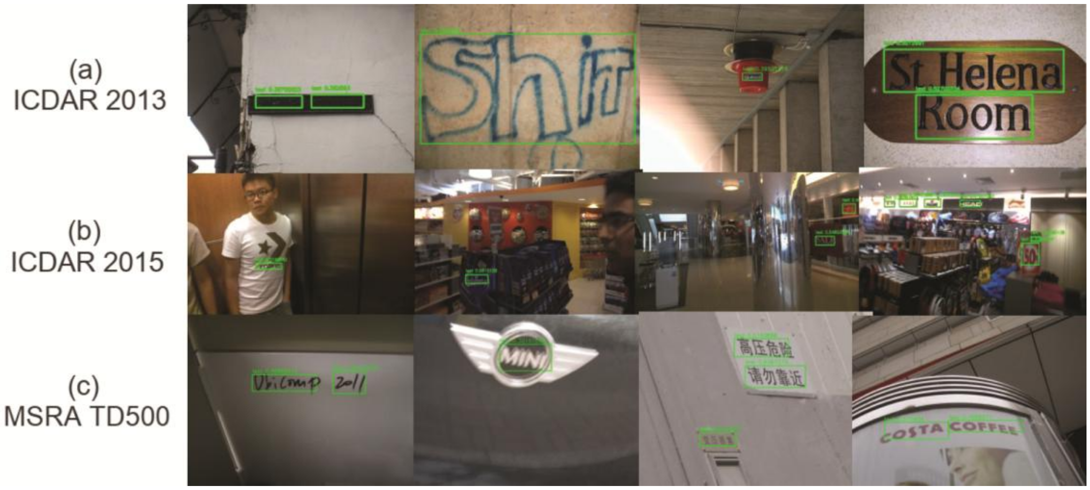

Fig. 9 — Some successful results on three benchmark quadrilateral-type datasets

performance of text detection. The predicted and ground truth bounding boxes are compared once the final predictions are determined after NMS. The IoU between two boxes is calculated as the ratio of area of overlap and area of union. A detected box is considered as a hit box if the IoU between detected and ground truth box is higher than the given IoU threshold of 0.3. TP, FP, and FN are the number of hit boxes, incorrectly identified boxes, and missed boxes, respectively. These are used in Eqs (4), (5) and (6) to calculate Precision, Recall and F-measure.

$$Precision = \frac{TP}{TP+FP} \quad \ldots (4)$$

$$Recall = \frac{TP}{TP+FN} \quad \ldots (5)$$

$$F - Measure = \frac{2\times Precision\times Recall}{Precision+Recall} \quad \ldots (6)$$

Ablation Study

A series of experiments and ablation studies are administered to display the efficacy of this system, a-MSER with STEMYN than the original YOLOv2. The dataset ICDAR 2013 is used for evaluation of the models on the basis of IoU, precision and recall which is given in Table 3. The presence of a-MSER is significant in the text detection process which is already shown in Fig. 6 wherein the text regions present in a non-uniformly illuminated image could be detected precisely. The experiments are carried out to show the importance of a-MSER with the help of ICDAR 2013 dataset. The anchors generated for the modified YOLOv2 model using the k-means algorithm with input images as output from a-MSER has greater IoU than those without a-MSER. Also, the distance between clusters converged to a minimum value in a lesser number of iterations. The STEMYN model is evaluated with and without a-MSER based on evaluation parameters IoU, Precision, Recall and results are tabulated in Table 4 indicating STEMYN with a-MSER is better than STEMYN without a-MSER.

Table 3 — Comparison with the original YOLOv2

Model	IoU	Recall	F-Measure
YOLOv2	0.85	0.89	0.81
STEMYN with a-MSER	0.89	0.93	0.88

Table 4 — Modified YOLOv2 with and without a-MSER

Model	IoU	Recall	F-Measure
STEMYN without a-MSER	0.86	0.90	0.82
STEMYN with a-MSER	0.89	0.93	0.88

Experimental Evaluation

The evaluation of a-MSER with STEMYN was carried out on standard benchmark datasets: ICDAR 2013, 2015 and MSRA TD500. The words are taken from ICDAR robust reading competition's focused scene text localization dataset[10], incidental scene text dataset[11], and multi-orientation text dataset.[12] The comparison of our model with other text detection methods are given in Tables 5, 6 and 7 for the datasets ICDAR 2013, 2015 and MSRA-TD500 respectively. From all these tables, the results are better when input training images are split and trained than when input images are given as a whole. The F-measure value acquired under the model STEMYN with a-MSER_splitds is on par with other text detection methods such as Pixellink[23] and He *et al.*[19] as shown in Table 5. Methods of Xie *et al.*[26], Lyu *et al.*[24], Ma *et al.*[37] and Baek *et al.*[45] produce better F-measure value.

The case is analogous with the detection of ICDAR 2015 dataset also and Table 6 shows the

Table 5 — Comparison with other text detection methods for ICDAR 2013 Focused Scene Text Dataset

Methods	Precision	Recall	F-Measure
Jaderberg *et al.*[46]	0.86	0.68	0.76
TextFlow[41]	0.85	0.76	0.80
SSD[3]	0.80	0.60	0.68
TextBoxes[34]	0.88	0.83	0.85
Gupta *et al.*[33]	0.93	0.76	0.84
Pixellink[23]	0.88	0.87	0.88
He *et al.*[19]	0.88	0.87	0.88
STEMYN with a-MSER_wholeds	0.92	0.79	0.85
STEMYN with a-MSER_splitds	0.93	0.83	0.88

Table 6 — Comparison with other text detection methods for ICDAR 2015 Incidental Scene Text Dataset

Methods	Precision	Recall	F-Measur
Shi *et al.*[38]	0.73	0.77	0.75
EAST[47]	0.83	0.78	0.81
He *et al.*[48]	0.82	0.80	0.81
Ma *et al.*[37]	0.84	0.77	0.80
Pixellink[23]	0.85	0.82	0.83
He *et al.*[19]	0.84	0.83	0.83
Lyu *et al.*[21]	0.85	0.81	0.83
TextBoxes++[35]	0.85	0.81	0.83
STEMYN with a-MSER_wholeds	0.85	0.76	0.80
STEMYN with a-MSER_splitds	0.87	0.78	0.83

Table 7 — Comparison with other text detection methods for MSRA-TD500 Text Dataset

Methods	Precision	Recall	F-Measur
He *et al.*[48]	0.770	0.700	0.740
Ma *et al.*[37]	0.820	0.690	0.750
Shi *et al.*[38]	0.860	0.700	0.772
EAST[47]	0.873	0.674	0.761
Pixellink[23]	0.830	0.732	0.778
Liao *et al.*[36]	0.870	0.730	0.790
STEMYN with a-MSER_wholeds	0.870	0.730	0.793
STEMYN with a-MSER_splitds	0.870	0.730	0.798

corresponding comparison. The F-measure value obtained under the model STEMYN with a-MSER_splitds is on par with many other text localization methods namely Pixellink[23], He *et al.*[19], Lyu *et al.*[24] and TextBoxes++.[35] Methods of Xie *et al.*[26], FOTS[20] and Baek *et al.*[45] have better F-measure value.

The comparison for MSRA TD 500 dataset is displayed in Table 7. The F-measure value received under the model STEMYN with a-MSER_splitds is better than other text localization methods namely Pixellink[23] and Liao *et al.*[36]. Methods of Lyu *et al.*[24] and Baek *et al.*[45] obtain better F-measure value.

Another significant point to be highlighted here regarding results on ICDAR 2013, 2015 and MSRA TD500 datasets is, though the results are not the best state-of-the-arts, this has been achieved by effectively training the model with the help of small datasets only with the proposed method in a cost-efficient manner employing CPU alone.

Conclusions

We have developed a method for text detection based on Maximally Stable Extremal Regions together with Convolutional Neural Network. The proposed method namely a-MSER with STEMYN (CNN architecture based on modified YOLOv2) could surpass the existing state-of-the-art methods for detecting text regions in natural scene images. The method a-MSER (amended MSER) is arrived at taking into account the intensity variations between text and background regions effectively. The output images from a-MSER are used as input for STEMYN model which is designed to overcome the limitations of original YOLOv2 object detection framework. To detect texts with smaller fonts better, we have introduced 1 × 1 layer with image size enhanced from 13 × 13 to 26 × 26. With the reconstructed classification loss using FL instead of CE, the performance of the text detection model is better. The repeated convolution layer in the deep layers is removed making the model less complex as it does not aid in the performance of the system. The entire text detection process employs small datasets only, but a-MSER together with STEMYN and multistage training has fetched these results in a cost-efficient manner using only CPU. Some of the slanting text regions could not be located precisely due to rectangular bounding boxes and in turn resulted in additional spaces. Hence, our future work is focused on detecting curved text regions with the help of polygon as bounding boxes so that minimal regions are encompassed.

Acknowledgement

This paper was funded from DST/INSPIRE Fellowship Program.

References

1 Matas J, Chum O, Urban M, Pajdla T, Robust wide-baseline stereo from maximally stable extremal regions, *Image Vis Comput*, **22(10)** (2004) 761–767.

2 Bochkovskiy A, Wang C Y & Liao H Y M, Yolov4: Optimal speed and accuracy of object detection, *arXiv:200410934*, (2020).

3 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y & Berg A C, SSD: Single shot multibox detector, *ECCV*, (2016) 21–37.

4 Redmon J & Farhadi A, Yolo9000: Better, faster, stronger, *CVPR*, (2017) 7263–7271.

5 Redmon J & Farhadi A, Yolov3: An incremental improvement, *arXiv:180402767*, (2018).

6 Redmon J, Divvala S, Girshick R & Farhadi A, You only look once: Unified, realtime object detection, *CVPR*, (2016) 779–788.

7 Girshick R, Donahue J, Darrell T & Malik J, Rich feature hierarchies for accurate object detection and semantic segmentation, *CVPR*, (2014) 580–587.

8 Girshick R, Fast r-cnn, *ICCV*, (2015) 1440–1448.

9 He K, Gkioxari G, Dollar P, Girshick R, Mask r-cnn, *ICCV*, (2017) 2961–2969.

10 Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda L G, Mestre S R, Mas J, Mota D F, Almazan J A, De Las Heras L P, Icdar 2013 robust reading competition, *Proc Int Conf Doc Anal Recog*, (2013) 1484–1493.

11 Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V R, Lu S, Shafait F, Uchida S & Valveny E, Icdar 2015 competition on robust reading, *Proc Int Conf Doc Anal Recog*, (2015) 1156–1160.

12 Yao C, Bai X, Liu W, Ma Y & Tu Z, Detecting texts of arbitrary orientations in natural images, *CVPR*, (2012) 1083–1090.

13 Chen X, Yuille A L, Zhu Y, Luo C & Wang T, Text recognition in the wild: A survey, *arXiv preprint arXiv:2005.03492*, (2020).

14 Ye Q & Doermann D, Text detection and recognition in imagery: A survey, *IEEE Trans Pattern Anal Mach Intell*, **37(7)** (2015) 1480–1500.

15 Jaderberg M, Vedaldi A & Zisserman A, Deep features for text spotting, *ECCV*, (2014) 512–528.

16 Wang T, Wu D J, Coates A & Ng A Y, End-to-end text recognition with convolutional neural networks, *Proc Int Conf Pattern Recognit*, (2012) 3304–3308.

17 Huang W, Lin Z, Yang J & Wang J, Text localization in natural images using stroke feature transform and text covariance descriptors, *ICCV*, (2013) 1241–1248.

18 Neumann L & Matas J, Real-time lexicon-free scene text localization and recognition, *IEEE Trans Pattern Anal Mach Intell*, **38(9)** (2015) 1872–1885.

19 He T, Tian Z, Huang W, Shen C, Qiao Y & Sun C, An end-to-end textspotter with explicit alignment and attention, *CVPR*, (2018) 5020–5029.

20 Liu X, Liang D, Yan S, Chen D, Qiao Y & Yan J, Fots: Fast oriented text spotting with a unified network, *CVPR*, (2018) 5676–5685.

21 Lyu P, Liao M, Yao C, Wu W & Bai X, Mask text spotter: An end-to-end trainable neural network for spotting text with arbitrary shapes, *ECCV*, (2018) 67–83.

22 Wei G, Rong W, Liang Y, Xiao X & Liu X, Toward arbitrary-shaped text spotting based on end-to-end, *IEEE Access*, **8**(2020) 159906–159914.

23 Deng D, Liu H, Li X & Cai D, Pixellink: Detecting scene text via instance segmentation, *Proc Conf AAAI Artif Intell*, (2018).

24 Lyu P, Yao C, Wu W, Yan S & Bai X, Multi-oriented scene text detection via corner localization and region segmentation, *CVPR*, (2018) 7553–7563.

25 Wang W, Xie E, Li X, Hou W, Lu T, Yu G, Shao S, Shape robust text detection with progressive scale expansion network, *CVPR*, (2019) 9336–9345.

26 Xie E, Zang Y, Shao S, Yu G, Yao C & Li G, Scene text detection with supervised pyramid context network, *Proc Conf AAAI ArtifIntell*, **33** (2019) 9038–9045.

27 Dai P, Zhang H & Cao X, Deep multi-scale context aware feature aggregation for curved scene text detection, *IEEE Trans Multimed*, **22 (8)** (2020) 1969–1984.

28 Chen H, Tsai S S, Schroth G, Chen D M, Grzeszczuk R & Girod B, Robust text detection in natural images with edge-enhanced maximally stable extremal regions, *Proc Int Conf Image Process*, (2011) 2609–2612.

29 Neumann L, Matas J, Real-time scene text localization and recognition, *CVPR*, (2012) 3538–3545.

30 Yin X C, Yin X, Huang K & Hao H W, Robust text detection in natural scene images, *IEEE Trans Pattern Anal Mach Intell*, **36(5)** (2014) 970–983.

31 Huang W, Qiao Y & Tang X, Robust scene text detection with convolution neural network induced MSER trees, *ECCV*, (2014) 497–511.

32 He T, Huang W, Qiao Y & Yao J, Text-attentional convolutional neural network for scene text detection, *IEEE Trans Image Process*, **25(6)** (2016) 2529–2541.

33 Gupta A, Vedaldi A & Zisserman A, Synthetic data for text localisation in natural images, *CVPR*, (2016) 2315–2324.

34 Liao M, Shi B, Bai X, Wang X & Liu W, Textboxes: A fast text detector with a single deep neural network, *Proc Conf AAAI ArtifIntell*, (2017).

35 Liao M, Shi B & Bai X, Textboxes++: A single-shot oriented scene text detector, *IEEE Trans Image Process*, **27(8)** (2018) 3676–3690.

36 Liao M, Zhu Z, Shi B, Xia Gs & Bai X, Rotation-sensitive regression for oriented scene text detection, *CVPR*, (2018) 5909–5918.

37 Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y & Xue X, Arbitrary-oriented scene text detection via rotation proposals, *IEEE Trans Multimed*, **20(11)** (2018) 3111–3122.

38 Shi B, Bai X & Belongie S, Detecting oriented text in natural images by linking segments, *CVPR*, (2017) 2550–2558.

39 VLFeat, Maximally stable extremal regions (MSER) feature detector, *http://www.vlfeat.org/overview/mser.html*, (2015).

40 Li Y, Jia W, Shen C & van den Hengel A, Characterness: An indicator of text in the wild, *IEEE Trans Image Process*, **23(4)** (2014) 1666–1677.

41 Tian S, Pan Y, Huang C, Lu S, Yu K & Lim Tan C, Text flow: A unified text detection system in natural scene images, *ICCV*, (2015) 4651–4659.

42 Kingma D P & Ba J, Adam: A method for stochastic optimization, *arXiv preprint arXiv:14126980*, (2014).

43 Lin T Y, Goyal P, Girshick R, He K & Dollar P, Focal loss for dense object detection, *ICCV*, (2017)2980–2988.

44 Tang Y & Wu X, Scene text detection and segmentation based on cascaded convolution neural networks, *IEEE Trans Image Process*, **26(3)** (2017) 1509–1520.

45 Baek Y, Lee B, Han D, Yun S & Lee H, Character region awareness for text detection, *CVPR*, (2019) 9365–9374.

46 Jaderberg M, Simonyan K, Vedaldi A & Zisserman A, Reading text in the wild with convolutional neural networks, *Int J Comput Vis*, **116(1)** (2016) 1–20.

47 Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W & Liang J, East: an efficient and accurate scene text detector, *CVPR*, (2017) 5551–5560.

48 He W, Zhang X Y, Yin F & Liu CL, Deep direct regression for multi-oriented scene text detection, *ICCV*, (2017) 745–753.