A deep learning approach for robust, multi-oriented and curved text detection

Ramin Ranjbarzadeh[1], Saeid Jafarzadeh Ghoushchi[2], Shokofeh Anari[3], Sadaf Safavi[4], Nazanin Tataei Sarshar[5], Erfan Babaee Tirkolaee[6,*], Malika Bendechache[7]

[1] School of Computing, Faculty of Engineering and Computing, Dublin City University, Ireland.
ramin.ranjbarzadehkondrood2@mail.dcu.ie

[2] Faculty of Industrial Engineering, Urmia University of Technology, Urmia, Iran.
s.jafarzadeh@uut.ac.ir

[3] Department of Accounting, Economic and Financial Sciences, Islamic Azad University, South Tehran Branch, Tehran, Iran.
shokofehanarii@gmail.com

[4] Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran.
sf.safavi@gmail.com

[5] Department of engineering, Islamic Azad university, Tehran north Branch, Tehran, Iran.
ab.tataee@gmail.com

[6] Department of Industrial Engineering, Istinye University, Istanbul, Turkey.
erfan.babaee@istinye.edu.tr

[7] Lero & ADAPT Research Centres, School of Computer Science, University of Galway, Ireland.
malika.bendechache@universityofgalway.ie

Abstract

Background Automatic text localization and segmentation in a normal environment with vertical or curved texts are core elements of numerous tasks comprising the identification of vehicles and self-driving cars, and preparing significant information from real scenes to visually impaired people. Nevertheless, texts in the real environment can be discovered with a high level of angles, profiles, dimensions, and colors which is an arduous process to detect.

Methods In this paper, a new framework based on a convolutional neural network (CNN) is introduced to obtain high efficiency in detecting text even in the presence of a complex background. Due to using a new inception layer and an improved ReLU layer, an excellent result is gained to detect text even in the presence of complex backgrounds. At first, four new m.ReLU layers are employed to explore low-level visual features. The new m.ReLU building block and Inception layer are optimized to detect vital information maximally.

Results The effect of stacking up inception layers (kernels with the dimension of 3×3 or bigger) is explored and it is demonstrated that this strategy is capable of obtaining mostly varying-sized texts further successfully than a linear chain of Convolution Layers (Conv layers). The suggested text detection algorithm is conducted in four well-known databases, namely ICDAR 2013, ICDAR

2015, ICDAR 2017, and ICDAR 2019.
Conclusions Text detection results on all mentioned databases with the highest Recall of 94.2%, Precision of 95.6%, and F-score of 94.8% illustrate that the developed strategy outperforms the state-of-the-art frameworks.
Keywords: Deep learning, Text detection, Curved texts, Convolutional neural networks, Text segmentation.

1. Introduction

Automatically understanding of scene text with edges and corner-point information in a real environment represent a significant influence in a diverse set of intelligent system applications in visual assistance, intelligent traffic systems, automatic driving car, and so on [1]. In contrast to algorithms based on the character or text applied to document images, which is sufficiently well addressed by Optical Character Recognition (OCR) system, text classification and localization in natural images are still an open complex problem [2]. Since text obtained from natural images typically include a wide variety of useful text contents surrounded by objects in comparison to graphics text, detecting target text in the real scene is a challenging task [3]. This is due to the fact that the system needs to reject irrelevant objects and finds the location of the texts.

There are a number of weaknesses in a text localization system that work in only one direction (horizontal), as a sizeable part of the texts achieved from natural images in the real world has a wide range of orientations, sizes, and fonts [4]. Such restriction would make it fail to extract the useful features contained in non-horizontal texts and thus seriously limits the efficiency and the scalability of these strategies [5].

Usually, text detection strategies are based on two major algorithms: (i) approaches based on the texture, and (ii) approaches based on the region [6]. The texture-based methods are based on exploring significant features in the whole image, while the region-based techniques only work on a part of the image to ease the problem of execution time [7]. The features in the region-based techniques are permanently distinctive in real scene text regions [8]. The two main strategies for doing this, are techniques based on the connected components and strategies based on the sliding windows [5]. The CC strategies mostly emphasize significant information such as edges that can be detected using an edge extracting algorithm or color-thresholding techniques and then combining the sub-Maximally Stable Extremal Regions (MSER) parts into a text-line or word area [9]. The mentioned strategies are capable to work in some hard detect scenarios including varying brightness or contrast, light flickering, recognized stroke characters, display reflections, and several joined characters. In the texture-based methods, by exploring the spreading of the textural features in a local or global area, a surrounding window related to the text area can be easily chosen and attached to text lines [10]. However, the significant disadvantages of strategies based on only textures can be described by a simple feature extraction method. Techniques employing sliding windows for feature extraction examine the image contents and try to extract numerous image rectangles. Nevertheless, these methods lead to an increase in complexity and computational cost [11].

The purpose of recognition of a wide range of texts is to identify and describe a sequence of characters and content details from a selected region inside the text images for recognizing signboards, license plates, and so on [12]. For recognizing the words, there is a need to detect them first [1]. Due to the wide disparity in languages used in different areas and in dissimilar language texts, significant parts of present scene text recognition approaches emphasize merely analyzing the obtained image from the most applicable language texts (limited characters) [6]. To this end, most of the text analyzing frameworks are widely investigated based on the English text and are assorted into two key classes: the word-based and the character-based approaches [13]. Directly, the word-based approach recognizes a similar pattern of the potential word inside the obtained image from the real scene [14]. As countless English words are presented in the obtained real scene images, common strategies cannot be able to determine a word or sentence directly without needing to consider any extra information [15]. Basically, the character-based approach determines all predefined characters inside a Region of Interest (ROI) by a character classifier. All extracted characters make one or more words that can be recognized by a combination of individual outcomes [16].

Recently, many Machine Learning (ML) methods are applied to various fields including social sciences [17], optimization [18], regulatory systems [19], data augmentation [20], stochastic systems [21], Internet of Medical Things [22], Internet of Things [23], Time series [24], medical data analysis [25], degenerative disorder [26], and recommendation systems [27].

In recent years, to solve the problem and difficulty of text localization and detection in a real scene due to the fuzzy boundary between text components and background, irregular shape, low contrast, and low intensity numerous ML-based complex strategies have been implemented [28]. All segmentation and recognition algorithms are classified into two main groups based on their characteristics, including semi-automatic techniques (interactive approaches) and automatic frameworks [29]. The interactive or semi-automatic frameworks normally can be employed by various Human-Machine Interactions (HMI) or user directions [30]. This kind of text detection is somehow impossible to use in an environment that needs a real-time response [31]. So, automatic frameworks have been employed in a number of applications to diminish the costs and time of analyzing and steadily develop accuracy [32].

Current automatic models mainly can be explored inside the two wide-ranging classes, including anti-learning and learning techniques [33]. The anti-learning frameworks regularly comprise the active contour, clustering, region-growing, graph cut, and level set methods [34]. Region-growing approaches are pixel-based image segmentation strategies that select the touching pixels iteratively with many similarities (homogeneities) in intensity, direction, color, or variance (adding the neighboring pixels) [35]. The efficiency of region-growing algorithms can be influenced by selecting the seed points, and they benefit from small calculation complexity and high speed [36]. Graph cut methods are powerful energy minimization (optimization) strategies that characterize the image to an undirected weighted graph. It means each input image can be represented as a graph of nodes. Due to the use of both boundaries and regional information, it has obtained a lot of attention [37]. In these approaches, there is a need to have prior information about the shape and size of the target object, and every location (pixel) $p \in I$ inside the image is implied as a node in the graph. Furthermore, every edge connects two adjacent nodes, therefore the weight of each edge defines the rate of the similarities among each pair [38].

In recent years, employing a neuron-based model as an automatic learning approach such as the Convolutional Neural Network (CNN/ConvNet) has been a surge of interest in text detection in the real scene [39]. There are different kinds of neural networks (NNs) in deep learning, such as artificial neural networks (ANN) [40], radial basis function (RBF) [41], convolutional neural networks (CNN) [42], recurrent neural networks (RNN) [43], etc. Unlike hand-crafted feature extraction models [44], these deep learning-based models are able to explore more informative information and hidden pattern inside the input data automatically [34].

To overcome the problem of text instances with arbitrary shapes, Liu et al [16] proposed a novel BezierAlign layer. The Bezier curve detection layer was employed to adaptively fit the oriented or curved text. Ma et al. [45] proposed a combination of the Rotated Bounding Box Representation method and Rotation Anchors technique to overcome the issues of text angle information. They used the convolutional layers of VGG-16 as sharable layers for extracting the low-level features, and the last convolutional layer is responsible for proposing the horizontal region. Moreover, a multi-modal algorithm has been proposed by [46] for Bib text/number recognition that is printed on cardboard tags or papers in Marathon natural images. This strategy combines text detection and torso detection to obtain an acceptable result. As torso detection focus on detecting the body parts such as the backside, stomach, and chest, there is no need to extract features related to the face. By integrating the binarization process at the post-processing step for segmenting texts, a Differentiable Binarization (DB) module is introduced in Liao et al. [47]. Moreover, they employed an efficient Adaptive Scale Fusion (ASF) module for improving the robustness of scale variation by fusing features of diverse scales adaptively.

To address the issue of the complex background, a Scale-based Region Proposal Network has been proposed by [48]. They investigated a two-stage pipeline to gain more accurate outcomes along with faster detection speed to understand the content of the image rather than analyzing the entire image. In the first stage, using a Scale-based Region Proposal Network, the location of the text is estimated. Next, a Fully Convolutional Network (FCN) is implemented to attain an accurate localization result. The described strategies suffer from intolerable outcomes in recognizing the vertical text in the real scene, especially in the images with low illumination and low contrast scenes. Also, these state-of-the-art techniques cannot properly identify the orientation and location of the text efficiently. These problems lead to uncertainty in some applications such as blind assistance systems and driver assistance systems. Therefore, to overcome these problems in this study, a deep learning strategy is proposed to reduce the bad influence of the complex background that is robust to variations in color, scale, and rotation. To address the problem of lacking color information like Red, Green, and Blue (RGB), a multi-channel MSER technique was introduced by [49]. Their model combined the enhanced multi-channel MSER focusing on the region and Canny edge detector concentrating on the edge, where the channels employed in MSER consist of B, G, and R channels of the RGB color space and the S channel of the Hue, Saturation, and Intensity (HSI) color space.

This study is structured as follows. In Section 2, the proposed methodology is described in detail. In Section 3, the experiment results and comparison with some recently published pipelines are investigated. Section 4 concludes the study and gives an outlook for future studies.

2. Methodology

In this study, a lightweight CNN architecture for text localization and detection is proposed that aims to detect texts in the real scene, even if the text rotation is 90°. This contribution aims to employ a convolutional neural network for localizing text more precisely. Moreover, some intermediate time-consuming phases including word partitioning, finding the most possible region of occurring, and text region formation are eliminated [50]. The proposed structure is demonstrated in Fig. 1. The developed methodology in this research is capable to detect both the location and rotation of the text and works well for the complex background. The structure of the combination of the 3×3 $new.mReLU$ block and the MaxPooling layer at the beginning of the network is used for low-level visual feature extraction and plays a key role in the final results [50]. This network for extracting mid-level and high-level features employs 5 $new.mReLU$ blocks followed by 10 inception blocks. As shown in Fig. 2, the output of the final inception and $new.mReLU$ blocks are considered as the input of the four 1×1 Convolution Layers (Conv layers). These four convolution layers and the next 5×5 convolution and negative layers aim to recognize the vertical text.

Furthermore, an additional layer is applied to increase the efficiency of the feature extraction. Moreover, at the beginning of the proposed structure a new m.ReLU is utilized [51]. The implemented 3×3 $new.mReLU$ block is illustrated in Fig. 2. The intermediate activation patterns in the CNNs are the main motivation for applying this module inside the proposed model [52]. In this part, the production results obtained from the Negation and Conv layers need to be concatenated [53]. Additionally, to ease the computational burden, a separated bias layer is applied which causes the correlated kernels capable of having dissimilar bias weights.

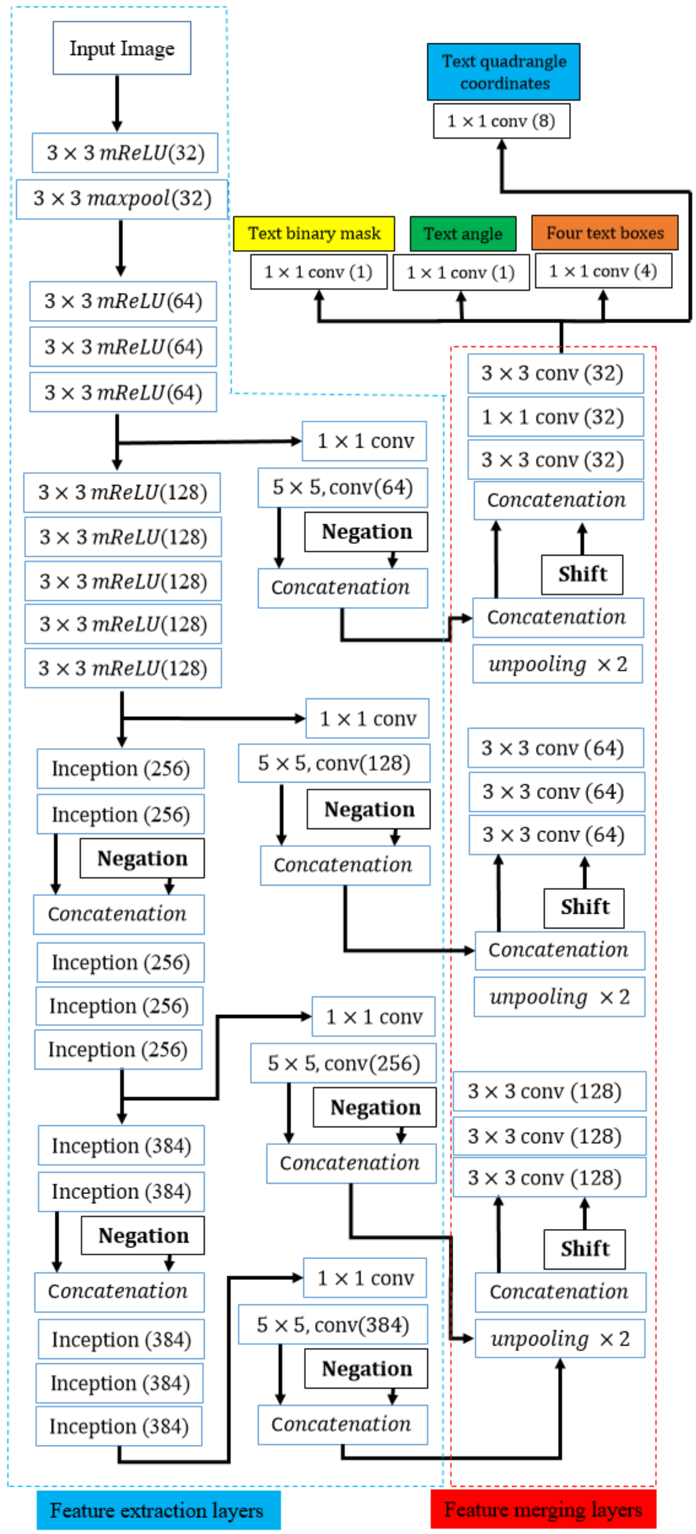

Fig. 1. Proposed pipeline.

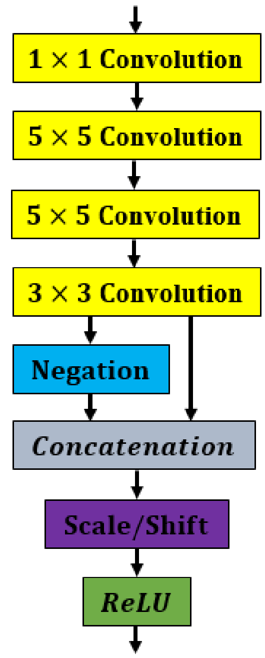

Fig. 2. Proposed 3×3 $new.mReLU$ building block.

In the $new.mReLU$ block, to overcome the limitation of the low contrast, the convolution output simply multiplies -1 by **Negation**. This Negation layer decreases the search space, compared to the recently published papers which predict four coordinates of an object [54]. Also, the trainable weights and biases can be applied to the next layer by **Scale/Shift** [55]. As the goal is to find a unique characteristic of text that can be determined for each text component at all levels, a **Scale/Shift** layer plays a key role in this purpose. It is also more dependable to estimate the rate of the text curvature to extract each character based on a distance ratio than identifying only a predefined distance between each character. To end this, the **Scale/Shift** layer needs to be after the concatenation layer. Using the $new.mReLU$ layer allows us to extract some low-level features suitably and causes robustness to font distortion and variation. Moreover, to address the issue of the complex background, three sequences of this layer at the beginning of the network have been employed [56].

A crucial step for achieving significant text detection results is exploring all potential areas inside the image including different scales, colors, and sizes [57]. It should be mentioned that extracted information from the different colorful textures plays a core role in improving the feature extraction procedure. This is because of intense color similarity in the most of characters and texts in the natural scene text (like warning traffic sign boards). Consequently, the proposed inception block is represented for improving the localization of the multi-scale and multi-orientation texts and preventing the production of more false-positive rates [55]. The proposed inception pipeline can be observed in Fig. 3.

Our new inception block is inspired by some of the suggestions implemented in Szegedy et al. [58] and comprises three parallel convolutional networks at the first, and one sequential concatenation and convolutional layers at the end of the block [59]. The core idea applied by the inception pipeline is eliminating the Conv layer and employing various parallel architectures to cover a larger region whereas a fine resolution can be obtained [60]. This approach forces multiplicity on the obtained features from each layer by merging feature maps at the end of the inception block and indicating a diminishing rate in the number of parameters. By overcoming the problems of changing the size of the text font employing this block, the accuracy of the final system output has successfully been improved. Here, it is realized that to attain an improvement in detecting largely varying-sized text, using stacking up inception layers is further useful than a simple linear chain of Conv layers [61]. Besides, to make the system more powerful to explore the location of the text with the minimum number of parameters the size of receptive fields is altered. Additionally, to ease the computational burden, two concatenation steps after extracting features in 3×3 Conv layers were employed.

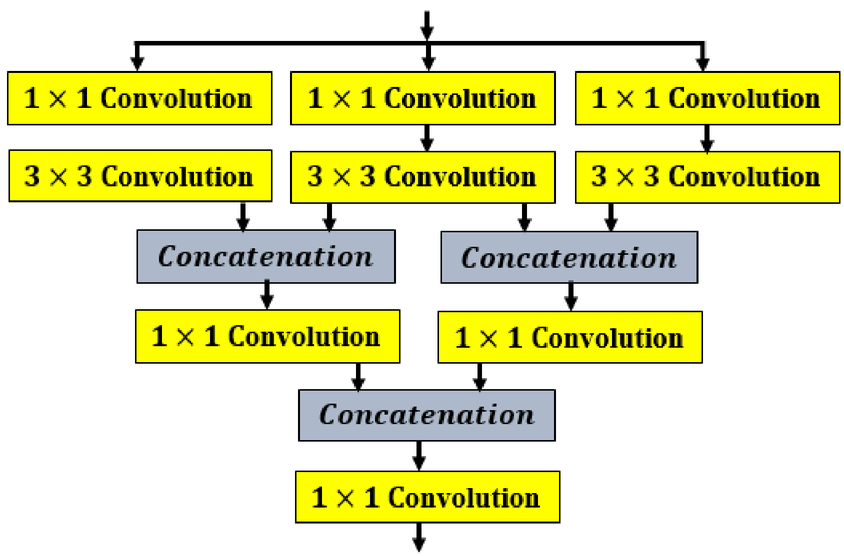

Fig. 3. Suggested inception building block.

As indicated in Fig. 1, stacking up inception layers are able to detect more varying-sized texts in an effective way compared to a chain of Conv layers. Owing to the use of the suggested inception building block, the output feature maps are produced with the same dimension of receptive fields. The output of the proposed structure can be implied as four detached vectors of features (both 1D and 2D vectors). The first feature vector is a 2D binary matrix (binary image) that is generated by considering the value of one for pixels inside the text window whereas others are represented by zero values. The next output vector is defined as the rotation of the text. The third output feature maps naming ***R*** matrices are represented by four **a**xis-**a**ligned **b**ounding **b**oxes (AABB). These output feature maps (2D vectors) can be considered as the distance of pixels to four corners of the obtained window that is fitted to the outer profile of the text [14]. Lastly, eight 1D vectors are generated to imply the corners of the text box's location (four corners) in the y and x directions. It means each corner can be defined by two distance variables: *dx* and *dy*. Then, the rotation map of text (rotation of each character) is calculated inside the described box with acceptable accuracy and demonstrated in a grayscale image. In order to attain the corners of the text box's location (eight channels), $LOC = \{p_i | i \in \{1,2,3,4\}\}$ is taken into account, where these vertices are defined by $p_i = \{x_i, y_j\}$. Besides, the reference length ref_i can be calculated for each vertex p_i as:

$$ref_i = \min\left(dist\left(p_i, p_{(i \bmod 4)+1}\right), dist\left(p_i, p_{((i+3) \bmod 4)+1}\right)\right), \tag{1}$$

where $dist(p_i, p_j)$ represents the Euclidean distance between p_j and p_i.

This binary mask (score map) can be produced by applying shrinking on each edge of the box by $0.38 \times ref_i$ and $0.38 \times ref_{(i \bmod 4)+1}$. In other words, to fit the obtained window around the text, the distances between these obtained 8 indices (or 8 channels) and 8 corners of the text inside the scene should be minimized. Hence, the value of 0.4 was chosen based on many experiments.

The loss function for text detection can be formulated as:

$$Loss_{total} = loss_{Score\ map} + \lambda_{Geometry}\, Loss_{Geometry}, \tag{2}$$

where $loss_{Score\ map}$ (oriented class-balanced cross-entropy) indicates the losses for the score map and $Loss_{Geometry}$ indicates the losses geometry. Moreover, $\lambda_{Geometry}$ implies the balancing weights t between two losses for achieving more robustness and accuracy.

The oriented class-balanced cross-entropy for minimizing the loss of score map is calculated using Equation (3):

$$loss_{Score\ map} = \varsigma\ out_{reference} \log out_{predicted} - (1 - \varsigma)(1 - out_{reference}) \log(1 - out_{predicted}), \tag{3}$$

where ς indicates the oriented text balancing factor between negative and positive samples, given by Equation (4):

$$\varsigma = \left(\frac{\sum_{corners \in out_{reference\ (i)}} corners - \sum_{corners \in out_{reference(i+1)}} corners}{\sum_{corners \in out_{predicted\ (i)}} Ccorners - \sum_{corners \in out_{predicted(i+1)}} corners} \right) \times \left(1 - \frac{\sum_{T \in out_{reference}} T}{|out_{reference}|} \right) \tag{4}$$

where i represents the current detected text and $i + 1$ demonstrates the adjacent detected text as shown in Fig. 4.

By considering the effect of distances between the current detected text and the adjacent detected text, the suggested network is able to predict the rotation of the text more efficiently. In this study, $\lambda_{Geometry}$ is set to 0.83 and the proposed structure was learned in 1,000 epochs with a batch size of 128, a learning rate of 0.01, and a weight decay of 0.0001. Furthermore, $Loss_{Geometry}$ can be described as:

$$Loss_{Geometry} = \alpha_\theta loss_\theta + Loss_{AABB}, \tag{5}$$

$$Loss_{AABB} = -\log\left(\frac{|reference \cap predicted|}{|reference \cup predicted|} \right), \tag{6}$$

where $predicted$ shows the calculated AABB geometry and $reference$ is the related Ground Truth (GT). Furthermore, by defining dis_1, dis_2, dis_3, and dis_4 as the distance from a pixel to the bottom, left, top, and right boundary of its corresponding window, the height and width of the intersected rectangle $|reference \cap predicted|$ are calculated using Equations (7) and (8):

$$\text{width} = \min\left(dis_{2(reference)}, dis_{2(predicted)}\right) + min\left(dis_{4(reference)}, dis_{4(predicted)}\right), \tag{7}$$

$$\text{height} = \min\left(dis_{1(reference)}, dis_{1(predicted)}\right) + min\left(dis_{3(reference)}, dis_{3(predicted)}\right). \tag{8}$$

Moreover, the union region can be calculated by Equation (9):

$$|reference \cup predicted| = |reference| + |predicted| - |reference \cap predicted|. \tag{9}$$

Then, the loss function of rotation angle is given by Equation (10):

$$L_\theta\left(\theta_{predicted}, \theta_{reference}\right) = 1 - \cos\left(\theta_{predicted} - \theta_{reference}\right). \tag{10}$$

Furthermore, it is identified that using only a 3×3 convolution layer after up-sampling and concatenation layers (see Fig. 1) causes a difficulty to precisely recognize the horizontal sides of words in the case of observing text on a curve. This is due to the fact that the distance of each character to the adjacent character is uneven for the upper and bottom parts of it which changes the shape of the components. In other words, as it is illustrated in Fig. 4, within a word in a curved line, it is a confusing task to find the exact distance between each character. Moreover, whatever the two borders of the text are closer together (see Fig. 4 (B)) the distance between the upper parts of the words or characters is bigger and vice versa. To overcome this problem, two 3×3 convolution layers have been utilized after the up-pooling and concatenation layers. Furthermore, as mentioned before, the first $new.mReLU$ layers are crucial for obtaining acceptable results. Hence, the impacts of the number of this layer are demonstrated in Fig. 5.

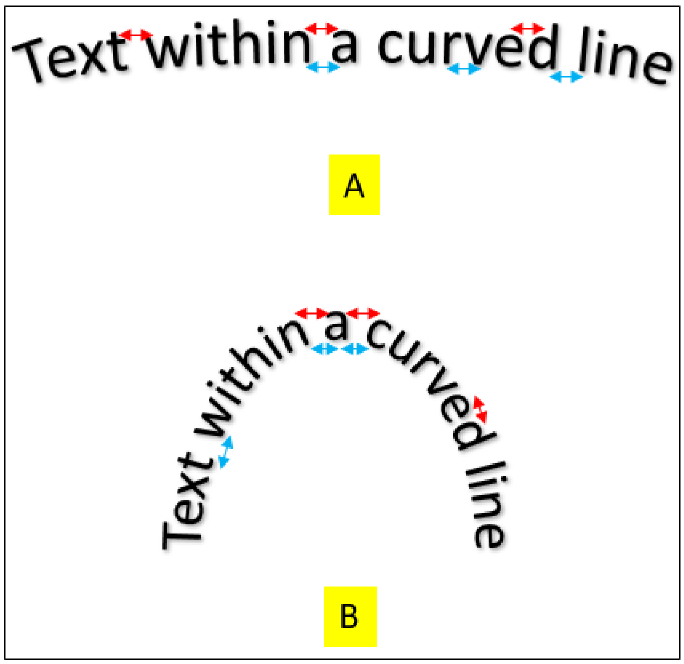

Fig. 4. Two examples of observing text in the real scene. (A) A sample text with a small curvature. (B) A sample text with high curvature. The red arrows indicate a bigger distance than the blue arrows. As it is clearly demonstrated the red arrows inside the (B) are bigger than the blue arrows, whilst these arrows are small differences in (A).

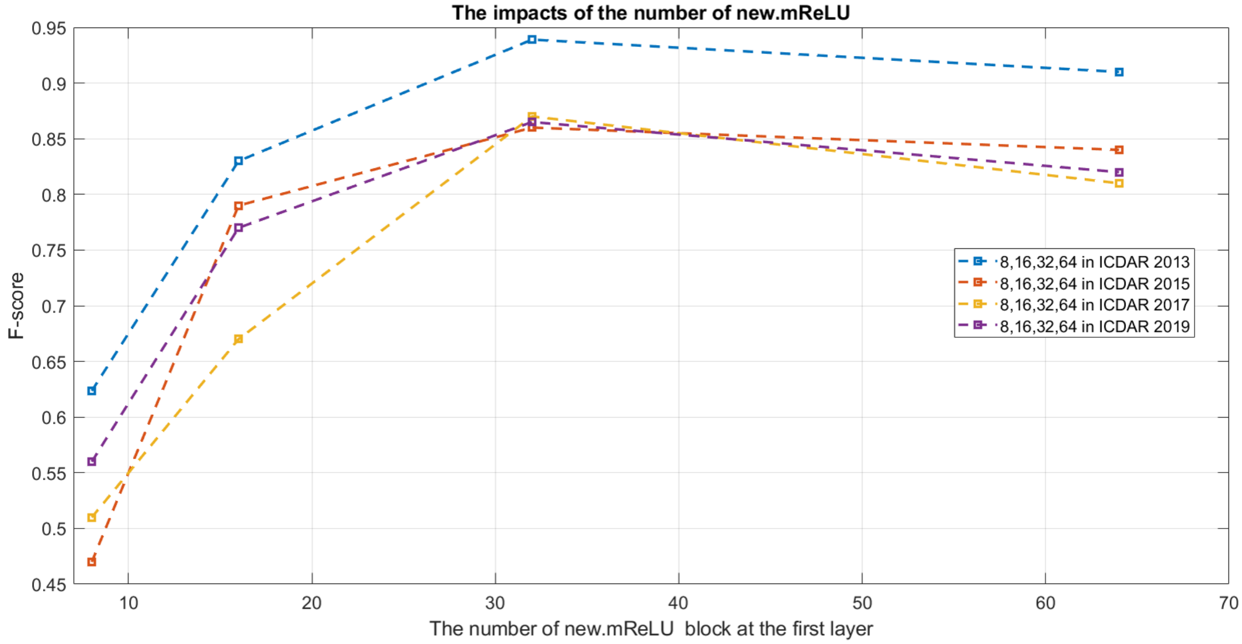

Fig. 5. Impacts of the number of the $new.mReLU$ blocks on all tested datasets.

3. Experiments

3.1 Datasets

In order to assess the proposed method, the datasets of ICDAR 2013 [62], ICDAR 2015 [63], ICDAR 2017 [64], and ICDAR 2019 [65] are utilized. They have been cited and used by several recent scene text research works. The ICDAR 2013 is based on the horizontal text which includes 229 and 223 images for training and testing, respectively. Also, the ICDAR 2015 is based on multi-oriented text with 1000 and 500 images for training and testing, respectively [56]. Moreover, the ICDAR 2017 consists of 1555 images with various text orientations. Finally, the ICDAR 2019 consists of 10,000 images for robust text locating [66]. Text detection results are illustrated in Figs. 6, 7, 8, and 9. It has been illustrated that text orientation and location can be successfully detected by the suggested algorithm.

3.2 Evaluation metrics

In this study, the following three measures, namely f-measure (F), recall (R), and precision (P), have been used to evaluate the developed model and compare the text detection results with some state-of-the-art approaches. These metrics can be defined as follows [67]:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{12}$$

$$F = \frac{2 \times Precision \times \ Recall}{Precision + \ Recall}, \tag{13}$$

where the FN, FP, and TP respectively represent the false negative, false positive, and true positive [68]. The outstanding results and experiments were accomplished utilizing Python on an Intel 3.2 GHz-Core I7 computer with a 64-bit operating system.

3.3 Experimental Results

In order to verify the performance and robustness of the suggested approach, it is compared with 10 state-of-the-art text localization pipelines. For a more clear understanding, vertical text localization is depicted in Fig. 8. Due to the trade-off between recall and precision result rate, the f-score is the best evaluation for analyzing the results of a text detection system. The outcomes are described and compared with the other pipelines in Tables 1-4. For each index in all tables, the highest values are highlighted in bold. By analyzing the indicated outcomes in Tables 1-4, it is obvious that the proposed pipeline has gained the best outcomes in comparison with all mentioned detection architectures. The notable obtained outcomes prove that the given strategy meaningfully improves the accuracy of the model even with the presence of texts with 90° orientation in the scene. Furthermore, to exemplify the importance of implementing the proposed network to accurately estimate the text location, Figs. 6-9 demonstrate the outcomes of the offered structure.

The effectiveness and accuracy of the proposed strategy are first investigated on a popular horizontal text dataset, namely the ICDAR 2013 dataset. As clearly shown in Table 1, the proposed pipeline obtains competitive performance both in terms of efficiency and accuracy. Although the CRAFT [15] and achieves the highest precision, the highest Recall and F-measure are obtained by the proposed methods. The Recall of TextBox MS [69] is only next to LocNet [57] and SRPN [48]. Moreover, Fast TextBox [69] obtains the worst results in all three measures. Examples of text detection on the ICDAR 2013 dataset are illustrated in Fig. 6.

Table 1. Results on ICDAR 2013.

Method	Recall	Precision	F-score
AF-RPN [70]	0.900	0.934	0.916
FTPN [5]	0.919	0.932	0.925
TextBox MS [69]	0.830	0.880	0.850
Fast TextBox [69]	0.740	0.860	0.800
Pyramid Context Network [28]	0.905	0.938	0.921
DeRPN [71]	0.774	0.867	0.818
LocNet [57]	0.875	0.940	0.906
CRAFT [15]	0.931	**0.974**	0.952
Delaunay Triangulation (DT) [72]	0.904	0.88	0.891
SRPN [48]	0.842	0.925	0.882
Multi-channel MSER [49]	0.937	0.894	0.915
The proposed approach	**0.942**	0.956	**0.948**

Fig. 6. Example of four text localization by the proposed method on ICDAR 2013. It is shown that the proposed method is capable of localizing the oriented text successfully.

By analyzing the outcomes achieved on ICDAR 2015 in Table 2, it is found out that there is not much difference between the minimum and maximum values based on the Recall criteria. Accordingly, it can noticeably be seen that the worst scores for Precision and Recall were obtained using EAST+VGG16 [14] and SegLink [73], respectively. Results obtained using PixelLink+VGG16 4s [12] and PixelLink+VGG16 2s [12] are very close to the proposed network regarding the Recall; however, it still failed to detect words as the Precision score demonstrates. Deep Direct Regression [4] and SegLink [73] methods cannot gain acceptable results, especially in the presence of a complex background. PixelLink+VGG16 2s [12], Multi-channel MSER [49], PixelLink+VGG16 4s [12], and Mask R-CNN [74] approaches are good to extract the oriented text when there is much similarity between two completely separated words, whilst they perform so poorly when encountering two close words. Moreover, EAST+PVANET2x MS [14] and EAST+PVANET2x [14] models are more prone to fail, especially when there are fuzzy boundaries. Finally, the developed approach reaches the best performance with the ICDAR 2015 dataset, followed by Mask R-CNN [74] which has a small difference in Precision score. Fig. 7 depicts that the suggested model has a powerful ability to detect curved texts. It is even able to read words within a short distance.

Table 2. Results on ICDAR 2015.

Method	Recall	Precision	F-score
EAST+VGG16 [14]	0.727	0.804	0.764
EAST+PVANET2x [14]	0.734	0.835	0.782
EAST+PVANET2x MS [14]	0.7563	0.7712	0.7516
SegLink [73]	0.768	0.731	0.750
Mask R-CNN [74]	0.815	0.908	0.859
Deep Direct Regression [4]	0.800	0.820	0.810
PixelLink+VGG16 4s [12]	0.817	0.829	0.823
PixelLink+VGG16 2s [12]	0.820	0.855	0.837
Direct Regression [2]	0.800	0.850	0.820
SRPN [48]	0.796	0.920	0.853
Adaptive scale fusion [47]	0.839	0.909	0.873
Kernel Proposal Network [75]	0.869	0.878	0.873
Multi-channel MSER [49]	0.922	0.894	0.903
The proposed approach	**0.931**	**0.924**	**0.927**

Fig. 7. Example of four text localization by the suggested method on ICDAR 2015. It is shown that the proposed method is capable of localizing the oriented text successfully.

As indicated in Table 3, text detection and segmentation by employing AF-RPN [70] and CLRS [76] imply the fewest match with the ground truth, especially when there are vertical texts. This is due to the fact that the vertical and horizontal texts exhibit different characteristics. Moreover, PSENet [77] obtains the worst Precision score amongst all evaluated approaches. Compared with previous state-of-the-art pipelines in the field of text localization, the developed pipeline in this work demonstrates the advantage in terms of Recall, Precision, and F-score. Delaunay Triangulation outperformed Mask R-CNN [74] and reached competitive outcomes against state-of-the-art algorithms (AF-RPN, PSENet, TSL, and ISNet). AF-RPN [70] and ISNet [78] models had issues identifying vertical word cases and when it does, they were detected with a very low confidence value. Delaunay Triangulation method [72] was very close to the developed approach regarding the Recall; however, it still failed to detect words as the Precision score demonstrates. Fig. 8 depicts that the suggested model has a powerful ability to detect curved texts. It is even able to read words within a short distance.

Table 3. Results on ICDAR 2017.

Method	Recall	Precision	F-score
AF-RPN [70]	0.667	0.794	0.725
PSENet [77]	0.753	0.691	0.721
TSL [79]	0.674	0.776	0.722
Delaunay Triangulation (DT) [72]	0.83	0.72	0.771
ISNet [78]	0.674	0.78	0.723
CLRS [76]	0.556	0.838	0.668
Mask R-CNN [74]	0.698	0.8	0.743
Multi-channel MSER [49]	0.806	0.764	0.784
The proposed approach	**0.874**	**0.867**	**0.870**

Fig. 8. Example of four text localization by the suggested method on ICDAR 2017. It is shown that the proposed method is capable of localizing the vertical and curved texts magnificently.

Experimental outcomes on the ICDAR 2019 illustrate that the developed pipeline outperformed well-known techniques, such as LOMO [80], Pyramid Context Network [28], and PixelLink+VGG16 2s [14] not only in effectiveness and accuracy but also in terms of the size of the network. As indicated in Table 4, text identification by applying Fast TextBox [69], EAST+PVANET2x [14], and TextBox MS [69] entails the fewest match with the ground truth, especially when there are vertical texts. From the obtained outcomes and Fig. 9, it can be observed that the identification is able to automatically be adapted to any kind of text, even with the different distances between characters. Concerning the best structures for detecting a text and their networks dimension, PixelLink+VGG16 2s [12], PixelLink+VGG16 4s [12], and Pyramid Context Network [28] achieved better results than Fast TextBox [69] and EAST+PVANET2x [14]; nevertheless, their models were larger than the proposed network. On the whole, the experimental outcomes imply the superiority of the developed approach.

Table 4. Results on ICDAR 2019.

Method	Recall	Precision	F-score
LOMO [80]	0.798	0.878	0.836
EAST+PVANET2x [14]	0.751	0.816	0.782
TextBox MS [69]	0.775	0.884	0.825
Fast TextBox [69]	0.753	0.845	0.796
Pyramid Context Network [28]	0.815	0.846	0.830
PixelLink+VGG16 4s [12]	0.823	0.821	0.821
PixelLink+VGG16 2s [12]	0.820	0.855	0.837
The proposed approach	**0.842**	**0.891**	**0.865**

Fig. 9. Example of four text localization by the suggested method on ICDAR 2019. It is shown that the proposed method is capable of localizing the oriented texts magnificently.

DL-based techniques have a key drawback that it is challenging for determining the basis of the proposed network judgment. The common technique to clarify the reason for the model prediction is visual description. The visual description technique illustrates an attention map that pictures an area in which the model concentrated as a heat map [81]. According to an achieved attention map, the reason for the segmentation or classification results can be understood and analyzed. In order to gain a clearer and more explainable attention map for a well-organized visual description, a number of techniques such as Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM) have been suggested in the field of computer vision [82].

In this study, the Grad-CAM method produces an attention map by utilizing gradient values computed at the backpropagation process. Fig. 10 illustrates example attention maps of Grad-CAM.

Fig. 10. Two examples of Grad-CAM of the suggested network. The first row and the second row indicate the original and Grad-CAM images, respectively. Color denotes the degree of activation: very low (blue), low (green), high (yellow) and very high (red).

4. Conclusion

In this research, a new real scene text detection pipeline was implemented based on an inception structure that can produce a location binary mask along with its rotation. The proposed structure overcame some problems such as local and global illumination variations, occlusion, a wide range of styles and colors, unpredictable orientations, and various sizes. This strategy was also capable of discovering even vertical texts in a real scene. By incorporating an extra layer for feature extraction and an optimized inception layer, the detector can find the text location more accurately.

Our structure was based on the combination of the $new.mReLU$ and inception structure. Because of utilizing $new.mReLU$ and inception blocks, text recognition also can be implemented more precisely and efficiently. Experimental comparisons with the state-of-the-art structures on four datasets; ICDAR2013, ICDAR2015, ICDAR2017, and ICDAR2019 depicted the efficiency and effectiveness of the developed approach for the text localization and recognition task. Each of these datasets is recorded in different environments with various image resolutions and light conditions. As there are some restrictions for detecting text in the presence of a complex background, the most important idea to extend this study is to use a transform learning approach for increasing the accuracy of the developed approach.

Funding This publication has emanated from research [conducted with the financial support of/supported in part by a grant from Science Foundation Ireland under Grant number No. 18/CRT/6183 and is supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106/_P2), Lero SFI Centre for Software (Grant 13/RC/2094/_P2) and is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Declarations Ethics Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Not applicable.

Conflict of Interest The authors declare that they have no conflict of interest regarding this work.

References

[1] S. Y. Arafat, N. Ashraf, M. J. Iqbal, I. Ahmad, S. Khan, and J. J. P. C. Rodrigues, “Urdu signboard detection and recognition using deep learning,” Multimed. Tools Appl., vol. 81, no. 9, pp. 11965–11987, Apr. 2022, doi: 10.1007/S11042-020-10175-2/FIGURES/14.

[2] W. He, X. Y. Zhang, F. Yin, and C. L. Liu, “Multi-Oriented and Multi-Lingual Scene Text Detection with Direct Regression,” IEEE Trans. Image Process., vol. 27, no. 11, pp. 5406–5419, Nov. 2018, doi: 10.1109/TIP.2018.2855399.

[3] A. Aiman, Y. Shen, M. Bendechache, I. Inayat, and T. Kumar, “AUDD: Audio Urdu Digits Dataset for Automatic Audio Urdu Digit Recognition,” Appl. Sci. 2021, Vol. 11, Page 8842, vol. 11, no. 19, p. 8842, Sep. 2021, doi: 10.3390/APP11198842.

[4] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, “Deep Direct Regression for Multi-Oriented Scene Text Detection,” 2017.

[5] F. Liu, C. Chen, D. Gu, and J. Zheng, “FTPN: Scene text detection with feature pyramid based text proposal network,” IEEE Access, vol. 7, pp. 44219–44228, 2019, doi: 10.1109/ACCESS.2019.2908933.

[6] G. Jawahar, M. Abdul-Mageed, and L. V. S. Lakshmanan, “Automatic Detection of Machine Generated Text: A Critical Survey,” Nov. 2020, Accessed: Dec. 26, 2020. [Online]. Available: http://arxiv.org/abs/2011.01314.

[7] N. Tataei Sarshar et al., “Glioma Brain Tumor Segmentation in Four MRI Modalities Using a Convolutional Neural Network and Based on a Transfer Learning Method,” pp. 386–402, 2023, doi: 10.1007/978-3-031-04435-9_39.

[8] W. ; Khan et al., “Introducing Urdu Digits Dataset with Demonstration of an Efficient and Robust Noisy Decoder-Based Pseudo Example Generator,” Symmetry 2022, Vol. 14, Page 1976, vol. 14, no. 10, p. 1976, Sep. 2022, doi: 10.3390/SYM14101976.

[9] L. Zou, Z. Wang, and D. Zhou, “Moving horizon estimation with non-uniform sampling under component-based dynamic event-triggered transmission,” Automatica, vol. 120, p. 109154, Oct. 2020, doi: 10.1016/j.automatica.2020.109154.

[10] R. Ranjbarzadeh and S. Baseri Saadi, “Corrigendum to ‘Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering’ [Measurement 150 (2020) 107086],” Measurement, vol. 151, p. 107230, Feb. 2020, doi: 10.1016/J.MEASUREMENT.2019.107230.

[11] S. Long, X. He, and C. Yao, “Scene Text Detection and Recognition: The Deep Learning Era,” Int. J. Comput. Vis., pp. 1–24, Aug. 2020, doi: 10.1007/s11263-020-01369-0.

[12] D. Deng, H. Liu, X. Li, and D. Cai, “PixelLink: Detecting Scene Text via Instance Segmentation,” 32nd AAAI Conf. Artif. Intell. AAAI 2018, pp. 6773–6780, Jan. 2018, Accessed: Dec. 24, 2020. [Online]. Available: http://arxiv.org/abs/1801.01315.

[13] X. Bian, C. Wang, W. Quan, J. Ye, X. Zhang, and D. M. Yan, “Scene text removal via cascaded text stroke detection and erasing,” Comput. Vis. Media 2021 82, vol. 8, no. 2, pp. 273–287, Dec. 2021, doi: 10.1007/S41095-021-0242-8.

[14] X. Zhou et al., “EAST: An Efficient and Accurate Scene Text Detector,” 2017.

[15] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character Region Awareness for Text Detection,” 2019.

[16] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, “ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network *,” 2020.

[17] E. Kropat, G.-W. Weber, and E. B. Tirkolaee, “Foundations of semialgebraic gene-environment networks,” J. Dyn. Games. 2020, Vol. 7, Pages 253-268, vol. 7, no. 4, p. 253, Jul. 2020, doi: 10.3934/JDG.2020018.

[18] A. Özmen, E. Kropat, and G. W. Weber, “Robust optimization in spline regression models for multi-model regulatory networks under polyhedral uncertainty,” vol. 66, no. 12, pp. 2135–2155, Dec. 2016, doi: 10.1080/02331934.2016.1209672.

[19] E. Kropat, A. Ozmen, G. W. Weber, S. Meyer-Nieberg, and O. Defterli, “Fuzzy prediction strategies for gene-environment networks – Fuzzy regression analysis for two-modal regulatory systems,” RAIRO - Oper. Res., vol. 50, no. 2, pp. 413–435, Apr. 2016, doi: 10.1051/RO/2015044.

[20] T. J. S.-H. Kumar, “Intra-Class Random Erasing (ICRE) augmentation for audio classification,” Proc. Korean Soc. Broadcast Eng. Conf., pp. 244–247, 2020.

[21] B. Kalaycı, A. Özmen, and G. W. Weber, “Mutual relevance of investor sentiment and finance by modeling coupled stochastic systems with MARS,” Ann. Oper. Res., vol. 295, no. 1, pp. 183–206, Dec. 2020, doi: 10.1007/S10479-020-03757-8/TABLES/1.

[22] I. A. Khan et al., “XSRU-IoMT: Explainable simple recurrent units for threat detection in Internet of Medical Things networks,” Futur. Gener. Comput. Syst., vol. 127, pp. 181–193, Feb. 2022, doi: 10.1016/J.FUTURE.2021.09.010.

[23] A. K. Sahu, S. Sharma, M. Tanveer, and R. Raja, “Internet of Things attack detection using hybrid Deep Learning Model,” Comput. Commun., vol. 176, pp. 146–154, Aug. 2021, doi: 10.1016/J.COMCOM.2021.05.024.

[24] Z. Yue et al., “Privacy-preserving Time-series Medical Images Analysis Using a Hybrid Deep Learning Framework,” ACM Trans. Internet Technol., vol. 21, no. 3, Jun. 2021, doi: 10.1145/3383779.

[25] R. Sharma, T. Goel, M. Tanveer, and R. Murugan, “FDN-ADNet: Fuzzy LS-TWSVM based deep learning network for prognosis of the Alzheimer’s disease using the sagittal plane of MRI scans,” Appl. Soft Comput., vol. 115, p. 108099, Jan. 2022, doi: 10.1016/J.ASOC.2021.108099.

[26] S. Dwivedi, T. Goel, M. Tanveer, R. Murugan, and R. Sharma, “Multimodal Fusion-Based Deep Learning Network for Effective Diagnosis of Alzheimer’s Disease,” IEEE Multimed., vol. 29, no. 2, pp. 45–55, 2022, doi: 10.1109/MMUL.2022.3156471.

[27] A. Chakraborty, D. Ganguly, A. Caputo, and G. J. F. Jones, “Kernel Density Estimation based Factored Relevance Model for Multi-Contextual Point-of-Interest Recommendation,” Inf. Retr. J., vol. 25, no. 1, pp. 44–90, Jun. 2020, doi: 10.48550/arxiv.2006.15679.

[28] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, “Scene text detection with supervised pyramid context network,” in 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019,

Jul. 2019, vol. 33, no. 01, pp. 9038–9045, doi: 10.1609/aaai.v33i01.33019038.

[29] A. Baghban, M. Bahadori, A. S. Lemraski, and A. Bahadori, “Prediction of solubility of ammonia in liquid electrolytes using Least Square Support Vector Machines,” Ain Shams Eng. J., vol. 9, no. 4, pp. 1303–1312, Dec. 2018, doi: 10.1016/J.ASEJ.2016.08.006.

[30] Z. Liu and A. Baghban, “Application of LSSVM for biodiesel production using supercritical ethanol solvent,” vol. 39, no. 17, pp. 1869–1874, Oct. 2017, doi: 10.1080/15567036.2017.1380732.

[31] M. T. S. T. R. B. and M. B. Teerath Kumar et al., “Forged Character Detection Datasets: Passports, Driving Licences and Visa Stickers,” Int. Artif. Appl. Vol.13,No.2, vol. 13, no. 2, p. 21, Mar. 2022, doi: 10.5121/IJAIA.2022.13202.

[32] B. Du, J. Ye, J. Zhang, J. Liu, and D. Tao, “I3CL: Intra- and Inter-Instance Collaborative Learning for Arbitrary-Shaped Scene Text Detection,” Int. J. Comput. Vis., vol. 130, no. 8, pp. 1961–1977, Aug. 2022, doi: 10.1007/S11263-022-01616-6/FIGURES/11.

[33] T. Kumar, J. Park, M. S. Ali, A. F. M. Shahab Uddin, J. H. Ko, and S.-H. Bae, “Binary-Classifiers-Enabled Filters for Semi-Supervised Learning,” IEEE Access, pp. 1–1, 2021, doi: 10.1109/ACCESS.2021.3124200.

[34] R. Ranjbarzadeh et al., “Nerve optic segmentation in CT images using a deep learning model and a texture descriptor,” Complex Intell. Syst. 2022, pp. 1–15, Feb. 2022, doi: 10.1007/S40747-022-00694-W.

[35] R. Ranjbarzadeh et al., “MRFE-CNN: multi-route feature extraction model for breast tumor segmentation in Mammograms using a convolutional neural network,” Ann. Oper. Res. 2022, pp. 1–22, May 2022, doi: 10.1007/S10479-022-04755-8.

[36] A. Aghamohammadi, R. Ranjbarzadeh, F. Naiemi, M. Mogharrebi, S. Dorosti, and M. Bendechache, “TPCNN: Two-path convolutional neural network for tumor and liver segmentation in CT images using a novel encoding approach,” Expert Syst. Appl., vol. 183, p. 115406, Nov. 2021, doi: 10.1016/J.ESWA.2021.115406.

[37] X. Liu and W. Wang, “An effective graph-cut scene text localization with embedded text segmentation,” Multimed. Tools Appl., vol. 74, no. 13, pp. 4891–4906, Jun. 2015, doi:

10.1007/s11042-013-1848-3.

[38] R. Ranjbarzadeh, S. B. Saadi, and A. Amirabadi, “LNPSS: SAR image despeckling based on local and non-local features using patch shape selection and edges linking,” Meas. J. Int. Meas. Confed., vol. 164, Nov. 2020, doi: 10.1016/j.measurement.2020.107989.

[39] Y. Tang and X. Wu, “Scene text detection using superpixel-based stroke feature transform and deep learning based region classification,” IEEE Trans. Multimed., vol. 20, no. 9, pp. 2276–2288, Sep. 2018, doi: 10.1109/TMM.2018.2802644.

[40] G. Nalcaci, A. Özmen, and G. W. Weber, “Long-term load forecasting: models based on MARS, ANN and LR methods,” Cent. Eur. J. Oper. Res., vol. 27, no. 4, pp. 1033–1049, Dec. 2019, doi: 10.1007/S10100-018-0531-1/FIGURES/9.

[41] S. Shamshirband, P. Saraei, N. Nabipour, and A. Baghban, “Hydrocarbons density estimates for a wide range of conditions using RBF-ANN and ANFIS strategies,” 2019, doi: 10.1080/15567036.2019.1704313.

[42] M. Turab, T. Kumar, M. Bendechache, and T. Saber, “Investigating Multi-Feature Selection and Ensembling for Audio Classification,” Jun. 2022, doi: 10.48550/arxiv.2206.07511.

[43] G. R. Kanagachidambaresan, A. Ruwali, D. Banerjee, and K. B. Prakash, “Recurrent Neural Network,” in EAI/Springer Innovations in Communication and Computing, Springer Science and Business Media Deutschland GmbH, 2021, pp. 53–61.

[44] S. M. Mousavi, A. Asgharzadeh-Bonab, and R. Ranjbarzadeh, “Time-Frequency Analysis of EEG Signals and GLCM Features for Depth of Anesthesia Monitoring,” Comput. Intell. Neurosci., vol. 2021, pp. 1–14, Aug. 2021, doi: 10.1155/2021/8430565.

[45] J. Ma et al., “Arbitrary-oriented scene text detection via rotation proposals,” IEEE Trans. Multimed., vol. 20, no. 11, pp. 3111–3122, Nov. 2018, doi: 10.1109/TMM.2018.2818020.

[46] P. Shivakumara, R. Raghavendra, L. Qin, K. B. Raja, T. Lu, and U. Pal, “A new multi-modal approach to bib number/text detection and recognition in Marathon images,” Pattern Recognit., vol. 61, pp. 479–491, Jan. 2017, doi: 10.1016/j.patcog.2016.08.021.

[47] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, “Real-Time Scene Text Detection with

Differentiable Binarization and Adaptive Scale Fusion," IEEE Trans. Pattern Anal. Mach. Intell., 2022, doi: 10.1109/TPAMI.2022.3155612.

[48] W. He, X. Y. Zhang, F. Yin, Z. Luo, J. M. Ogier, and C. L. Liu, "Realtime multi-scale scene text detection with scale-based region proposal network," Pattern Recognit., vol. 98, p. 107026, Feb. 2020, doi: 10.1016/j.patcog.2019.107026.

[49] G. Tong, M. Dong, X. Sun, and Y. Song, "Natural scene text detection and recognition based on saturation-incorporated multi-channel MSER," Knowledge-Based Syst., vol. 250, p. 109040, Aug. 2022, doi: 10.1016/J.KNOSYS.2022.109040.

[50] S. Hong, B. Roh, K.-H. Kim, Y. Cheon, and M. Park, "PVANet: Lightweight Deep Neural Networks for Real-time Object Detection," Nov. 2016, Accessed: Dec. 26, 2020. [Online]. Available: http://arxiv.org/abs/1611.08588.

[51] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units," 33rd Int. Conf. Mach. Learn. ICML 2016, vol. 5, pp. 3276–3284, Mar. 2016, Accessed: Dec. 26, 2020. [Online]. Available: http://arxiv.org/abs/1603.05201.

[52] S. J. Ghoushchi, R. Ranjbarzadeh, A. H. Dadkhah, Y. Pourasad, and M. Bendechache, "An Extended Approach to Predict Retinopathy in Diabetic Patients Using the Genetic Algorithm and Fuzzy C-Means," Biomed Res. Int., vol. 2021, pp. 1–13, Jun. 2021, doi: 10.1155/2021/5597222.

[53] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-Attentional Convolutional Neural Network for Scene Text Detection," IEEE Trans. Image Process., vol. 25, no. 6, pp. 2529–2541, Jun. 2016, doi: 10.1109/TIP.2016.2547588.

[54] S. Anari, N. Tataei Sarshar, N. Mahjoori, S. Dorosti, and A. Rezaie, "Review of Deep Learning Approaches for Thyroid Cancer Diagnosis," Math. Probl. Eng., vol. 2022, pp. 1–8, Aug. 2022, doi: 10.1155/2022/5052435.

[55] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, "PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection," ArXiv, vol. 2012, pp. 1–7, Aug. 2016, Accessed: Dec. 26, 2020. [Online]. Available: http://arxiv.org/abs/1608.08021.

[56] F. Naiemi, V. Ghods, and H. Khalesi, “A novel pipeline framework for multi oriented scene text image detection and recognition,” Expert Syst. Appl., vol. 170, p. 114549, May 2021, doi: 10.1016/j.eswa.2020.114549.

[57] Z. Zhong, L. Sun, and Q. Huo, “Improved localization accuracy by LocNet for Faster R-CNN based text detection in natural scene images,” Pattern Recognit., vol. 96, p. 106986, Dec. 2019, doi: 10.1016/j.patcog.2019.106986.

[58] C. Szegedy et al., “Going Deeper with Convolutions,” 2015.

[59] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, “Rethinking the Inception Architecture for Computer Vision,” 2016.

[60] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, “Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images,” Sci. Rep., vol. 11, no. 1, p. 10930, Dec. 2021, doi: 10.1038/s41598-021-90428-8.

[61] S. Baseri Saadi, N. Tataei Sarshar, S. Sadeghi, R. Ranjbarzadeh, M. Kooshki Forooshani, and M. Bendechache, “Investigation of Effectiveness of Shuffled Frog-Leaping Optimizer in Training a Convolution Neural Network,” J. Healthc. Eng., vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/4703682.

[62] D. Karatzas et al., “ICDAR 2013 robust reading competition,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2013, pp. 1484–1493, doi: 10.1109/ICDAR.2013.221.

[63] D. Karatzas et al., “ICDAR 2015 competition on Robust Reading,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Nov. 2015, vol. 2015-November, pp. 1156–1160, doi: 10.1109/ICDAR.2015.7333942.

[64] N. Nayef et al., “ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Jul. 2017, vol. 1, pp. 1454–1459, doi: 10.1109/ICDAR.2017.237.

[65] N. Nayef et al., “ICDAR2019 robust reading challenge on multi-lingual scene text detection

and recognition-RRC-MLT-2019,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Sep. 2019, pp. 1582–1587, doi: 10.1109/ICDAR.2019.00254.

[66] S. Saha et al., “Multi-lingual scene text detection and language identification,” Pattern Recognit. Lett., vol. 138, pp. 16–22, Oct. 2020, doi: 10.1016/j.patrec.2020.06.024.

[67] A. Fateh, M. Rezvani, A. Tajary, and M. Fateh, “Persian printed text line detection based on font size,” Multimed. Tools Appl., pp. 1–26, Jun. 2022, doi: 10.1007/S11042-022-13243-X/FIGURES/17.

[68] F. Tasnim, S. U. Habiba, N. Nafisa, and A. Ahmed, “Depressive Bangla Text Detection from Social Media Post Using Different Data Mining Techniques,” Lect. Notes Electr. Eng., vol. 834, pp. 237–247, 2022, doi: 10.1007/978-981-16-8484-5_21/COVER.

[69] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “TextBoxes: A Fast Text Detector with a Single Deep Neural Network,” 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 4161–4167, Nov. 2016, Accessed: Dec. 24, 2020. [Online]. Available: http://arxiv.org/abs/1611.06779.

[70] Z. Zhong, L. Sun, and Q. Huo, “An anchor-free region proposal network for Faster R-CNN-based text detection approaches,” in International Journal on Document Analysis and Recognition, Sep. 2019, vol. 22, no. 3, pp. 315–327, doi: 10.1007/s10032-019-00335-y.

[71] L. Xie, Y. Liu, L. Jin, and Z. Xie, “DeRPN: Taking a further step toward more general object detection,” in 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Jul. 2019, vol. 33, no. 01, pp. 9046–9053, doi: 10.1609/aaai.v33i01.33019046.

[72] S. Roy, P. Shivakumara, U. Pal, T. Lu, and G. H. Kumar, “Delaunay triangulation based text detection from multi-view images of natural scene,” Pattern Recognit. Lett., vol. 129, pp. 92–100, Jan. 2020, doi: 10.1016/j.patrec.2019.11.021.

[73] B. Shi, X. Bai, and S. Belongie, “Detecting Oriented Text in Natural Images by Linking Segments,” 2017.

[74] Z. Huang, Z. Zhong, L. Sun, and Q. Huo, “Mask R-CNN with pyramid attention network

for scene text detection," in Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Mar. 2019, pp. 764–772, doi: 10.1109/WACV.2019.00086.

[75] S. X. Zhang, X. Zhu, J. B. Hou, C. Yang, and X. C. Yin, "Kernel Proposal Network for Arbitrary Shape Text Detection," IEEE Trans. Neural Networks Learn. Syst., 2022, doi: 10.1109/TNNLS.2022.3152596.

[76] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation," 2018.

[77] W. Wang et al., "Shape Robust Text Detection with Progressive Scale Expansion Network," 2019.

[78] P. Yang et al., "Instance Segmentation Network with Self-Distillation for Scene Text Detection," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2978225.

[79] Z. Zhong, L. Sun, and Q. Huo, "A teacher-student learning based born-again training approach to improving scene text detection accuracy," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Sep. 2019, pp. 281–286, doi: 10.1109/ICDAR.2019.00053.

[80] C. Zhang et al., "Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes," 2019.

[81] A. Yildiz, H. Zan, and S. Said, "Classification and analysis of epileptic EEG recordings using convolutional neural network and class activation mapping," Biomed. Signal Process. Control, vol. 68, p. 102720, Jul. 2021, doi: 10.1016/J.BSPC.2021.102720.

[82] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization." pp. 618–626, 2017, Accessed: Oct. 29, 2021. [Online]. Available: http://gradcam.cloudcv.org.