

A novel ensemble deep network framework for scene text recognition

Sunil Kumar Dasari, Shilpa Mehta, Diana Steffi

Department of Electronics and Communication Engineering, School of Engineering, Presidency University, Bangalore, India

Article Info

Article history:

Received Mar 3, 2023

Revised Sep 23, 2023

Accepted Sep 29, 2023

Keywords:

Autoencoder

Customized CNN

EDN-proposed system

Ensemble deep network

Scene text recognition

ABSTRACT

In recent years, scene text recognition (STR) has always been considered a sequence-to-sequence problem. Attention-based techniques have a greater potential for context-semantic modelling, but they tend to overfit inadequate training data. STR is one of the most important and difficult challenges in image-based sequence recognition. A novel framework ensemble deep network (EDN) is proposed, EDN comprises customized convolutional neural network (CNN), and deep autoencoder. Customized CNN is designed by introducing the optimal spatial transformation module for optimizing the input of irregular text to read for same size. Further, deep autoencoder is introduced with effective attention mechanism utilizing the inherent features. The proposed ensemble deep network-proposed system (EDN-PS) approach outperforms the existing state-of-art techniques for both irregular and regular scene-texts and upon further simulations, the proposed model generates better results for IIT5K, ICDAR-13, ICDAR-15, and CUTE dataset in comparison with the existing system hence our proposed EDN-PS model outperforms the existing state-of-art methods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sunil Kumar Dasari

Department of Electronics and Communication Engineering, School of Engineering, Presidency University
Bangalore, Karnataka, India

Email: sunilkumar.d@presidencyuniversity.in

1. INTRODUCTION

The text contains many details about concepts used in various applications of artificial intelligence, including image storage, navigation, and translation. Recognizing text in natural situations has been important in understanding the vision system because text is widely used in communication and is versatile. Problems arise due to the information of the text, including the readability of the text. Reading begins with text searching by finding text in images and then using text recognition to convert those events into a readable word. On-the-spot reading has many applications in daily life, including machine translation to overcome the challenges of language restrictions and allow text to be read and translated instantly. Using text-to-speech, visual aids can be used to help visually impaired people read instructions on signs, automatic teller machine (ATM) machines or books. Many applications include multimedia retrieval, object identification, and intelligent analysis [1].

The existing text on the site is classified as an optical character recognition (OCR) problem after text detection and segmentation. Studies on literature have yielded monotonically positive results in some areas. While different fonts are often used to represent text, the background shows researchers a way to unravel the complexity of writing. Image distortion is very difficult to achieve and is affected by the limitations of the image medium, including blur, image brightness, and orientation. These features always reveal unique characteristics of the image. Another potential problem is identifying relevant features from a region of the

text. In addition, the change in text size occurring in the image scene causes the next difficulty to be color difficulty, orientation, and page layout difficulty [2]–[5]. When looking at photographs in nature, the effects that affect recognition are the structure and size of the text, changes in light, color, texture, and direction. The above challenges are not language-specific; However, to solve these problems, speech forms need to be placed in the region. Therefore, considering the complexity of the text in the database, different methods are proposed for local text, text extraction, and recognition, as described in [6]–[8].

Scene text recognition (STR) has always been problematic in many ways. This is based on problems that occur in almost all computer vision tasks, including image noise, background complexity, brightness differences, and visibility. Text in each language comes in different shapes, styles, and font sizes. In addition, natural text shows many things that affect performance, such as atypical expressions. However, it can also be seen outside the plane and in plane curvature with its change appearance. All these situations require us to focus on recognizing natural scripts. A deep learning based STR framework generally consists of four main steps. It starts with pre-processing as the first step and several models use transformations as well as image editing for easy recognition. The next step is feature extraction using convolutional neural network (CNN) to extract features from the image. The extracted results are then processed according to a sequential process and the last step after that is word prediction. reported research on the state-of-the-art in STR [3]. Recently, many strategies for STR have been proposed in [1], [2], [4]–[8], especially for written languages including Latin and English.

Major changes in STR methods have been successfully described in the literature in recent years, taking advantage of advances in deep learning. Recognizing text from images using traditional techniques is done by finding individual characters and then using CNN to recognize each character after cropping [4]. However, the effectiveness of network recognition is reduced due to the large number of different attributes. This process is based on natural defense mechanisms. Recurrent neural network (RNN) is used to solve the recommendation sequence to sequence problem. But the RNN model found that illicit text recognition presented a more challenging task. STR has developed rapidly in recent years. But there are still two major challenges to overcome: i) there are flaws in the representation of many ideas and ii) the large size of the text makes it difficult for the network to learn from patterns. This leads to adjustments to word images that will reduce clutter and make word recognition easier. Still or static images improve the accuracy of text, especially in documents where text is often inaccurate.

- Here we propose a new method called integrated ensemble deep networks-proposed system (EDN-PS), which essentially uses multiple neural networks such as CNN and RNN to constrain non-consistent features, providing more accurate readings.
- Create a custom CNN using the spatial translation module that aims to treat false text as normal text.
- A deep autoencoder architecture is integrated with the transformation to ensure accurate rendering of text; the autoencoder works in bidirectional operation.
- Measure deep correlations using accuracy as a metric to measure regular and irregular data.

This study is organized as follows: the first part of the study starts from the background of the text and the problems related to it. The second part of the research focuses on various deep learning techniques for text recognition. In the third part of this study, the deep integration network and its mathematical model and design are introduced. Finally, fifth part evaluates the ensemble deep network (EDN) decision variables to demonstrate the effectiveness of the proposed model.

2. RELATED WORK

Recently, STR models have been divided into two groups: language-based methods and language-based methods. The second is just images, patterns, and colors. Determines visual aspects such as the appearance of text, including the former, in turn, also learn speech rules that refer to the promise of speech in a way that forms meaningful patterns at the level of behavior, rather than retaining the verb as a word. Traditional methods for recognizing text in natural images rely on recognition of whole words or individual symbols [7], [8] describes the process of using the floating window symbol, shows the link [9], [10] describes the tree model, and width-to-line conversions [11] are used for special characters. Character classifiers are combined with various descriptors, including CNN [8], histogram of gradient [7], and random fern [12], intermediate features combined with random forest multiple forms please. The unique characters identified are classified into a word in combination using some fixed dictionary. According to Jaderberg *et al.* [8], CNN are used to generate text and non-text, character-sensitive and case-insensitive characters, and explicit binary maps to detect and reconstruct scene inherent text.

Recently text recognition problems are seen from series to series. According to Shi *et al.* [13], feature coding integration, text transcription method, and modeling sequences are integrated into an end-to-end framework that learns to recognize text from scene images. According to Zhang *et al.* [14], text

recognition is done through fragment-to-fragment matching based on listening to natural text images. The audio-based encoder automatically focuses on the most relevant parts of the text. According to Lei *et al.* [15], query recognition is considered as a convolutional problem, which is a combination of CNN and RNN. Multiple variants were extracted from the sequence-tagged architecture and analyzed. According to Sheng *et al.* [16], encoder and decoder models with sequence-to-sequence self-tracking stacking were used. Additionally, a transformation method can be used that can transform 2D images of natural phenomena into 1D features.

According to Ahmed *et al.* [17], a network architecture model called "transformer" for machine translation, which does not include convolution and recursion and is based on a tracking mechanism, is introduced. By calculating the position pair through a neural network, the individual behavior of the module can be linked to a specific function of the system, thus achieving a more and less parallel network connection. According to Dong *et al.* [18], a transformer architecture was adopted to solve the speech recognition problem. Similarly, Yu *et al.* [19], a network was designed to compensate for reading difficulties by combining convolutional techniques into a self-monitoring system. The transformer's frame is inspired by the combination model. According to Dehghani *et al.* [20], the transformer architecture is based on a design called "universal transformer" to process strings and other translations based on the length of the string found at runtime. Considering the transformer model in [21], a piece-to-piece acyclic STR framework with self-monitoring as an important part of the encoder and decoder architecture is designed for better understanding. According to Yang *et al.* [22], a stronger and easier to implement STR network based on holistic representation is proposed. According to Mu *et al.* [23], random blur unit (RBU) was proposed that divides the fuzzy function into different classes. The pixels of a unit have similar characteristics. Similarly, RBU provides additional reads to train the model and deliver more effective models. According to Zou *et al.* [24], the Bi-Long-term memory algorithm was found and implemented from the heat map and information about the behavior of the plates.

3. PROPOSED METHOD

The plan has three stages; one is a customized CNN and the other is an autoencoder EDN architecture is shown in Figure 1. Figure 1 shows the full operation and design of deep integration; it consists of three parts. The first part uses CNN to enhance the normal text to be like the normal input type and extract some important text. The second and third involve the encoding and decoding process of deep autoencoders. Additionally, each component and its mathematical model are discussed in the subsections.

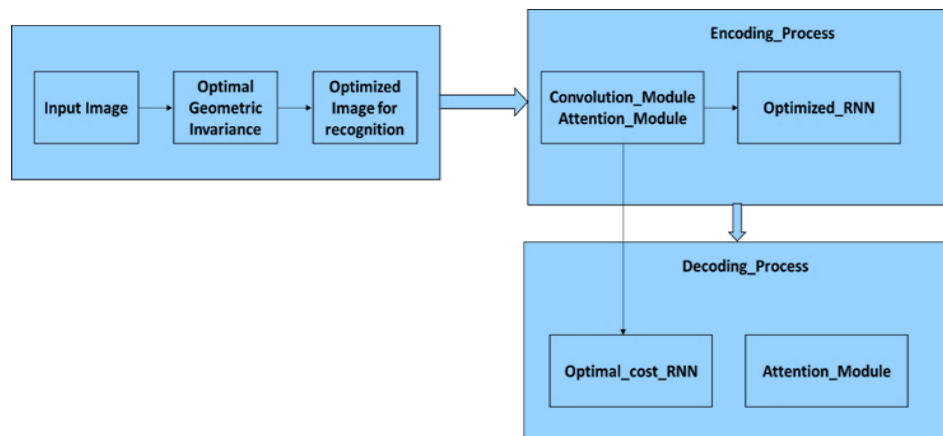


Figure 1. EDN

3.1. Customized convolutional neural network

The CNN rule here is a lightweight network that transforms input into larger-sized, more readable images. A two-stage autoencoder has two stages; uses image editing and puts the output into two representations. Finally, the output is decoded by the autoencoder. This model has been proven to be very good by simply detecting the text image gradient and then from this stage which allows the neural network to learn the transformer network is obtained and one has three levels to find the right one, network, generators, and standards. The process controls the content and output using a sampling engine. The original image is converted into an edited image and captured by the sampler.

3.1.1. Localized network

A transformation is evaluated across two sets of control points with similar size l , allocated by G^a and G^q . However $G^a = [G_1^a, \dots, G_l^a] \in A^{2 \times l}$ is the source of the control point, here $G_i^a = [s_i^a, n_i^a]^a$ is the i -th point. G^q is the target point which is the same as G^a , the G^q is placed at the top and bottom border of the output image at a specific position. Henceforth we obtain the G^a allotted to the localized network. This network is responsible to process the input image for the convolutional and pooling layer. This descends down G^a through a connected layer whose dimension is of size $2l$. This is noted down based on the neural network model, which is not trained by the back-propagation model.

3.1.2. Location mapping

Here G^a and G^q , results in the generation of the sample grid which maps each location to the rectified image through the input image. The transformation is shown in (1):

$$G^a = x + l^q \gamma(G^q) + y^q G^q \quad (1)$$

here $x \in A^{2 \times 1}$, $y \in A^{2 \times 2}$, $l^q \in A^{l \times 2}$ and γ is a function calculated by (2) and (3):

$$\gamma(G^q) = (\varpi(G^q - g_1^q), \varpi(G^q - g_2^q), \dots, \varpi(G^q - g_l^q))^q \quad (2)$$

$$\varpi(s) = \|s\|^2 \log(|s|) \quad (3)$$

a linear model is formed with specific boundary conditions.

$$\begin{aligned} l &= 0 \\ G_s^q l_1 &= 0 \\ G_t^q l_2 &= 0 \end{aligned} \quad (4)$$

The parameters x , y , w are evaluated, l_1 and l_2 is the first and second column l . G_s^q and G_t^q through the co-ordinates and t for G^q . The (1) is rewritten as (5):

$$\begin{bmatrix} l \\ x \\ y \end{bmatrix} = \begin{pmatrix} G^v \\ 0 \\ 0 \end{pmatrix} \begin{bmatrix} V & 1^{1 \times 1} & G^q \\ 1^{l \times 1} & 0 & 0 \\ (G^q)^q & 0 & 0 \end{bmatrix} \quad (5)$$

in the above equation $V \in A^{l \times l}$ is the matrix and $V_{u,v} = \varpi(|g_u^q - g_v^q|)$.

3.1.3. Interpolation of an input image

This example ensures that the pixel values of the edited image are imported to reflect the input image. The position is fixed outside the input image; Its value is cropped to save the image. A bilinear correlation is used to measure the pixel values of the image corrected for the newest pixels. Similar to the local network, the model is responsible for differentiation and allows CNN to leverage gradient-based algorithms.

3.2. Deep auto-encoder

Virtual feature extraction mechanism used in autoencoders. Scene text images are biased. The face mechanism created by the autoencoder suppresses the background and expands the foreground process accordingly. Constraints from the reception domain in the convolutional process. The integration network here is responsible for elaborating the regional context behind the visual extraction mechanism. Conventional labels indicate the disadvantages of RNN, both extract features from various retrieval domains. In the first paragraph, the visual feature map of the evaluation network feeds the output of the decoder. The second part is responsible for processing the specific maps determined by the best combination before being fed to the autoencoder.

3.2.1. Optimized attention residual

In the auto-encoder phase, adopt an attention-based residual block to extract visual features. The architecture of the residual block is shown below. An attention-based mechanism is implemented before integrating trunks and shortcuts. These channels are responsible for separate spatial attention across negligible parameter overhead. This is plugged through different residual blocks. The proposed approach

consists of two attention modules, channel attention and spatial attention. An intermediate feature map FM, evaluates a pooling operation determined as H_{avg}^g and the peak value is denoted as H_{max}^g , these are evaluated parallel to show-case global distinctive features. The channel attention mask $AM_g(FM) \in A^{I \times I \times G}$ is succeeded by the set of convolutional layers via descriptors. Consequently to AM_g , the FM is obtained by channel-wise max-pooling and average-pooling. They are connected and processed by a convolutional layer to allocate the mask $AM_x(FM) \in A^{L \times J \times 1}$.

$$AM_g(FM) = act_func(C(H_{avg}^g) + C(H_{max}^g)) \quad (6)$$

$$AM_x(FM) = act_func(h^{3 \times 3}([H_{avg}^v; H_{max}^v])) \quad (7)$$

Here FM allocates the broadcast model multiplied by AM_g to AM_x to generate an attentional feature map. This model here is capable of modelling the long-term dependencies for input sequence features. The output of this model at each timestamp depends on the present input and previous inputs. The feature map here works in a single direction and the ability of the single layer is constricted. Given an input $FM' \in A^{L' \times Z' \times G'}$, the size is varied accordingly to $(L'Z') \times U'$, here U' determines the hidden units. The auto-encoder-based text recognition specifically relies on two different types, which are the inherent text features in the text images for semantic dependency in-between, the characters. To evaluate these two aspects by taking into account the advantages to adopt these techniques by the auto-encoder incorporating the attention mechanism. This feature accommodates the visual FM from the attention-based ensemble network of the auto-encoder. This attention mechanism focuses on the semantic context features; this utilizes the output through the stacked auto-encoder. These losses are focused on (P_{At} and CP_{enc}) the weights are added for backpropagation in training. The total loss is determined by P_{tot} in (8):

$$P_{tot} = CP_{enc} + P_{At} \quad (8)$$

here, C is the hyper-parameter set to a proximal value. There exist many advantages as parallel training and parameter-free decoding. The scene-text recognition here is responsible for selecting the most probable character sequence. The dimension of the output is given by the class symbol represented by $(K + 1)$. The input sequence for feature s of length Q, the probability is aligned in the sequence of the output fetched. One sequence of the labelp is represented by various modifications, the probability distribution on p by summing up the probability over the possible alignment π . The probability $\xi(p|s)$ labelled as p on s is determined as follows. Here P_q is the probability of time t and $Y^{-1}(p)$ represents the sequence set that is mapped to p by Y. Upon assuming this the relevant labelling for the input sequence is allotted by (9)-(11):

$$\xi(\pi|s) = \cup_{q=1}^Q P_q, \forall \pi \in (K + 1)^Q \quad (9)$$

$$\xi(p|s) = \sum_{\pi \in Y^{-1}(p)} \xi(\pi|s) \quad (10)$$

$$z(s) = \underset{p \leq q}{argmax} prob(p|s) \quad (11)$$

The result is exponentially proportional to Q, for the decoding mechanism is predictive. This sequence-based prediction translates feature sequence to character sequence through varied lengths through a separate mechanism. This attention mechanism takes the visual features by taking into account to model out the dependencies and its ability to achieve effectiveness in capturing the output and dependency simultaneously. The attention mechanism uses the output at each step through the attention mechanism. This process is utilized for o steps to create a sequence $N = (n_1, n_2, \dots, n_a)$ for length A for the entire sequence. For the o - th step based on the output for the division $Z = (z_1, z_2, \dots, z_n)$, and n_a is predicted through the (12)-(16):

$$n_a = SAM(y_o + E_o v_a) \quad (12)$$

$$v_a = AEC(n_{a-1}, v_{a-1}, d_a) \quad (13)$$

$$d_a = \sum_{n=1}^Q z_n (\beta_{an}) \quad (14)$$

$$\beta_{an} = exp(f_{an}) (exp(\sum_{m=1}^Q f_{am}))^{-1} \quad (15)$$

$$f_{an} = x^o (HTF)(Xz_n + y + Ev_{a-1}) \quad (16)$$

here E_o , y_o , x , E , X and y are all learnable parameters and v_a is the hidden state in the decoder at step o . f_{an} is the alignment for the model that selects the inputs through this position a and the output of this position a and the output is given at position n . SAM indicates soft_arg_max function and HTF indicates hyperbolic tangent function.

4. PERFORMANCE EVALUATION

This section of the research focuses on evaluating the proposed framework considering the irregular and regular benchmark dataset. The data sets consisting images with multi lingual content in it with major irregularities. EDN-PS utilizes the deep learning libraries with system configuration of 4 GB CUDA enabled graphics packed with 16 GB of RAM on windows platform.

4.1. Dataset details

4.1.1. Irregular dataset

The IIIT5K dataset [9], which was produced with the use of inter-network data, contains the 3k clipped word 641-test pictures. The 643 words in each picture are broken down into 50 small words, 1,000 words in length. The remaining 644 words were from the dictionary, but some of them were randomly selected. The majority of the (ICDAR-13) [25] dataset image 658 samples come from IC03, which is its successor. Words with non-alphanumeric characters were 660 for a fair 659 comparison is taken out of the dataset. 661 1015-cropped word pictures without lexicons linked with 662 of them make up the filtered test dataset.

4.1.2. Regular dataset

The dataset ICDAR-15 [26] contains 6,545 cropped 664 text images, 2,077 testing photos, and 4,468 training images. It is not associated with any word. Almost all of the 666 within-word expressions. Images from the ICDAR-15 dataset 668 were captured using Google glasses without the requisite positioning or focus. The collection consists of high-resolution pictures taken in realistic locations. It comes with 288 cropped text pictures for testing purposes [27]. Because the majority of the word graphics are made up of randomly shaped letters, CUTE is the most challenging dataset to analyze. This dataset is unrelated to any lexicon.

4.2. Experimental analysis

In the above section, we evaluate our proposed model EDN-PS with two types of datasets irregular dataset and regular dataset. We have considered two datasets for an irregular dataset that is IIIT5K dataset and ICDAR-13 dataset, whereas the regular dataset we consider two datasets that is ICDAR-15 dataset and CUTE dataset and the results for each are shown below: furthermore, in order to prove the model efficiency. Comparison is carried out with several state of art technique such as ACNN [28], RARE [5], Char-Net [29], AON [30], FAN [31], EP [6], AGRU [32], MORAN [33], CA-FCN [34], MBAN [35], ESIR [36], ASTER [4], OCR [21], SCATTER [37], mask text spotter [38], EPAN [39], STAN [40], 2D-CNN [22], SRN [41], MASTER [22], PMMN [42], 2DPE_ES [43].

4.2.1. IIIT5K dataset

The IIIT5K dataset, which was produced with the use of inter-network data, contains the 3k clipped word 641-test pictures. We can see in Table 1 and Figure 2 result the proposed EDN-PS model is evaluated with the existing state-of-art techniques in the form of accuracy for irregular IIIT5K dataset and results are plotted in the form of a graph, in the end to conclude we can state that our proposed EDN-PS model outperforms the existing state-of-art techniques and generates an accuracy value of 98.3.

4.2.2. ICDAR-13 dataset

The majority of the ICDAR-13 dataset image 658 samples come from IC03, which is its successor. Words with non-alphanumeric characters, we can see in the below result our proposed EDN-PS model is evaluated with the existing state-of-art techniques in the form of accuracy, and results are plotted in the form of a graph for the irregular ICDAR-13 dataset. Table 2 and Figure 3 shows the accuracy comparison for ICDAR-13 dataset. In the end to conclude we can state that our proposed EDN-PS model outperforms the existing state-of-art techniques and generates an accuracy value of 98.42.

Table 1. Accuracy comparison for IIIT5K dataset

Dataset	Accuracy	Dataset	Accuracy
ACNN	81.8	ESIR	93.3
RARE	81.9	FAN	87.4
Char-Net	83.6	EP	88.3
AON	87	OCR	93.6
AGRU	89.5	SCATTER	93.7
MORAN	91.2	Mask text spotter	93.9
CA-FCN	91.9	ASTER	93.4
MBAN	93.2	STAN	94.1
2D-CNN	94.7	PMMN	96.2
SRN	94.8	2DPE_ES	97.7
MASTER	95	EDN-PS	98.3

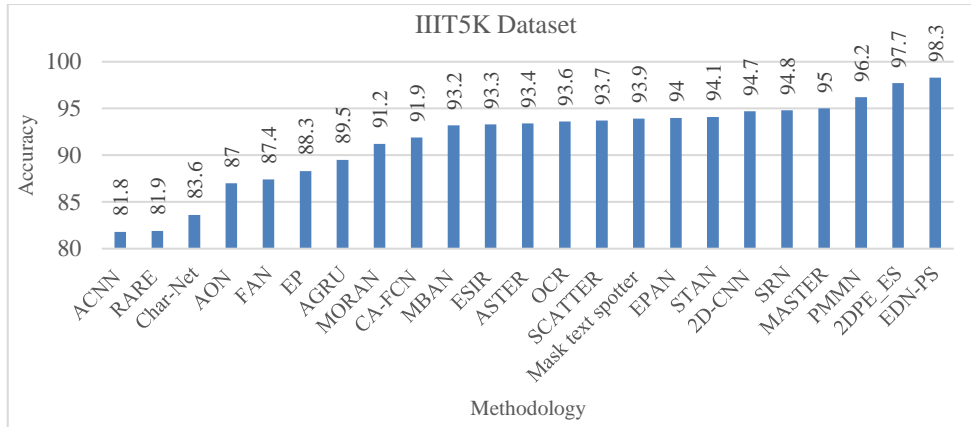


Figure 2 Accuracy comparison for IIIT5K dataset

Table 2. Accuracy comparison for ICDAR-13 dataset

Dataset	Accuracy	Dataset	Accuracy
ACNN	88	CA-FCN	91.5
RARE	88.6	ASTER	91.8
AGRU	90.1	MORAN	92.4
Char-Net	90.8	MBAN	92.8
ESIR	91.3	OCR	92.8
STAN	92.8	EPAN	94.5
2D-CNN	93.2	Mask text spotter	95.3
FAN	93.3	MASTER	95.3
SCATTEER	93.9	SRN	95.5
EP	94.4	PMMN	97.7
2DPE_ES	98	EDN-PS	98.42

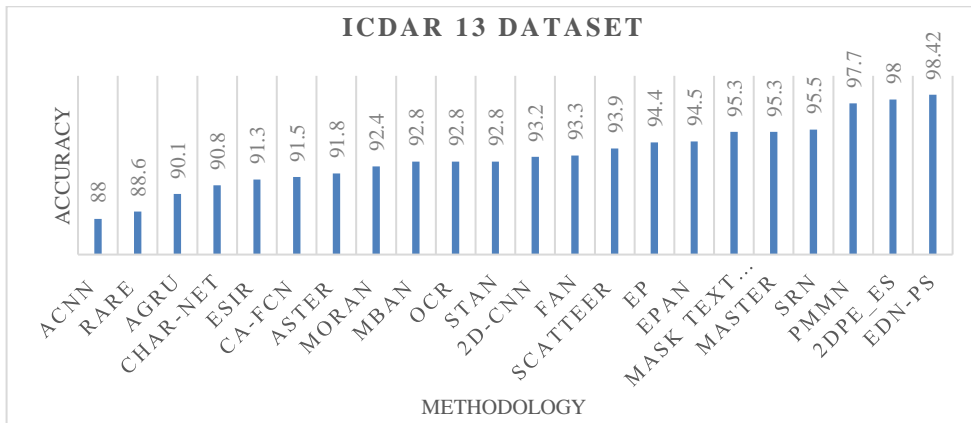


Figure 3. Accuracy comparison for ICDAR-13 dataset

4.2.3. ICDAR-15 dataset

The ICDAR-15 dataset contains 6,545-cropped 664 text images, 2,077 testing photos, and 4,468 training images. It is not associated with any word, we can see in the below result our proposed EDN-PS model is evaluated with the existing state-of-art techniques in the form of accuracy, and results are plotted in the form of a graph for the irregular ICDAR-15 dataset. Table 3 and Figure 4 shows the accuracy comparison of ICDAR-15 dataset. In the end to conclude we can state that our proposed EDN-PS model outperforms the existing state-of-art techniques and generates an accuracy value of 90.72.

4.2.4. CUTE dataset

The collection consists of high-resolution pictures taken in realistic locations. It comes with 288 cropped text pictures for testing purposes, we can see in the below result our proposed EDN-PS model is evaluated with the existing state-of-art techniques in the form of accuracy, and results are plotted in the form of a graph for irregular CUTE dataset. However, MORAN [27] approach gives the least performance in terms of accuracy of 77.4 and AON [32] approach gives an accuracy of 76.8. whereas the SCATTER [30] method generates an accuracy value of 87.5 and MASTER [37] method generates a value of 87.5, and the ES [41] gives a high-performance value of 91.3 and in the end to conclude we can state that our proposed EDN-PS model outperforms the existing state-of-art techniques and generates an accuracy value of 98.96. Table 4 and Figure 5 shows the accuracy comparison for CUTE dataset.

Table 3. Accuracy comparison for ICDAR-15 dataset

Dataset	Accuracy	Dataset	Accuracy
AON	68.2	ASTER	76.1
MORAN	68.8	MBAN	76.6
EP	73.9	AGRU	76.6
2D-CNN	74	STAN	76.7
ESIR	76.9	OCR	80
Mask text spotter	77.3	SCATTER	82.2
MASTE	79.4	SRN	82.7
EPAN	79.4	FAN	85.3
PMMN	85.5	EDN-PS	90.72
2DPE_ES	88.2		

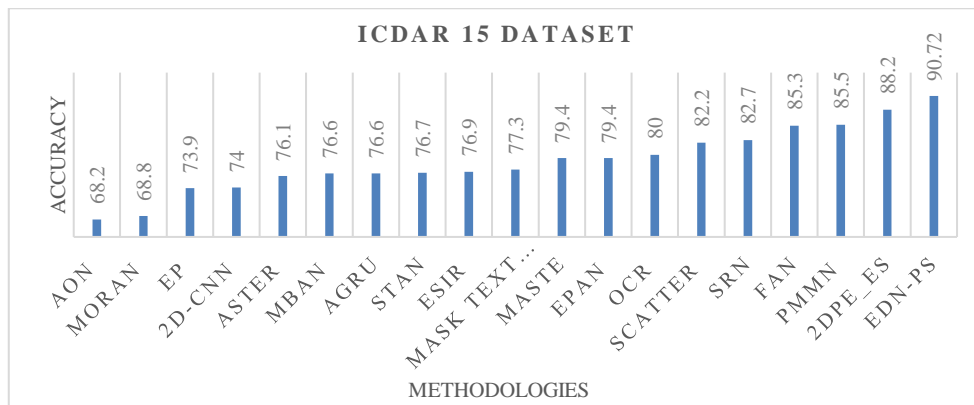


Figure 4. Accuracy comparison for ICDAR-15 dataset

Table 4. Accuracy comparison for CUTE dataset

Dataset	Accuracy	Dataset	Accuracy
RARE	59.2	AGRU	78
EPAN	73.9	ASTER	79.5
AON	76.8	CA-FCN	79.9
MORAN	77.4	MBAN	82.6
STAN	83.3	SCATTER	87.5
ESIR	83.3	MASTER	87.5
OCR	83.6	SRN	87.8
2D-CNN	85.4	Mask text spotter	87.8
2DPE-ES	91.3	EDN-PS	98.96
PMMN	91.9		

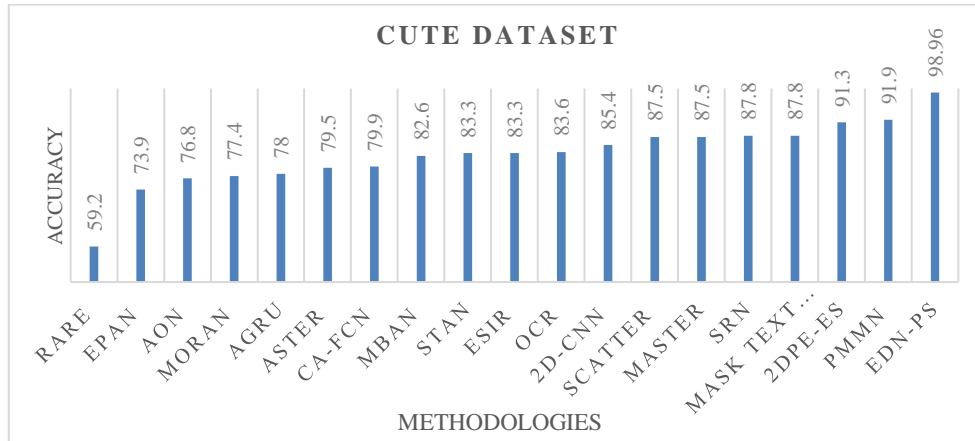


Figure 5. Accuracy comparison for CUTE dataset

4.3. Comparative analysis

Comparative analysis is made with existing and planned methods and the level of improvisation for erroneous and irregular data is calculated. While the accuracy value of the existing method for the IIIT5K dataset is 97.7, the accuracy value of the proposed EDN-PS is 98.3 and gives 0.612% improvisation. For the ICDAR-13 dataset, the actual value of the existing system is 98, while the actual value of the proposed EDN-PS is 98.42, giving a result of 0.42765% improvisation. The comparison analysis is shown in Table 5. For the ICDAR-15 dataset, the correct value of the existing method is 88.2, while the correct value of the proposed EDN-PS is 90.72, giving a result of 2.8169% improvisation. For the CUTE dataset, the accuracy value of the existing method is 91.3, while the accuracy value of the proposed EDN-PS is 98.96, resulting in an improvisation of 0.42765%. Finally, we can conclude that our design works better than the current one.

Table 5. Comparative analysis

Dataset	Existing system	Proposed system	Improvisation in (%)
IIIT5K	97.7	98.3	0.612
ICDAR-13	98	98.42	0.42765
ICDAR-15	88.2	90.72	2.8169
CUTE	91.3	98.96	8.05214

5. CONCLUSION

This research work develops an EDN for STR considering regular and irregular text; EDN comprises customized CNN for irregular text feature extraction and deep autoencoder for enhancement of accuracy in text recognition through optimizing the cost. In the first step, the arbitrary image is converted into a more reliable form henceforth the complexity to reduce the feature extraction process. In the second step, the feature extraction model extracts the feature representation accommodating the sequential transformation from a rectified image. In the third step, the auto-encoder performs feature extraction and transformation simultaneously. Our proposed approach EDN-PS is evaluated for IIIT5K dataset and an improvisation of 0.612% is achieved, whereas for the ICDAR-13 dataset, the improvisation is 0.42765%, for the ICDAR-15 dataset, the improvisation is 2.8169% and for CUTE dataset the improvisation over the existing system is 8.05214%, henceforth we can conclude that our proposed EDN-PS model outperforms the existing state-of-art techniques. Future work could focus on real time video-based STR.

REFERENCES





- [1] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: recent advances and future trends," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 19–36, 2016, doi: 10.1007/s11704-015-4488-0.
- [2] Q. Ye and D. Doermann, "Text detection and recognition in imagery: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1480–1500, 2015, doi: 10.1109/TPAMI.2014.2366765.
- [3] S. Long, X. He, and C. Yao, "Scene text detection and recognition: the deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021, doi: 10.1007/s11263-020-01369-0.
- [4] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: an attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2019, doi: 10.1109/TPAMI.2018.2848939.

- [5] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168–4176, doi: 10.1109/CVPR.2016.452.
- [6] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit probability for scene text recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1508–1516, doi: 10.1109/CVPR.2018.00163.
- [7] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1457–1464, doi: 10.1109/ICCV.2011.6126402.
- [8] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, pp. 512–528, doi: 10.1007/978-3-319-10593-2_34.
- [9] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," *BMVC 2012 - Electronic Proceedings of the British Machine Vision Conference 2012*, 2012, doi: 10.5244/C.26.127.
- [10] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2961–2968, doi: 10.1109/CVPR.2013.381.
- [11] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2963–2970, doi: 10.1109/CVPR.2010.5540041.
- [12] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: a learned multi-scale representation for scene text recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4042–4049, doi: 10.1109/CVPR.2014.515.
- [13] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017, doi: 10.1109/TPAMI.2016.2646371.
- [14] Y. Zhang, S. Nie, W. Liu, X. Xu, D. Zhang, and H. T. Shen, "Sequence-to-sequence domain adaptation network for robust text image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2735–2744, doi: 10.1109/CVPR.2019.00285.
- [15] Z. Lei, S. Zhao, H. Song, and J. Shen, "Scene text recognition using residual convolutional recurrent neural network," *Machine Vision and Applications*, vol. 29, no. 5, pp. 861–871, 2018, doi: 10.1007/s00138-018-0942-y.
- [16] F. Sheng, Z. Chen, and B. Xu, "NRTR: a no-recurrence sequence-to-sequence model for scene text recognition," *Proceedings of the International Conference on Document Analysis and Recognition*, 2019, pp. 781–786, doi: 10.1109/ICDAR.2019.00130.
- [17] S. Bin Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated Arabic scene character recognition," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, IEEE, Apr. 2017, pp. 46–51, doi: 10.1109/ASAR.2017.8067758.
- [18] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2018, pp. 5884–5888, doi: 10.1109/ICASSP.2018.8462506.
- [19] A. W. Yu *et al.*, "QaNet: combining local convolution with global self-attention for reading comprehension," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018, pp. 1–16.
- [20] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal transformers," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.
- [21] Y. Chen, H. Shu, W. Xu, Z. Yang, Z. Hong, and M. Dong, "Transformer text recognition with deep learning algorithm," *Computer Communications*, vol. 178, pp. 153–160, 2021, doi: 10.1016/j.comcom.2021.04.031.
- [22] L. Yang, P. Wang, H. Li, Z. Li, and Y. Zhang, "A holistic representation guided attention network for scene text recognition," *Neurocomputing*, vol. 414, pp. 67–75, 2020, doi: 10.1016/j.neucom.2020.07.010.
- [23] D. Mu, W. Sun, G. Xu, and W. Li, "Random blur data augmentation for scene text recognition," *IEEE Access*, vol. 9, pp. 136636–136646, 2021, doi: 10.1109/ACCESS.2021.3117035.
- [24] Y. Zou *et al.*, "A robust license plate recognition model based on Bi-LSTM," *IEEE Access*, vol. 8, pp. 211630–211641, 2020, doi: 10.1109/ACCESS.2020.3040238.
- [25] S. M. Lucas *et al.*, "ICDAR 2003 robust reading competitions: entries, results, and future directions," *International Journal on Document Analysis and Recognition*, vol. 7, no. 2–3, pp. 105–122, 2005, doi: 10.1007/s10032-004-0134-3.
- [26] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015, pp. 1156–1160, doi: 10.1109/ICDAR.2015.7333942.
- [27] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, Dec. 2014, doi: 10.1016/j.eswa.2014.07.008.
- [28] Y. Gao, Y. Chen, J. Wang, M. Tang, and H. Lu, "Reading scene text with fully convolutional sequence modeling," *Neurocomputing*, vol. 339, pp. 161–170, Apr. 2019, doi: 10.1016/j.neucom.2019.01.094.
- [29] W. Liu, C. Chen, and K. Y. K. Wong, "Char-Net: a character-aware neural network for distorted scene text recognition," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 7154–7161, 2018, doi: 10.1609/aaai.v32i1.12246.
- [30] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: towards arbitrarily-oriented text recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5571–5579, doi: 10.1109/CVPR.2018.00584.
- [31] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: towards accurate text recognition in natural images," *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5086–5094, doi: 10.1109/ICCV.2017.543.
- [32] Y. Huang, Z. Sun, L. Jin, and C. Luo, "EPAN: effective parts attention network for scene text recognition," *Neurocomputing*, vol. 376, pp. 202–213, 2020, doi: 10.1016/j.neucom.2019.10.010.
- [33] C. Luo, L. Jin, and Z. Sun, "MORAN: a multi-object rectified attention network for scene text recognition," *Pattern Recognition*, vol. 90, pp. 109–118, 2019, doi: 10.1016/j.patcog.2019.01.020.
- [34] N. Lu *et al.*, "MASTER: multi-aspect non-local network for scene text recognition," *Pattern Recognition*, vol. 117, pp. 1–10, Sep. 2021, doi: 10.1016/j.patcog.2021.107980.
- [35] C. Wang and C. L. Liu, "Multi-branch guided attention network for irregular text recognition," *Neurocomputing*, vol. 425, pp. 278–289, 2021, doi: 10.1016/j.neucom.2020.04.129.
- [36] F. Zhan and S. Lu, "ESIR: end-to-end scene text recognition via iterative image rectification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2054–2063, doi: 10.1109/CVPR.2019.00216.





- [37] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor, and R. Manmatha, "Scatter: selective context attentional scene text recognizer," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11959–11969, doi: 10.1109/CVPR42600.2020.01198.
- [38] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: an end-to-end trainable neural network for spotting text with arbitrary shapes," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, pp. 71–88, doi: 10.1007/978-3-030-01264-9_5.
- [39] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 8714–8721. doi: 10.1609/aaai.v33i01.33018714.
- [40] Q. Lin, C. Luo, L. Jin, and S. Lai, "STAN: a sequential transformation attention-based network for scene text recognition," *Pattern Recognition*, vol. 111, pp. 1–9, 2021, doi: 10.1016/j.patcog.2020.107692.
- [41] D. Yu *et al.*, "Towards accurate scene text recognition with semantic reasoning networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12110–12119, doi: 10.1109/CVPR42600.2020.01213.
- [42] Y. Zhang, Z. Fu, F. Huang, and Y. Liu, "PMMN: pre-trained multi-modal network for scene text recognition," *Pattern Recognition Letters*, vol. 151, pp. 103–111, 2021, doi: 10.1016/j.patrec.2021.07.016.
- [43] Y. Wu *et al.*, "Sequential alignment attention model for scene text recognition," *Journal of Visual Communication and Image Representation*, vol. 80, pp. 1–8, Oct. 2021, doi: 10.1016/j.jvcir.2021.103289.

BIOGRAPHIES OF AUTHORS







Sunil Kumar Dasari     have done B.E. in ECE from Andhra University, M.Tech. from JNTUH received gold medal for academics, pursuing Ph.D. from Presidency University and worked in Sreenidhi Institute of Science and Technology, GITAM University and also worked as assistant professor in Presidency University. Areas of interest: signal processing, image processing using AI, neural network, and machine learning algorithms. He can be contacted at email: sunilkumar.d@presidencyuniversity.in.



Dr. Shilpa Mehta     is a B.E. gold medalist and a professor with teaching experience of 30 years. She completed her bachelor of engineering in 1991 and masters in 1997. She completed her Ph.D. in ECE from JNTU Anantapur (Andhra Pradesh). He has been working in teaching field from 1992 onwards. She had published her first national conference paper in 1994 in national conference at IIT Roorkee, and first international paper at Nanyang University, Singapore in 1995. She has numerous journal and conference papers and has guided may award winning projects for undergraduate students. She can be contacted at email: shilpamehta@presidencyuniversity.in.



Diana Steffi     is a research scholar in Presidency University, Bangalore. Her research interests include image/signal processing, object detection, robotics, and path planning. She can be contacted at email: steffiseelan@gmail.com.