

Text detection and recognition through deep learning-based fusion neural network

Sunil Kumar Dasari, Shilpa Mehta

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

Article Info

Article history:

Received Apr 1, 2022

Revised Jun 17, 2022

Accepted Jan 30, 2023

Keywords:

Character recognition rate
Convolutional neural network
Fusion neural network
Recurrent neural network
Text recognition
Word recognition rate

ABSTRACT

Text recognition task involves recognizing the text from the natural image; it possesses various application, which aids information extraction through data mining from street view like images. Scene text recognition involves two stages i.e., text detection and text recognition, in the past several mechanisms has been proposed for accurate identification, these mechanisms are either traditional approach or deep learning-based. All the existing deep-learning methodology fails as this comprises character data and image data, further this research develops an optimal architecture fusion neural network (FNN) for text identification and recognition. FNN comprises several layers of convolutional neural network (CNN) as well as recurrent neural network (RNN). Within FNN architecture convolutional layer is utilized for the feature extraction and recurrent layer is utilized for attaining the feature classification prediction. Further, an optimal training architecture is established for the enhancement of classification accuracy. Here Devanagari MLT-19 dataset is utilized for the evaluation of FNN. Three different parameters are considered during evaluation i.e., script word identification, character recognition rate (CRR) and word recognition rate (WRR). Further comparison with existing models is performed to establish the proposed model efficiency and it shows FNN methodology observes 98.67% of script identification accuracy, 84.65% of WRR and 92.93% of CRR.

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, Presidency University

Bangalore, India

Email: sunilkumar@presidencyuniversity.in

1. INTRODUCTION

Over the years, digitalization has caused massive growth in image processing for every field. The advancement of technology results in huge demand for its application overall domains worldwide. There has been a massive transition toward technological advancements in every aspect of life. Before the occurrence of this change, information was normally stored in books, on pages, printed, or through newspapers [1]. The information that is stored in this form can be reprocessed by digitally storing it for further use and improvisations. In the field of medical science, clinical data and laboratory reports can be digitally stored with the help of image processing models such as text recognition [2]. Text recognition is one of the major parts while data extraction in image processing and it has several applications such as on traffic signals cameras are installed and they capture the image of registration plate of vehicles and further details of driver can be extracted by image text recognition technique [3]. One of the major applications comes in the health field that prescription given by doctors could be make readable through text recognition technique. Text extraction from medical reports can be easily implemented for diagnosis purpose, which improvise the quality of assessment [4].

Text recognition techniques are also utilized for several purpose such as tracking or for gathering information. From videos and text, we can extract several information through this process and it may contain the information, which we required. Therefore, it is necessary to perform the extraction and recognition of text. One of the major applications of text recognition is traffic cameras [5]. As applied on several application it can be said that deep learning architecture is applicable for text recognition. In terms of recognition as well as identification and prediction of text, several methodologies are implemented. Text recognition approach can be performed by utilizing of characters of text as well as by utilizing the characters sequence. Images frames can be regular or irregular, which are utilized for text recognition. When we refer to irregular text this means for those texts, which are not uniform in writing such as vertical way written text or might be it can be curved as well as horizontal. A text recognition architecture comprises character identification along with text prediction and recognition and all together clubbed into one single model. Nature of the model is bidirectional in nature so as the decoders which are generally applied also bidirectional. Proposed text recognition architectures are also utilized for training of proposed network. Text recognition techniques are implemented and tested through extensive experiments within fixed dataset [6], [7].

Here Figure 1 displays the variety of images to which the proposed model can be applied. Those images comprise several types of images such as (a) character images, (b) registration plate of vehicle, (c) written medical prescription, (d) printed medical reports as well as, (e) name boards of shops, and (f) people along with signboards. All useful information extracted from these images are processed for further process in model. All these processes are performed automatically. Machine learning models have been implemented for the text recognition and extraction process. Figure 1(a) to 1(f) show the various applications of text recognition such as extracting information of vehicle number plates through traffic cameras, the medical reports or prescriptions of patients can be stored and used for medical purposes, signboards that contain information are a few such examples where the text recognition models can be applied. Once the textual information has been converted digitally, editing and changes that have to be made to it are also easily performed. The application of text recognition is also applied to the food industry for checking ingredients, and the life cycle of the goods. This is also applied to drug industries as well [8]. Other than printed documents, other handwritten documents also have high informational value. The characters in these cursive written images must be also recognized correctly. Hence, recognition models have been implemented by machine learning methodologies [9].



Figure 1. Types of images that are used for text recognition (a) text image, (b) vehicle number plate, (c) written doctor's prescription, (d) printed medical report, (e) name board, and (f) signboard

Motivation and contribution of research work. Within every field, text identification techniques have huge demand. This is useful for automatic recognition, making work easier and for people who are visually challenged. Those images, which contains text, are captured and further processed for extracting the useful information. The traffic department in every major city uses traffic cameras to spot rules offered through their licence plate. One of the major applications of text recognition comes in medical science. This helps in storing the data of patients as well as helps in analysing reports for quality diagnosis. Apart from that, food industries are using this recognition technique to gather the information about product life cycle. Considering all

organisations, that run digitally require data that is also stored digitally. The conversion of information digitally can be performed by the use of text recognition models. Moreover, considering the above motivation, this research work has the following contribution,

- A fusion neural network (FNN) model is established which comprises convolutional layer as well as recurrent layers. Feature extraction is done with images and further converted into text sequences.
- Each character sequences are divided into several frames for efficient prediction and recognition of text. And those frames are used as input.
- Above all the layer deep translation layer comes at top and implemented for translating each frame those are extracted from the FNN model.
- Output frames are considered as recognized sequences and further FNN is computed by implementing Devanagari MLT-19. Evaluation is performed by considering three different parameters that are word recognition rate (WRR), as well as character recognition rate (CRR) and script word identification.

This research work comprises four sections and ends with a conclusion that states the research outcome. First section elaborates the existing text recognition architecture and it emphasises the character and text identification models along with their process of feature extraction in detail. This section states the motivation as well as contribution to be carried out in this research. The second section of this research elaborates about the existing models along with the technologies utilized and gap in their research. Third phase comprises on the development of architecture for extraction of features and FNN. Fourth and last section of this research shows the outcomes obtained from the research. At the end it comprises with conclusion of the research.

2. RELATED WORK

In the previous year's several research is performed for scene text recognition techniques. However, most of the researches are not that much concerned about the arbitrary-orientation scene in text semantics and this extensively exist within several real-world applications. These semantically arbitrary-orientation scene texts majorly obtained from two aspects. Paper [10] focuses on text recognition of a cursive script that is not in the Latin family. The Urdu language is used in the text recognition in the natural scenes. There is a three-step process that is proposed in this paper namely, detection, prediction and recognition of the images. Optical character recognition of a photo is used for the evaluation of its three main phases. The detection phase of this problem uses Resnet50 and recurrent neural network (RNN) model for prediction. In paper [11], a framework is proposed for text recognition in mobile systems where two various artificial neural network (ANN) networks are used along with dynamic programming. The training parameters used in this ANN model are of a small number to satisfy constrain difficulties since the model is fitted into an embedded system. The proposed work shows high accuracy used on natural datasets. The methodology proposed shows the highest accuracy in comparison to various long-short-term memory (LSTM) that have been used prior. In paper [12], multi scripted images are used for text recognition as well as detection. A methodology where an image is represented with bits of byte that are used for every pixel is termed plane bit slicing. The main aim of this is to convert an image into a binary image. The constraints that arise due to concave and convex to identify the plane are called the iterative symmetry of the nearest neighbor. The recognition process is based on the character size, which gives the relationship between bands of fused and high wavelet frequency. Paper [13] mainly focuses on pattern and text recognition for Arabic writing. This is considered a challenging task, as there are many constraints involved such as writing in cursive in nature, characters that are alike and limitless terminologies. Hence, a deep learning system is introduced that consists of a framework that is used for feature extraction on an auto-encoder and text recognition are performed with the help of a hidden Markov model. In paper [14], a framework of encoder and decoder is proposed where a convolutional neural network (CNN) model is used for the conversion of the image into a sequence of features. An LSTM model is further applied to the sequence of features to convert it into feature code. An integrated module is developed for decoding and producing an output. In the paper [15], the focus lies on the medical health records that are stored and accessed electronically. Entire health records are not always available due to problems that arise in the electronic health record (EHR) systems. Hence, a deep learning method is proposed for textual information to be extracted from medical reports. This uses text detection as well as text recognition that is based on the uses of a patch strategy. In paper [16], unlabeled data is used in comparison to labelled data since the use of unlabeled data is more easily available and inexpensive, easy for collection. This data is used for the text recognition that is performed with semi-supervised methodologies. The string quality that is generated is evaluated using reinforcement learning by comparing the original label and the string generated. In paper [17], emphasizes the image recognition process that is guided by text using neural networks. CNN model is used for feature concentration, and influence on semantic features is used to train the CNN model irrespective of the image having the presence of texts in it. Paper [18]–[20], focuses on Arabic English text recognition due to its extensive use in the gulf regions. An extremal region

stability technique (MSER) is proposed for the extraction of invariant features using regions that are detected by MSER. An adapted LSTM model is deployed to deal with the cursive form of Arabic writing. In papers [20]–[23], a CNN model is proposed for the extraction of features and an LSTM model is proposed for the process of sequence recognition as well as the decoding performed by a transcription layer. The combination of these two models is incorporated into forming a network for text recognition. This methodology that is later proposed is compared to various other algorithms and RNN based algorithms for performance-based comparison of the achieved datasets. Paper [24]–[27], proposes a combination of a CNN model as well as an RNN model for text detection. For text adaptation on various scales, a feature pyramid is implemented and is applied to a part of the CNN model after which an LSTM model is used for the encoding of these features for the generation of output texts. In paper [28]–[30], the main focus lies on image text recognition for the Arabic language. An Arabic language dataset has been introduced in this paper that sets a benchmark for public availability, a model that is trained on bilingual recognition and detection is proposed to handle text that is bilingual for which a bidirectional LSTM model is used.

3. PROPOSED METHODOLOGY

The use of a CNN for similar works has been applied recently but the CNN model cannot be directly applied to sequence text prediction and recognition due to a fixed dimension being present in the operation of the inputs and outputs. Hence, the recognition and prediction process cannot be applied to a varying sequence label length. Therefore, a methodology that is proposed in this paper uses a combination of two various types of neural networks namely, a CNN followed by a RNN. This combination of both these neural networks that are used in this methodology is collectively termed an FNN. FNN architecture comprises of deep-CNN model that is particularly utilized for the detection of patterns from the images as well as videos. In comparison to CNN network, we use the FNN model as per the proposed architecture and it can be trained by utilizing the label sequences as without any need for any kind of annotations. Here it can produce sequence of several labels. Proposed model achieves a better performance for text recognition and detection and the parameters which we utilize in the FNN architecture are few than those applied in the CNN model.

The text images that are used in the recognition process are split into frames out of which the feature extraction process is performed using the proposed network. Figure 2 consists of three distinct parts. The first two parts of this model consist of the FNN, which consists of a CNN and the RNN followed by a translation layer. The last part is the translation of the frames into labels.

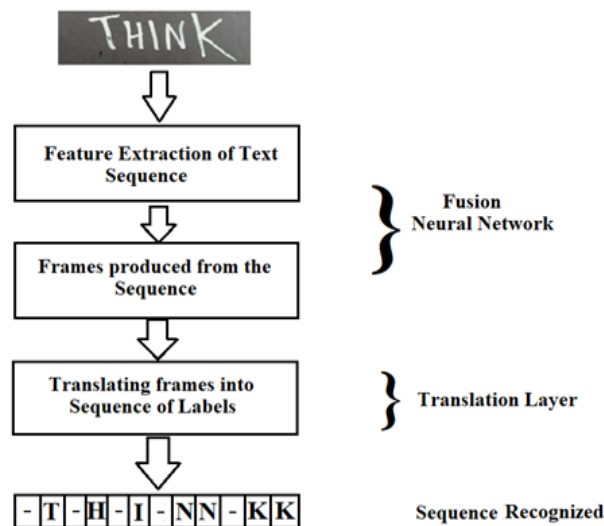


Figure 2. Proposed work architecture

The first step is to upload the input image that comprises of such images where text sequences should be recognized. Further, FNN is applied for the feature extraction process of text sequences that has been retrieved through input images. Each text sequences are splitted into several frames for effective text recognition as well as text prediction process. All the sequences, which are given as output after sequence formation that is applied for prediction process. After which the translation layer at the top is used for translating each of the frames that are retrieved from the FNN into a sequence of labels. The output of this is

produced as a recognized sequence. This entire framework is proposed in this paper for the efficient recognition of text from images.

3.1. Feature extraction of text sequence

A basic CNN layer is made up of four distinct types of layers, namely the convolutional layer, the pooling layer followed by the rectified liner activation unit (ReLU) layer for correction and lastly the fully connected layer. This model is constructed by using the convolutional layer and the max-pooling layer of the CNN architecture model. These two layers of the FNN are used in the process of feature extraction. The input images have to be properly resized to a set constrain value before being fed as input for the network. The layers that are mentioned extract the sequence as features from the input image that has been deployed. After which vectors are used in the network where a vector of n-dimension is given various features that represent different objects. These feature vectors are extracted from the sequences by the convolutional layer. This is used as the input for the next part of the FNN, which includes the RNN. Sequence features are generated from the left side to the right side and sorted in columns. Every single pixel value is the fixed value for the width of the column.

The sequences that are produced are translation invariant; this means even if the inputs are translated the network still produced the same output. In other words, the system will produce the same response even if the input sequence is shifted. Each column is in its respective feature sequence from left to right. Every feature vector is associated with the sequence that is derived from the original image. This process constitutes the first phase of the FNN, which is the CNN part of FNN this cannot completely be used for the problem that has been proposed in this paper because of the definite dimension that has to be used for CNN models. Since the implementations are performed based, on the text sequences, this is not efficient and hence a RNN is needed to be used with the existing network, which is termed an FNN.

3.2. Labelling sequence text

The FNN uses the output from the previous part of the process as input for this phase. This phase uses RNNs since they do not require prior knowledge of the input data that is being deployed, other than the choice of the input and output being represented. This network is trained based on time series, as the general mechanism that is deployed for this phase is powerful. This phase of the network is used to provide a sequence label for the already segmented data. The labelling classifications that are applied are independent. The FNN at this phase is used for label prediction and is placed above the CNN layers. The RNN is strongly used when contextual information is involved with sequencing. The use of contextual indications for images relating to text format during feature extraction for this model makes it more stable and efficient. The sequence recognition based on textual images is given more accuracy by using a recurrent network in the FNN. It is easier for the recognition process to occur in sequence than considering individual symbols. The characters that are used in text vary in width and some characters may be wider than others may. Therefore, the wider characters occupy more frames than the less wide ones. Some characters can also be recognized by their appearance such as their height, for example, the height difference between “rl” is recognizable. Hence, the recognition process is easier and more accurate when performed in sequence as compared to being performed individually. FNN also allows training of the model, which is done by the convolutional layer. An error that is calculated is back propagated to the input, training the model for better accuracy. The fusion network is collectively trained by the convolutional layer, which also includes training of the recurrent layer.

3.3. FNN-optimal training

This network also sanctions the training of model those who have varying length in text sequencing. No any specific constraints are applied for the length sequence. This provide the network through a wide character range and this is applied for training purpose. The FNN predicts distribution of the labels l_t for each frame f_t . In (1) calculate the extracted features.

$$S = f_1, \dots, \dots, f_T \quad (1)$$

The FNN architecture comprises of hidden layer between the first phase output and second phase output. Every time the frame f_t is received in a sequence from FNN. The internal state I_t along with the function, which comprises both the frames, that is considered as present input f_t as well as the I_{t-1} that is the previous state. The internal state given as below mentioned in (2).

$$I_t = \partial(f_t, I_{t-1}) \quad (2)$$

This means that the prediction of the label l_t is done based on the internal state I_t . The previous texts are also used to train the current network. The general machine learning algorithms that use backpropagation

normally encounter the vanishing gradient problem, which has been taken care of by the recurrent layer of the FNN. To address this problem in detail the recurrent layers that are being used are associated with three gates, namely, forget gate, input gate and output gate along with a memory cell. When a network that uses the output of the first phase containing the convolutional layers as an input to the other phase of the network requires storing the data. Since there are no constraints on the variable length of the sequence that is being used the characters need to be stored. The FNN stores the past sequence in the memory cell, whereas the input gate and the output gate allow sequences to be stored over a long time. The stored memory can be erased from time to time by using the forget gate. The proposed FNN uses both forward and backward propagation since the sequence from not just past contents but is to be used in both directions. This makes the network that is proposed more accurate and efficient, especially for image related sequences. Hence, the recurrent layer that is deployed in this network uses propagation in both directions. For a better more intense and accurately performing network multi back and front propagating models can be used making it a deep network. These models normally result in a higher performance rate as well as significant improvements, especially in relation to speech and text recognition models.

3.4. Deep error differentiation

The error differentiation that is calculated in the FNN is produced in the backward direction, which is also termed back propagation in relation to time. At the lower end of the FNN, the calculated error trains the model and converts the frames into sequence features. There is a customized layer in the FNN that is used to merge the two phases of the proposed network where the feature sequence that is produced is mapped for the formation of labels. The error rate is calculated to perform propagation in both directions; the updating of the network for better efficiency is performed by calculating the error for the label sequence and updating the training model.

$$\text{Error}(g, R') = \frac{1}{Y} \sum_{(w,Y) \in R'} \text{CD}(g(w)) \quad (3)$$

In the (3), the error rate of the labels is calculated using the equation. The total count of the final labels is given as Y in R'. Whereas the test set of sequence that is given for the error to be calculated is denoted as R'. The denotation CD(o, p) is the distance between two sequences, namely o and p.

3.5. Optimal labeling

The annotation g is used to represent the classifier state for training the model. For a sequence that is given as W and the length of the sequence is given as S, the FNN has i inputs and j outputs and the feature vector is given as v. Then the mapping of the sequence into labels is performed.

$$M_S: (\mathbb{N}^i)^S \rightarrow (\mathbb{N}^j)^S \quad (4)$$

Let f be the sequence of the output that results from the FNN. Hence, the value of f is given as,

$$f = M_S(W) \quad (5)$$

$$\rho(w|\pi) = \prod_{s=1}^S f_{\pi_s}^s \quad \text{where } \pi \text{ belongs to } L^S \quad (6)$$

the (6) is used to calculate the output of the network that is used for improving the values of labels, this is denoted as ρ . The labels that are used in this equation are denoted as L, whereas L^S is the label that is produced for the particular sequence of length denoted as S. The model is trained according to the FNN that is proposed. The calculation and updating of sequence values for labels that are given by error rate are essential for proper working and higher accuracy of the FNN. In the next part of the developed model, the translation layer at the top is used for translating each of the frames that are retrieved from the FNN into a sequence of labels. The output of this is produced as a recognized sequence.

3.6. Deep translation layer

In machine learning, retrieval of information is performed using sample sequence data of the system, this is termed lexicon. Another form of information retrieval is performed without the use of lexicons. The basic example of lexicon use is a spelling check in a dictionary. A non-lexicon model of information retrieval is performed by considering the maximum probability of the sequence. The proposed architecture utilizes a non-lexicon model for transcription. In this phase of research, all those frames, which are produced through the FNN, are translated and we get the outcome in sequence recognition. Her all the sequences features are developed from left to right side along with in sorted columns. Each single pixel value will be fixed value

regarding the width of column that is produced by the same response either the input sequence is even shifted. Each column appears in the feature sequence in left to right manner. Each feature vector that is related with those sequences that derived from the original image, after which the sequence is recognized.

4. PERFORMANCE EVALUATION

A FNN model is applied for the text recognition approach and it comprises of two diverse deep learning layers in combination i.e., recurrent layer as well as convolutional layer. Globally within major cities, they extensively perform multi-lingual uses of texts as several scripts needs to be read as well as recognised. Hence, the FNN architecture, which is proposed in this research, is evaluated on multi lingual text (MLT) comprises of scene text images. There are several phases of FNN evaluation that includes, i) detection of text, ii) cropped classification of word script, and iii) recognition of text. FNN is designed on visual studio IDE using python as a programming language on Windows 10 platform. System configuration includes 8 GB of RAM packed with 2 GB of NVidia CUDA enabled graphics card. This section evaluates the proposed FNN, evaluation is carried out considering different parameters such as classification accuracy, WRR, CRR; also, script cropping and identification is presented.

4.1. Dataset details

The dataset that is used in this proposed research is robust reading challenge–multi-lingual scene text-2019 (RRC-MLT-2019), which consists of multi-lingual scene images that consist of various languages and scripts. This dataset helps for text detection and recognition of script and contains added scene text images as compared to other text image datasets. The RRC-MLT-2019 dataset consists of 20,000 scene text images that have over 10 different language scripts. This text recognition model is trained to recognise text from images that are captured at various natural scenes such as signboards, name boards of shops, medical prescriptions, printed medical reports, which are in various script languages [22]. Below mentioned Figure 3 shows the text scene images of the dataset RRC-MLT-2019.



Figure 3. Text scene images of the dataset RRC-MLT-2019

The images in the RRC-MLT-2019 dataset undergo three different phases that result in the recognized output, which namely are,

- Detection of the text: The image dataset that consists of images has to be detected in comparison to images that do not consist of texts. The dataset that is used to train the data has 10,000 images.
- Classification of word script after cropping: in this phase of the model, the script that has been cropped from the image is identified. The dataset for training and testing of this phase has cropped images.
- Recognition of text: In this phase, each word of the image is combined and identified. The dataset used to train and test in this phase consist of 10,000 images each.

4.2. Script cropping

The text script has to be cropped from the scene text image after the presence of text in the image is detected. In this phase of the model, the script that has been cropped from the image is identified. The dataset for training and testing of this phase has cropped images. Table 1 shows the input image, cropped words and recognised words respectively in three columns.

Table 1. Script cropping and recognition

Original image	Cropped words	Recognition results
	   	सावधान गतिरोधक पुढे आहे
	    	ज्ञानदिप कॉलनी वाहने सावकाश चालवा
	   	डंख छोटा धोका मोठा
	  	ज्ञानराज फोटो स्टुडीओ
	     	वासुदेव बळवंत फडके मार्ग सेक्टर word_84895.png, क्र-१

4.2.1. Script word identification

Based on classification accuracy we evaluate script word identification. Here H is considered as the ground truth class as well as U is considered as predicted class. In (7), it shows the computation of classification accuracy.

$$\text{accuracy} = \frac{1}{n} \sum_{j=1 \text{ to } n} \begin{cases} 1 & \text{if } h_j = u_j \\ 0 & \text{else} \end{cases} \tag{7}$$

As per the (7) mentioned, h_j and u_j indicates towards the image. Table 2 displays the script word identification comparison. Existing methodology attains 94.02 as classification accuracy as the existing model applies CNN with self-attention along with RNN. However, the proposed fusion technique observes 98.67% of accuracy.

Table 2. Script word identification comparison

Methodologies	Accuracy (in percentage)
ResNet_TPS	90.90%
SCUT_DLVC	90.96%
GSPA_HUST	91.01%
CNN based classifier	91.65%
DPPR	94.02%
Fusion neural network (FNN)	98.67%

4.3. Word recognition rate

WRR is one of the comparison parameters for model evaluation. It is computed as the number of the correctly identified word in an identified image. Figure 4 shows the comparison of existing STAR-Net with 64.55% and proposed FNN with 84.65% of WRR.

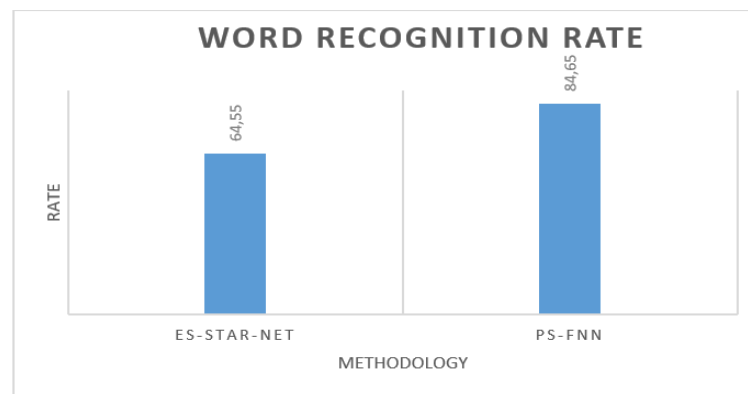


Figure 4. Word recognition rate comparison

4.4. Character recognition rate

CRR is yet another comparison parameter for model evaluation. The stands for total number of correctly recognized characters in the text. Figure 5 shows the comparison of existing methodologies STAR-Net with 85.587% of CRR and proposed fusion neural network (PS-FNN) with 92.93% of CRR.



Figure 5. Character recognition rate comparison

4.5. Comparative analysis and discussion

This section discusses the improvisation of the fusion approach over the existing methodologies. At first considering the script word identification, the proposed fusion approach observes 4.65% of improvisation

over the existing approach. Similarly, considering the WRR, the proposed fusion approach observes 20.10% of improvisation over the existing model. At last, considering CRR, the proposed fusion approach observes 7.06% of improvisation over the existing model.

5. CONCLUSION

In our day-to-day life, text plays an important role within several applications such as root navigation as well as automatic translation and as an assistance for people who are visually impaired. This research proposes an optimized FNN that's consists of different layers of diverse neural networks. Here two different layers are fused so that two different set of data can be exploit. All layers have to perform own task such as first convolutional layer is utilized for extraction of features then next recurrent layer is applied for attaining feature sequence prediction. Further, optimal training architecture is established for the enhancement of classification accuracy. Here we utilize dataset RRC-MLT-2019, which is available publicly for FNN evaluation. Several parameters are considered while evaluating the classification accuracy that are WRR as well as CRR. In addition, comparative analysis of the same parameter is carried out with the existing methodology and marginal improvisation is observed by FNN. FNN possesses a great advantage by using the two distinctive neural networks for text recognition, however, this research work is limited to well-mannered images and any distortion in the image due to natural causes makes it harder to recognize, also dataset remains one of the major issues. Thus, future scope relies on enhancing further recognition.





REFERENCES

- [1] 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.
- [2] W. Xue, Q. Li, and Q. Xue, "Text detection and recognition for images of medical laboratory reports With a deep learning approach," *IEEE Access*, vol. 8, pp. 407–416, 2020, doi: 10.1109/ACCESS.2019.2961964.
- [3] Y. Cao, S. Ma, and H. Pan, "FDTA: fully convolutional scene text detection with text attention," *IEEE Access*, vol. 8, pp. 155441–155449, 2020, doi: 10.1109/ACCESS.2020.3018784.
- [4] H. Liu *et al.*, "A natural language processing pipeline of chinese free-text radiology reports for liver cancer diagnosis," *IEEE Access*, vol. 8, pp. 159110–159119, 2020, doi: 10.1109/ACCESS.2020.3020138.
- [5] S. Tian, X.-C. Yin, Y. Su, and H.-W. Hao, "A unified framework for tracking based text detection and recognition from web videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 542–554, Mar. 2018, doi: 10.1109/TPAMI.2017.2692763.
- [6] 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, Sep. 2019, doi: 10.1109/TPAMI.2018.2848939.
- [7] X.-Y. Zhang, C.-L. Liu, and C. Y. Suen, "Towards robust pattern recognition: a review," *Proceedings of the IEEE*, vol. 108, no. 6, pp. 894–922, Jun. 2020, doi: 10.1109/JPROC.2020.2989782.
- [8] P. Sahare and S. B. Dhok, "Multilingual character segmentation and recognition schemes for indian document images," *IEEE Access*, vol. 6, pp. 10603–10617, 2018, doi: 10.1109/ACCESS.2018.2795104.
- [9] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542.
- [10] S. Y. Arafat and M. J. Iqbal, "Urdu-text detection and recognition in natural scene images using deep learning," *IEEE Access*, vol. 8, pp. 96787–96803, 2020, doi: 10.1109/ACCESS.2020.2994214.
- [11] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Two-step CNN framework for text line recognition in camera-captured images," *IEEE Access*, vol. 8, pp. 32587–32600, 2020, doi: 10.1109/ACCESS.2020.2974051.
- [12] K. S. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal, and T. Lu, "Multi-script-oriented text detection and recognition in video/scene/born digital images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1145–1162, Apr. 2019, doi: 10.1109/TCSVT.2018.2817642.
- [13] N. Rahal, M. Tounsi, A. Hussain, and A. M. Alimi, "Deep sparse auto-encoder features learning for arabic text recognition," *IEEE Access*, vol. 9, pp. 18569–18584, 2021, doi: 10.1109/ACCESS.2021.3053618.
- [14] L.-Q. Zuo, H.-M. Sun, Q.-C. Mao, R. Qi, and R.-S. Jia, "Natural scene text recognition based on encoder-decoder framework," *IEEE Access*, vol. 7, pp. 62616–62623, 2019, doi: 10.1109/ACCESS.2019.2916616.
- [15] C. Yang, M. Chen, Z. Xiong, Y. Yuan, and Q. Wang, "CM-Net: concentric mask based arbitrary-shaped text detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2864–2877, 2022, doi: 10.1109/TIP.2022.3141844.
- [16] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Semi-supervised scene text recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 3005–3016, 2021, doi: 10.1109/TIP.2021.3051485.
- [17] Z. Zhang, P. Chen, X. Shi, and L. Yang, "Text-guided neural network training for image recognition in natural scenes and medicine," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1733–1745, May 2021, doi: 10.1109/TPAMI.2019.2955476.
- [18] S. Bin Ahmed, S. Naz, M. I. Razzak, and R. B. Yusof, "A novel dataset for english-arabic scene text recognition (EASTR)-42K and Its evaluation using invariant feature extraction on detected extremal regions," *IEEE Access*, vol. 7, pp. 19801–19820, 2019, doi: 10.1109/ACCESS.2019.2895876.
- [19] Q. Liang, S. Xiang, Y. Wang, W. Sun, and D. Zhang, "RNTR-Net: a robust natural text recognition network," *IEEE Access*, vol. 8, pp. 7719–7730, 2020, doi: 10.1109/ACCESS.2020.2964148.
- [20] J. Deng, X. Luo, J. Zheng, W. Dang, and W. Li, "Text enhancement network for cross-domain scene text detection," *IEEE Signal Processing Letters*, vol. 29, pp. 2203–2207, 2022, doi: 10.1109/LSP.2022.3214155.
- [21] H. Hassan, A. El-Mahdy, and M. E. Hussein, "Arabic scene text recognition in the deep learning era: analysis on a novel dataset," *IEEE Access*, vol. 9, pp. 107046–107058, 2021, doi: 10.1109/ACCESS.2021.3100717.





- [22] N. Nayef *et al.*, “ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition--RRC-MLT-2019,” in *2019 International conference on document analysis and recognition (ICDAR)*, Jul. 2019, pp. 1582–1587.
- [23] S. Gunna, R. Saluja, and C. V Jawahar, “Towards boosting the accuracy of non-latin scene text recognition,” in *ICDAR 2021: Document Analysis and Recognition – ICDAR 2021 Workshops*, 2021, pp. 282–293.
- [24] N. Gandhewar, S. R. Tandan, and R. Miri, “Deep learning based framework for text detection,” in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Feb. 2021, pp. 1231–1236, doi: 10.1109/ICICV50876.2021.9388529.
- [25] P. Adarsh, P. Rathi, and M. Kumar, “YOLO v3-tiny- object detection and recognition using one-stage improved model,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 687–694, doi: 10.1109/ICACCS48705.2020.9074315.
- [26] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, “Curved scene text detection via transverse and longitudinal sequence connection,” *Pattern Recognition*, vol. 90, pp. 337–345, Jun. 2019, doi: 10.1016/j.patcog.2019.02.002.
- [27] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, “Real-time scene-text-detection with differentiable binarization,” in *Proceedings of the AAAI conference on artificial intelligence*, Nov. 2020, pp. 11474–11481.
- [28] L. Deng, Y. Gong, X. Lu, Y. Lin, Z. Ma, and M. Xie, “STELA: a real time scenetext-detector with learned anchor,” *IEEE Access*, vol. 7, pp. 153400–153407, Sep. 2019, doi: 10.1109/ACCESS.2019.2948405.
- [29] L. Gao, X. Yi, Y. Liao, Z. Jiang, Z. Yan, and Z. Tang, “A deep learning-based formula detection method for PDF documents,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Nov. 2017, pp. 553–558, doi: 10.1109/ICDAR.2017.96.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.

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 PhD 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@presidencyuniversity.in.



Dr Shilpa Mehta     is a BE 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 PhD in ECE from JNTU Ananthapur (Andhra Pradesh). Dr Shilpa Mehta 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.