A Deep Learning-Based Approach for Robust Text Detection and Recognition in Natural Scene Images

Somnath Saha [1], Dr. Narendra Chaudhari [2]

[1,2] Department of Computer Science & Engineering, Mansarovar Global University, Sehore, MP, India.

ABSTRACT

Recent years have seen a maturation of scene text detection and recognition research methodologies, with an increased focus on improving accuracy, due to the growing relevance of this field in real life. It is challenging to satisfy the application's needs using multiple ways when dealing with complicated backgrounds in natural settings, variable font layouts, lighting, etc. Text identification and recognition in complicated natural scene photos is addressed in this study by introducing a very effective deep learning approach. The suggested system captures contextual information in character sequences using a Bidirectional Long Short-Term Memory (BLSTM) network and a ResNet-based convolutional neural network (CNN) for deep visual feature extraction. The DenseNet architecture is used for text recognition because of its high feature propagation and reuse capabilities. The final character categorization is done by a Softmax classifier. Partitioned into training, validation, and test sets in a 3:1:1 ratio, the VOCdevkit and MSRA-TD500 datasets are used to train and assess the model. Evaluative metrics included F-measure, Precision, and Recall, and training was executed over 50 epochs using an NVIDIA Tesla M40 GPU with Keras and TensorFlow. The experimental findings show that our strategy outperforms other current methods on the benchmark datasets ICDAR 2011 and ICDAR 2013. As an added bonus, DenseNet-based identification reached a peak accuracy of 94%, proving that our system is reliable and resilient in identifying English and Chinese letters in a variety of stressful visual environments.

Keywords: *Text, Deep Learning, Detection, Recognition, Scene.*

I. INTRODUCTION

A major area where computer vision has advanced thanks to recent developments in AI and deep learning is the ability to detect and recognize text in photos of real scenes. Text detection and recognition is the art and science of extracting text from photographs, often captured in natural settings. Autonomous driving, digital document processing, social media monitoring, and human-computer interaction are just a few of the many applications that rely on this specialized skill, scene text recognition. In the past, methods for text detection and recognition in natural scenes relied primarily on traditional image processing techniques and handcrafted features. However, with the recent integration of deep learning methods, the accuracy and efficiency of these tasks have been greatly enhanced.

Due to their intrinsic complexity, natural scene photos represent a distinct difficulty in text identification and recognition compared to controlled situations like scanned papers. Identification and extraction of text from these photos is extremely difficult due to the text's embedded orientation, size, font, and backdrop variations. Further obstacles to obtaining high identification accuracy include noise, distortions, illumination, occlusions, and crowded backdrops. Deep learning techniques, such as CNNs, RNNs, and, more lately, Transformer-based models, have become effective tools for scene text identification in response to these difficulties. In terms of managing real-world aberrations, contextual information from photos, and different text appearances, these models have proven to be superior.

Figuring out where text is located in pictures is a major goal of text recognition in natural situations. In the beginning, methods for detecting text used linked component analysis, edge detection, and segmentation. However, when faced with complicated backdrops or text that varied in orientation and typeface, these approaches were prone to mistakes. Thanks to convolutional neural network (CNN) based models, end-to-end learning frameworks that can locate and identify text in pictures have been made possible with the rise of deep learning. These techniques, like R-CNN and its variations, have shown to be quite effective in text localization, displaying high levels of accuracy and resilience in difficult situations.

Recognizing the textual content follows the detection of text sections. It is necessary to extract meaningful character sequences from the localized text areas in order to perform scene text recognition. Optical character recognition (OCR) methods, which entail pre-processing, feature extraction, and classification steps, were once employed in conventional procedures. But complicated and noisy language was sometimes too much for these algorithms to process. Integrating RNNs with Long Short-Term Memory (LSTM) networks has completely changed this step. Now, models can collect text dependencies sequentially, which is crucial for picture text recognition because letters aren't always aligned or cleanly separated. Recognizing text of varied lengths and orientations has been made easier using RNNs' capacity to process sequential input, leading to improved recognition accuracy in pictures of natural scenes.

Text recognition in scenes has been considerably improved with the application of attention processes in deep learning models. The model is able to process text by concentrating on pertinent areas of the image, much like people do while reading: via attention processes. amid situations where the text is partially obscured or situated amid chaotic backdrops, this feature becomes even more helpful. Using self-attention processes, transformer-based models have shown remarkable improvement in scene text recognition tests. This is because the model is better able to grasp the image's contextual information and long-range relationships.

In order to increase the accuracy of text in photos of natural scenes, post-processing techniques are typically necessary, in addition to detection and identification. Improving the output and decreasing mistakes may be achieved by including post-processing methods like language modeling, spell-checking, and contextual corrections into the recognition pipeline. By integrating deep learning models with language models that comprehend syntax and grammar, for instance, recognition performance may be greatly improved, particularly in noisy settings. Successful text identification and recognition using deep learning techniques has been greatly aided by the increasing availability of large-scale annotated datasets. Some datasets that have been useful for training and evaluating deep learning models are MSRA-TD500 and ICDAR (International Conference on Document Analysis and Recognition). These datasets include large collections of natural scene photos complete with text annotations. Thanks to these datasets and developments in transfer learning, sophisticated text detection and recognition systems are now possible. Models may be trained on massive quantities of data and adjusted for individual tasks.

A number of obstacles persist in the creation of reliable text detection and identification systems for pictures of nature scenes, notwithstanding the remarkable advancements achieved in the area. Managing text in hostile environments, such as low-resolution pictures, weak text-to-background contrast, or severely deformed text, is a significant obstacle. An extra degree of difficulty arises when dealing with natural settings that contain writing in more than one language or script. Models should be able to deal with a wide variety of characters, as well as complicated language structures and varied writing styles.

The processing of data in real-time is another major obstacle. Addressing concerns relating to compute efficiency, latency, and memory limits is necessary for deploying deep learning-based models for real-time applications like autonomous cars or mobile devices, even if these approaches have demonstrated outstanding performance in offline environments. Deep learning model optimization using methods like model compression, quantization, and pruning has been the subject of much study in an effort to fulfill these goals by reducing model size and processing costs without sacrificing accuracy.

The significance of text identification and detection in photographs of natural scenes goes well beyond the realm of academia and has several practical uses. Recognition of road signs, traffic lights, and other textual information from street sceneries is essential for autonomous driving's navigation and safety. The ability to accurately recognize text is crucial in the field of augmented reality as it allows for the seamless integration of the virtual and physical realms. People who are blind or have low vision can also benefit from assistive technology that use scene text recognition to interpret text displayed in their environment. There are several potential applications for picture text recognition, including content-based image retrieval systems, digital archiving, and monitoring.

II. REVIEW OF LITERATURE

Li, Xuexiang. (2022) In this research, we build a deep learning model for detecting and recognizing text in real scenes by studying scenes in great detail and building models. In cases when recognizing natural scene text is difficult owing to the significant backdrop and text complexity, this work presents a scene text recognition approach that relies on connection time classification and attention mechanisms. To circumvent the issue of diminished recognition performance as a whole caused by the difficulties of character segmentation, the approach shifts the focus from text recognition in natural settings to sequence recognition. Reducing network complexity and improving recognition accuracy are both achieved with the use of the attention mechanism. This research compares the performance of an enhanced PSE-based text identification method using two curved text datasets, SCUT-ctw1500 and ICDAR2017, in real-world scenarios. Excluding the pre-training module, the findings reveal that the suggested method attains 81.3% in the F1 value index, 77% in recall, and 88.5% in accuracy. This paper tests an improved CRNN-based text recognition algorithm on the ICDAR2017 natural scene dataset; under no constraints, the results show an accuracy rate of 94.5%, demonstrating good performance. The algorithm can adequately detect text in any direction without incorporating the pre-training module.

Ali, Asghar et al., (2022) Recognizing text in photographs of natural scenes is one of the most difficult problems in computer vision. Due to factors such as varying text sizes, colors, fonts, orientations, complicated backdrops, occlusion, illuminations, and uneven lighting conditions, text detection in real scene photos is more challenging than optical character recognition (OCR). Our approach to the challenge of cursive text detection, with an emphasis on Urdu language in real situations, is a deep convolutional recurrent neural network-based, segmentation-free solution. Urdu text recognition is more challenging than non-cursive scripts because of the script's stylistic variances, many character forms, linked text, ligature overlapping, stretched, diagonal, and compressed text. Without first segmenting the input image into individual letters, the suggested model takes an entire word picture as input and converts it into a series of relevant attributes. For feature extraction and encoding, our model employs a deep convolutional neural network (CNN) with shortcut connections. For feature decoding, we use a recurrent neural network (RNN). Finally, for mapping predicted sequences into target labels, we employ a connectionist temporal classification (CTC). In order to improve the text recognition accuracy even further, we investigate deeper convolutional neural network (CNN) architectures such as VGG-16, VGG-19, ResNet-18, and ResNet-

34. Our goal is to identify the aspects of Urdu text that are most relevant to the language and then compare the findings. The studies are carried out using a newly-developed large-scale benchmark dataset consisting of cropped photographs of Urdu words in realistic settings. The suggested deep CRNN network with shortcut connections outperforms existing network topologies, according on the testing findings.

Sun, Weiwei et al., (2022) The emergence of smart cities has made it possible to precisely find and detect text content inside images. This has important applications in areas like as real-time translation, picture retrieval, identification of card surfaces, and license plate recognition. As a result, people's lives and jobs will be made easier and more pleasant. Textual characteristics from photos are difficult to recognize since text may be in many different orientations, forms, and sizes. Hence, to better recognize and identify slanted text in photos, we provide an enhanced EAST detection method. A controller for a recurrent neural network is trained using reinforcement learning in the suggested approach. In order to extract text features at several scales, the best fully convolutional neural network structure is chosen. The data is then imported into the output module, where it is enhanced using the Generalized Intersection over Union method so that the text bounding box has a stronger regression impact. The enhanced text identification results are then generated after adjusting the loss function to achieve a balance between positive and negative sample classes. The suggested approach is able to enhance the poor recall rate in target recognition and solve the issue of category homogeneity, according to the experimental findings. The suggested technique outperforms competing image detection algorithms in detecting slanted text in photos of real-world scenes. Lastly, it excels at text recognition even in complicated settings.

Khan, Tauseef et al., (2021) The enormous success of deep learning models in recent decades has greatly improved text detection in the wild. This thriving age of deep learning has seen the emergence and transformation of computer vision applications. The field of text identification from natural scene photos has seen tremendous changes in the last decade, with models based on deep neural networks making tremendous strides in terms of coverage, performance, and methodology. Specifically, this paper provides the following: (1) an extensive literature review on scene text detection using deep learning; (2) a critical evaluation of appropriate deep frameworks for this task; (3) a classification study of publicly available scene image datasets and relevant standard evaluation protocols outlining their advantages and disadvantages; and (4) comparative findings and analysis of reported methods. We also specifically identify potential future scopes and thrust areas of deep learning techniques towards word identification from natural scene photos that forthcoming researchers may concentrate on, based on our evaluation and analysis.

Cao, Dongping et al., (2020) The identification of text in scenes is quickly becoming a hot subject in the field of machine vision. Recent advances in deep learning and the mobile Internet of things have allowed researchers to make great strides in the field of text identification. Aiming to synthesize and evaluate the main obstacles and noteworthy advancements in scene text detection research, this review provides a comprehensive overview. We begin by providing a brief overview of scene text detection's development and evolution, before moving on to a detailed classification of both conventional and deep learning-based approaches, highlighting the relevant challenges and solutions. After that, we provide evaluation methodologies and widely used benchmark datasets, and then we compare them to find the best algorithms. We conclude by reviewing the results and making some predictions about where the field may go from here in terms of future studies.

V, Anantha et al., (2020) The computer vision issue of text identification and recognition in photographs of natural scenes has long been a source of frustration for computer engineers. Innovations in deep learning have completely altered the landscape of computer vision. In order to decipher text in photos of real-world scenes, this research seeks to construct a text detection and identification model using Deep

Learning (DL). Identifying potential text regions, extracting those regions, and recognizing the text are the three phases that make up the suggested model. Prior to processing the natural scene picture, it is sent via the candidate text region identification technique, which identifies possible areas that contain text characters. In stage two of processing, areas that were presented in stage one but contain non-text are filtered out. After the second step, the set of text areas is identified in the final stage. When looking for potential regions of text to include, the MSER algorithm is used. The proposed approach applies two convolutional neural networks: one for text area extraction and another for text recognition. Recognizing text in real-life settings is harder than it seems. Text character identification and recognition in natural scene photos is challenging for several reasons, including the great variety of both the text and the environment itself, the existence of diverse disturbances, varying lighting conditions, and variations in text color, size, and area. To train and validate our models, we employ the ICDAR-2011, ICDAR-2013, CHARS-74K, and CIFAR-100 datasets.

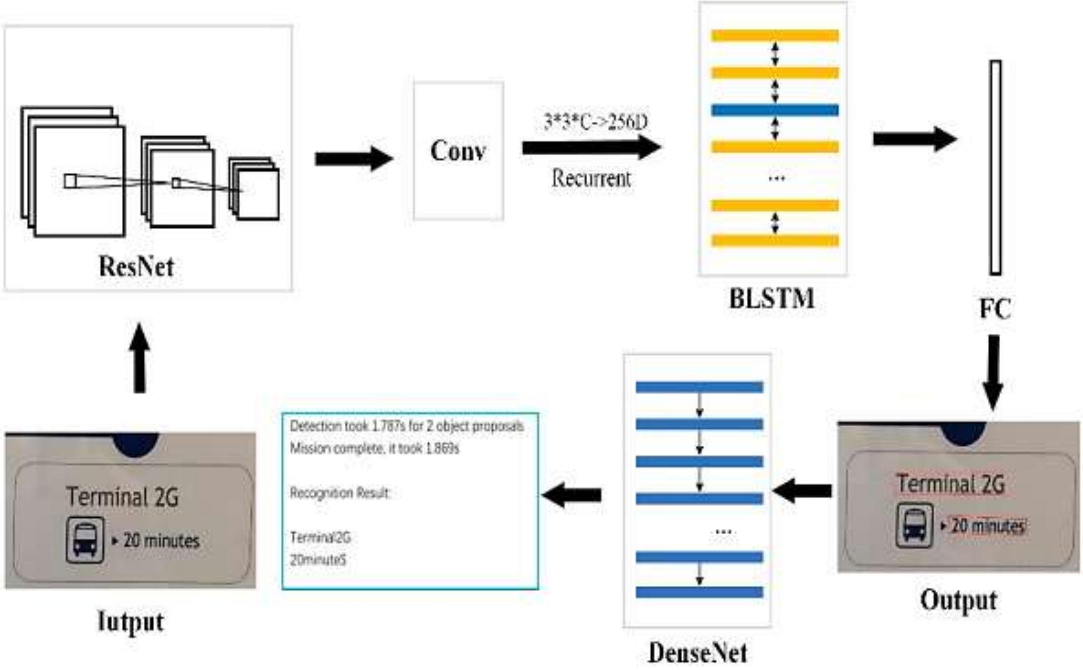
Long, Shangbang et al., (2018) The emergence and advancement of deep learning have significantly revolutionized and redefined computer vision. Scene text identification and recognition, a significant domain in computer vision, has been profoundly impacted by this revolutionary wave, hence transitioning into the deep learning age. In recent years, the community has experienced significant progress in mentality, methodology, and performance. This survey seeks to summarize and analyze the principal developments and notable advancements in scene text identification and recognition over the deep learning era. This essay aims to: (1) present novel insights and concepts; (2) emphasize modern methodologies and benchmarks; (3) explore forthcoming developments. We will highlight the significant disparities introduced by deep learning and the substantial issues that persist. We anticipate that this review article will function as a reference work for scholars in this domain.

Xu, Yin et al., (2013) Text identification in natural scene photographs is a crucial precondition for several content-based image analysis applications. This work presents a precise and resilient approach for identifying text in natural scene photographs. An efficient pruning approach is developed to find Maximally Stable Extremal Regions (MSERs) as character candidates by reducing regularized variations. Character candidates are categorized into text candidates using the single-link clustering technique, whereby distance weights and the clustering threshold are autonomously determined using an innovative self-training distance metric learning algorithm. The posterior probability of text candidates associated with non-text are assessed using a character classifier; candidates with elevated non-text probabilities are discarded, while texts are recognized using a text classifier. The suggested system is assessed using the ICDAR 2011 Robust Reading Competition database, achieving an F-measure above 76%, significantly surpassing the state-of-the-art performance of 71%. Experiments on multilingual, street view, multi-orientation, and born-digital databases further validate the efficacy of the suggested strategy.

III. EXPERIMENTAL SETUP

The purpose of this research is to offer a text detection and recognition system that is capable of handling complicated natural scene photos accurately. When text detection is conducted, a deep convolutional neural network (CNN) with a ResNet backbone is utilized. This type of CNN is able to successfully extract high-level characteristics from complex visual input. Both a BLSTM layer and a Faster R-CNN with vertical anchors are utilized in order to effectively detect text boundaries. This is done in order to capture the context of character sequences. Due to its capacity to improve feature propagation and reuse through cross-layer connections, a DenseNet-based model is applied for recognition and is therefore deployed.

As the last output layer for character classification, a Softmax classifier is utilized there. Images are mapped using ResNet, 3×3 convolutions are applied, and sequential features are processed using a 256D BLSTM, which is then followed by a fully connected layer. This architecture is depicted in figure 1. Text is the category that labels anchors that have scores that are more than 0.8. Once the model has been trained, it is then utilized to identify and recognize text in natural scene photographs that have not been seen before.



```
Detection took 1.787s for 2 object proposals
Mission complete, it took 1.869s

Recognition Result:

Terminal2G
20minuteS
```

Figure 1: Model Architecture for Detection and Recognition

Dataset Used

In this experiment, we utilized the VOCdevkit and MSRA-TD500 datasets, randomly partitioning them into training, validation, and test sets in a 3:1:1 ratio. The model underwent training for 50 epochs, with validation conducted after each epoch, followed by a final assessment on the test set. Performance was evaluated in three phases: text detection, text recognition, and study of factors affecting outcomes. Precision, Recall, and F-score were employed to assess detection accuracy. The experiment was conducted in Python 2.7 utilizing TensorFlow and Keras, with training executed on an NVIDIA Tesla M40 GPU with the SGD optimizer.

IV. RESULTS AND DISCUSSION

Image Text Detection

Using the Annotation, Image Sets, and JPEG photos files, we converted all of the images into a VOC data set. In the Annotation folder, you'll find primarily xml files that represent images; these files contain the location and category information of the tagged targets; and their names are typically identical to the original images.

In contrast, ImageSets simply require access to the Main folder, which has a few text files (e.g., train.txt, test.txt, etc.) that contain the names of the images to be trained or tested (without suffixes or paths). The JPEG Images folder holds the original images that we've named according to the uniform rules. We feed the dataset into the network during training, run 50 iterations with 30 images per batch (batchsize=30), check our work together after each iteration, and then store the best model as a detection model.

The ICDAR 2011 and ICDAR 2013 benchmark data sets served as the basis for our evaluation of the test findings. Various perspectives, tiny sizes, and poor resolution make these test data sets difficult.

Our evaluation outcomes for two publicly available datasets are displayed in Table 1. Our approach provides optimal performance on both data sets when compared to other current methods [Stroke Feature Transform, Unified text, Robust text, FASText, Local text, Symmetry text, TextFlow].

Table 1: Results on the ICDAR 2011 and ICDAR 2013

ICDAR 2011				ICDAR 2013			
Method	**Precision**	**Recall**	**F-Measure**	**Method**	**Precision**	**Recall**	**F-Measure**
Stroke Feature Transform	0.79	0.74	0.74	Robust text	0.86	0.65	0.74
Unified text	0.79	0.68	0.74	Local text	0.80	0.70	0.75
Robust text	0.84	0.70	0.78	FASText	0.82	0.67	0.75
Symmetry text	0.82	0.74	0.81	Symmetry text	0.86	0.72	0.82
TextFlow	0.85	0.74	0.83	TextFlow	0.87	0.75	0.82
ResNet (Proposed)	0.89	0.79	0.85	ResNet (Proposed)	0.94	0.82	0.85

When tested on the ICDAR 2011 and ICDAR 2013 datasets, the suggested ResNet-based approach outperformed several state-of-the-art methods for text identification in natural scene photos, as shown in Table 1. Outperforming approaches like Stroke Feature Transform, Unified Text, and TextFlow—which display lower accuracy and recall scores—on the ICDAR 2011 dataset, the suggested method obtains a recall of 0.79, an F-measure of 0.85, and a precision of 0.89.

The suggested method also achieves respectable results on the more difficult ICDAR 2013 dataset, with a recall of 0.82, F-measure of 0.85, and accuracy of 0.94. This is a huge leap forward over former top approaches like Symmetry Text and TextFlow, which only achieved maximum accuracy ratings of 0.86 and 0.87, respectively. With the most recent approach TextFlow, it has shown remarkable improvement in Recall/F-measure and raised the P-value from 0.87 to 0.94, particularly on the ICDAR 2013 data set.

Figure 2 shows the results of our method's testing on the VOCdevkit data set. Red boxes demarcate the areas where our technology automatically detects text in photos of natural scenes. It turns out that our method is foolproof in these tough cases, even if many other methods have failed miserably in them. It is clear that our approach works well for complicated and difficult natural scene image identification tasks.

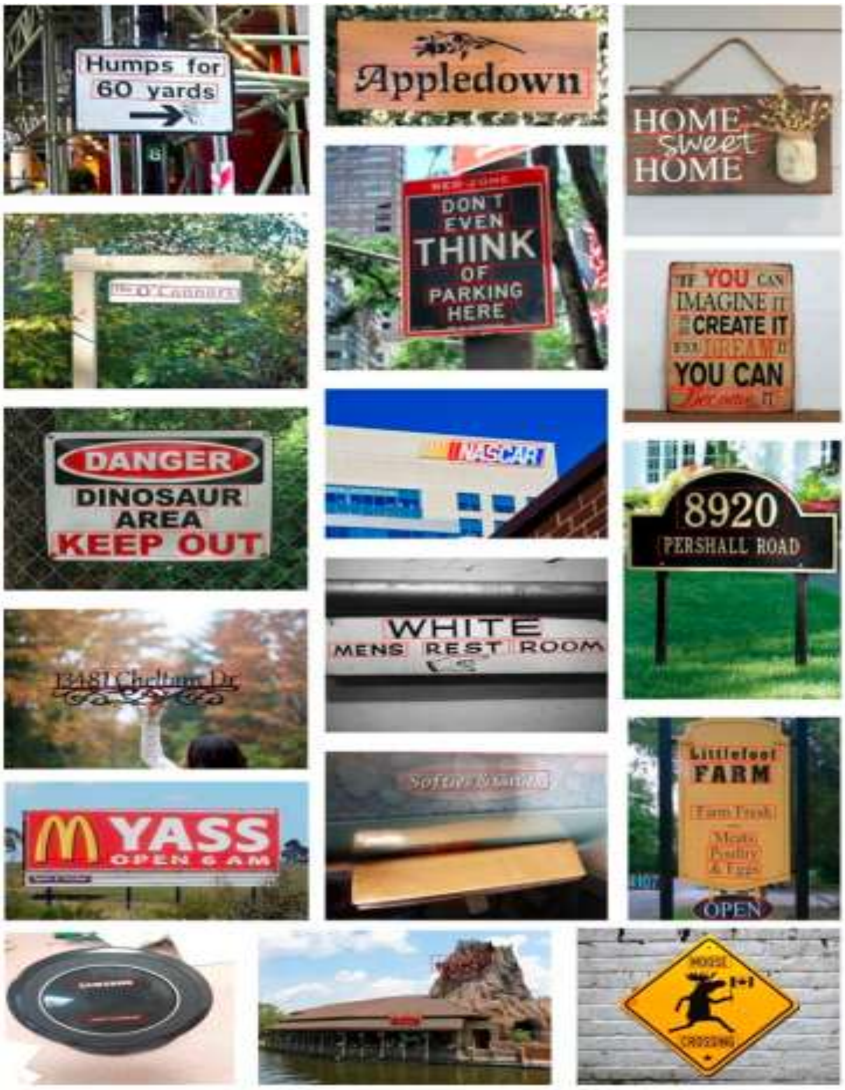


Figure 2: Visual Effect for Text Detection

Image Text Recognition

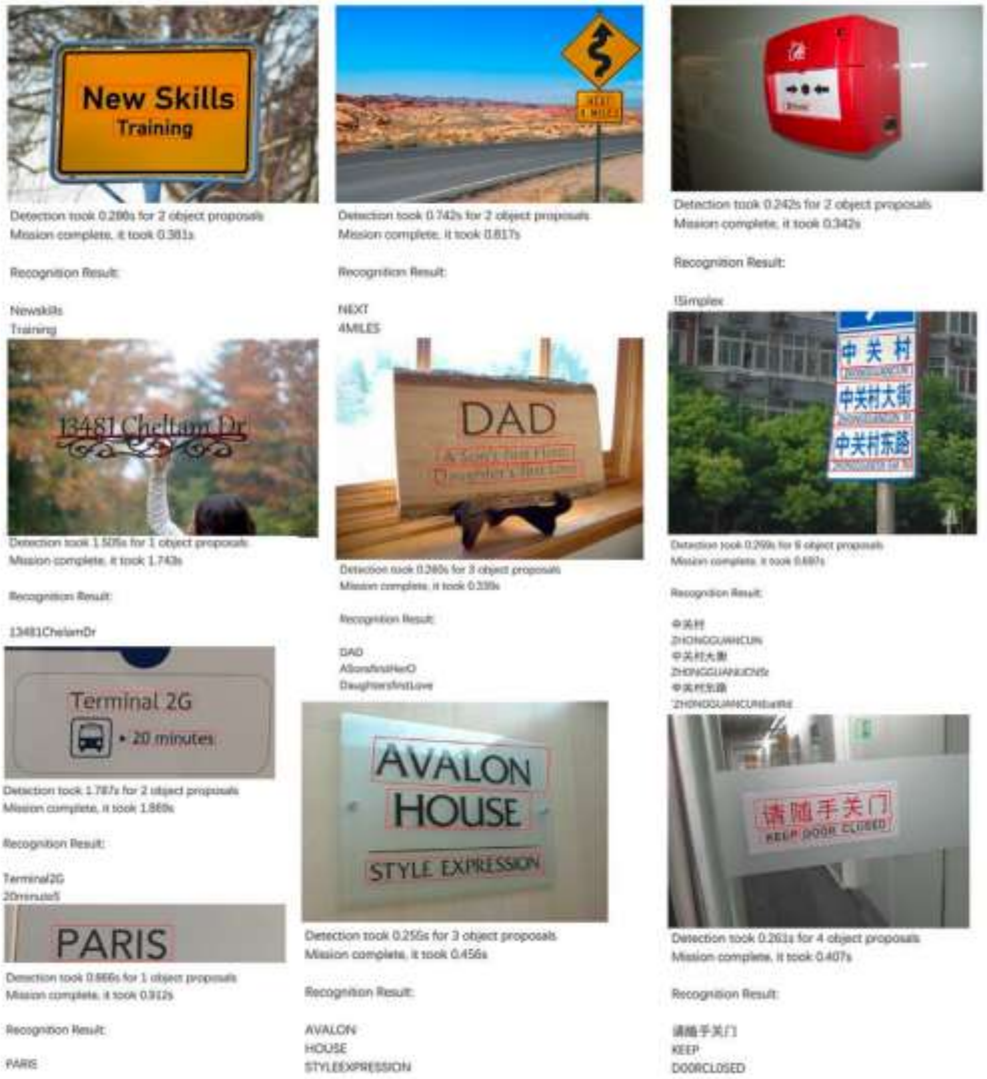
In order to train a model for text recognition, we use DenseNet. Using VOCdevkit, our training data set iteratively inputs 30 photos (batch size=30) at a time, verifies each iteration, and then saves the best model as the prediction model after 50 epochs. We use the image's text area detection to feed the trained recognition model, and then we integrate it with the corpus to identify the associated characters.

Table 2: Comparing Recognition Results Between Existing Methods and Proposed Methods

Model	Training Sets	Testing Sets	Accuracy
LeNet	254500	9000	0.869
NinNet	254500	9000	0.825
VggNet	254500	9000	0.932
DenseNet (Proposed)	254500	9000	0.940

Our final findings for recognition are shown in Table 2. Comparing the LeNet and NinNet methods on the VOCdevkit data set, we find that the former achieves a recognition rate of 0.869 and the latter of just 0.825. Using the VGGNet technique improves the accuracy to 0.932. The incorporation of DenseNet into our technique yields an accuracy of 0.940, demonstrating the efficacy of our approach.

Figure 3 shows a recognition effect map on the MSRA-TD500 and VOCdevkit datasets. Our technique performs well when it comes to text recognition of real-life photos, as seen in the figure. Almost all Chinese characters can be deciphered, in addition to English ones. There are still some blunders, but there are also many objective considerations. Things like lighting, camera angles, and font distortions etc. Furthermore, it is not possible to accurately determine how the text area affects the following character recognition.



```
Detection took 0.280s for 2 object proposals
Mission complete, it took 0.381s

Recognition Result:

Newskills
Training
```

```
Detection took 1.505s for 1 object proposals
Mission complete, it took 1.743s

Recognition Result:

13481ChelamDr
```

```
Detection took 1.787s for 2 object proposals
Mission complete, it took 1.869s

Recognition Result:

Terminal2G
20minute5
```

```
Detection took 0.866s for 1 object proposals
Mission complete, it took 0.912s

Recognition Result:

PARIS
```

```
Detection took 0.742s for 2 object proposals
Mission complete, it took 0.817s

Recognition Result:

NEXT
4MILES
```

```
Detection took 0.280s for 3 object proposals
Mission complete, it took 0.339s

Recognition Result:

DAD
ASonsfirstHerO
DaughtersfirstLove
```

```
Detection took 0.255s for 3 object proposals
Mission complete, it took 0.456s

Recognition Result:

AVALON
HOUSE
STYLEEXPRESSION
```

```
Detection took 0.242s for 2 object proposals
Mission complete, it took 0.342s

Recognition Result:

ISimplex
```

```
Detection took 0.269s for 6 object proposals
Mission complete, it took 0.687s

Recognition Result:

中关村
ZHONGGUANCUN
中关村大街
ZHONGGUANJCNSt
中关村东路
'ZHONGGUANCUNEastRd
```

```
Detection took 0.261s for 4 object proposals
Mission complete, it took 0.407s

Recognition Result:

请随手关门
KEEP
DOORCLOSED
```

Figure 3: Visual Effect for Text Recognition

V. CONCLUSION

This study indicates that a deep learning-based system can effectively detect and recognize text in complex images of natural situations. By integrating ResNet for robust feature extraction, BLSTM for contextual understanding, and DenseNet for enhanced recognition accuracy, the proposed system outperforms earlier methods significantly. This is particularly true when handling complex and diverse circumstances like low resolutions, small sizes, and variable text orientations. This model shows significant potential for practical usage in sectors like document analysis, autonomous automobiles, and augmented reality with its remarkable text recognition accuracy and outstanding performance on benchmark datasets like ICDAR 2011 and ICDAR 2013.

REFERENCES

1. X. Li, "A Deep Learning-Based Text Detection and Recognition Approach for Natural Scenes," *J. Circuits Syst. Computer*, vol. 32, no. 05, pp. 1–19, 2022.
2. A. Ali, M. Asikuzzaman, M. Pickering, and M. Leghari, "Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 10062–10078, 2022.
3. W. Sun *et al.*, "Deep-Learning-Based Complex Scene Text Detection Algorithm for Architectural Images," *Mathematics*, vol. 10, no. 20, pp. 1–22, 2022.
4. P. Pathak, P. Gupta, N. Kishore, N. Yadav, and Dr. Chaudhary, "Text Detection and Recognition: A Review," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 5, pp. 2733–2740, 2022.
5. R. Harizi, R. Walha, and F. Drira, "Deep-learning based end-to-end system for text reading in the wild," *Multimedia Tools Appl.*, vol. 81, no. 12, pp. 1–12, 2022.
6. Q. Zhao, "Researches advanced in Natural Scenes Text Detection Based on Deep Learning," *Highlights Sci. Eng. Technol.*, vol. 16, pp. 188–197, 2022.
7. R. Ranjbarzadeh *et al.*, "A Deep Learning Approach for Robust, Multi-oriented, and Curved Text Detection," *Cogn. Comput.*, vol. 16, no. 4, pp. 1–13, 2022.
8. T. Khan, R. Sarkar, and A. Mollah, "Deep learning approaches to scene text detection: a comprehensive review," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 1–60, 2021.
9. S. Mahajan and R. Rani, "Text detection and localization in scene images: a broad review," *Artif. Intell. Rev.*, vol. 54, no. 2, pp. 1–24, 2021.
10. S. Long, X. He, and C. Yao, "Scene Text Detection and Recognition: The Deep Learning Era," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 1–16, 2021.
11. M. Chaitra, D. Ramegowda, M. T. Gopalakrishna, and B. Prakash, "Deep-CNNTL: Text Localization from Natural Scene Images Using Deep Convolution Neural Network with Transfer Learning," *Arab. J. Sci. Eng.*, vol. 47, no. 9, pp. 1–28, 2021.
12. D. Cao, Y. Zhong, L. Wang, Y. He, and J. Dang, "Scene Text Detection in Natural Images: A Review," *Symmetry*, vol. 12, no. 1, pp. 1–27, 2020.
13. A. V, M. Kumar, and V. Tamizhazhagan, "Text Region Detection and Recognition in Natural Scene Images Using MSER and Convolutional Neural Network," *SSRN Electron. J.*, vol. 2, no. 1, pp. 451–461, 2020.
14. A. Agrahari and R. Ghosh, "Multi-Oriented Text Detection in Natural Scene Images Based on the Intersection of MSER With the Locally Binarized Image," *Procedia Comput. Sci.*, vol. 171, pp. 322–330, 2020.
15. U. Yasmeen *et al.*, "Text Detection and Classification from Low Quality Natural Images," *Intell. Autom. Soft Comput.*, vol. 26, no. 4, pp. 1251–1266, 2020.
16. X. Liu, G. Meng, and C. Pan, "Scene text detection and recognition with advances in deep learning: a survey," *Int. J. Doc. Anal. Recognit.*, vol. 22, no. 5, pp. 143–162, 2019.
17. S. Long, X. He, and C. Ya, "Scene Text Detection and Recognition: The Deep Learning Era," *J. LaTeX Class Files*, vol. 10, no. 10, pp. 1–24, 2018.
18. Z. Zhao, C. Fang, Z. Lin, and Y. Wu, "A Robust Hybrid Method for Text Detection in Natural Scenes by Learning-based Partial Differential Equations," *Neurocomputing*, vol. 168, pp. 1–19, 2015.
19. A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014.
20. Y. Xu, X. Yin, K. Huang, and H. H. Hao, "Robust Text Detection in Natural Scene Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, pp. 1–10, 2013.