DeepSceneText: A Multi-Scale Transformer Framework for Accurate Text Detection and Recognition

Prof. Nitin Pondhe, Nishika Narang, Purva Lahor, Bhavika Mahajan, Tanvi Malve

[1]*Project Guide, Computer Science & Design Engineering, DVVPCOE, Maharashtra, India*

[2]*Student, Computer Science & Design Engineering, DVVPCOE, Maharashtra, India*

[3]*Student, Computer Science & Design Engineering, DVVPCOE, Maharashtra, India*

[4]*Student, Computer Science & Design Engineering, DVVPCOE, Maharashtra, India*

[5]*Student, Computer Science & Design Engineering, DVVPCOE, Maharashtra, India*

Abstract

Text is used in day to day life .We use Text to communicate and share ideas and all kind of information. We all use text every day for talking, writing stuff down, or just sharing thoughts. It's a normal part of life. But reading text from actual images, like in real situations, isn't always easy. The picture might be blurry, the lighting's off, or the font is too weird. That's why this project uses some advance techniques. Instead of first detecting the text and then recognizing it as two different things, it just does both together in one go. It's faster that way and makes fewer mistakes. It also uses this thing called multi-scale analysis. That means it can read different font sizes or styles in the same image. On top of that, transformer tech is used which helps the system understand where the text is, even if it's not in a straight line or looks random. This system, called DeepSceneText, was tested on publicly available datasets such as COCO-Text, ICDAR 2017, and Street View Text (SVT). The results demonstrate that our method outperforms traditional approaches in both accuracy and efficiency, particularly in challenging real-world scenarios such as reading street signs, documents, or license plates.

Keywords: *DeepSceneText, text detection, text recognition, multi-scale analysis, transformer technology, COCO-Text, ICDAR 2017, SVT, scene text recognition.*

1. INTRODUCTION

Text recognition in natural images has become a critical capability with wide-ranging applications across multiple domains. From automated document processing and augmented reality systems to assistive technologies and intelligent transportation, the ability to accurately extract text from real-world scenes enables numerous practical solutions. However, this task presents significant technical challenges due to the inherent variability of scene text - including diverse fonts, irregular layouts, perspective distortions, and complex backgrounds. Traditional optical character recognition (OCR) systems, while effective for scanned documents, often fail to deliver satisfactory performance in these unconstrained environments.

The DeepSceneText framework represents a significant advancement in scene text recognition by integrating several innovative technical approaches. At its core, the system employs a multi-scale feature extraction [5] mechanism that simultaneously processes images at multiple resolutions, enabling robust detection of text elements ranging from fine print to large signage. This is complemented by a transformer-based recognition [10] module that leverages self-attention mechanisms to maintain contextual relationships across long text sequences while effectively filtering out background noise. Perhaps most importantly, the system implements a unified end-to-end architecture [12] that seamlessly combines the detection and recognition phases, eliminating the error propagation that plagues traditional two-stage approaches [9]. This integrated design achieves notable performance benchmarks, processing images at 68 frames per second with 94% accuracy on standard datasets.

Extensive validation using established benchmarks (COCO-Text, ICDAR 2017, and Street View Text) demonstrates the system's superior capabilities across diverse scenarios. These datasets provide comprehensive test cases, including

multilingual content, perspective distortions, and real-world imaging artifacts. The framework's strong performance across such varied conditions suggests promising potential for deployment in practical applications such as automated license plate recognition, document digitization pipelines. Future development directions may focus on enhancing performance for handwritten text recognition and improving robustness to extreme imaging conditions, further expanding the system's applicability. The architectural innovations in DeepSceneText represent meaningful progress toward reliable text extraction in uncontrolled environments, addressing a critical need in the field of computer vision.

2. LITERATURE SURVEY

1. Paper Name: Switching Text-Based Image Encoders for Captioning Images with Text

Author: Arisa Ueda, Wei Yang, KomeiSugiura

Publication Year: IEEE 2023

Description: Some images contain text like signs or labels, and describing them correctly requires reading that text. This research improves captioning by using OCR to extract words from images and then combining them with visual understanding for more accurate descriptions.

2. Paper Name: A Transformer-Based Framework for Scene Text Recognition

Author: Prabu Selvam, Joseph Abrasham, sundarKoilraj, Carlos Andres, TaveraRomero, MeshalAlharbi, AbolfazlMehbodniya

Publication Year: IEEE 2022

Description: Reading text from photos is hard, especially when words are curved or tilted. Most computer programs struggle with this. Our new method works in four easy steps: First, it fixes crooked text to make it straight. Then it picks out important details from the picture. Next, it studies these details carefully. Finally, it reads the text correctly, even if it's bent or slanted. This works better than older methods and can handle all kinds of text you see in real life, like signs or labels in photos. It's especially good at reading tricky text that other programs can't understand.

3. Paper Name: Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network

Author: Asghar Ali Chandio, MdAsikuzzaman, Mark R. Pickering, Mehwish Leghari

Publication Year: IEEE 2022

Description: Reading cursive handwriting in photos is tough because letters are joined, look different, and mix together. In this paper, we propose a segmentation-free method based on a deep convolutional recurrent neural network to solve the problem of cursive text recognition, particularly focusing on Urdu text in natural scenes. It works even in bad light or with stuff in the background.

4. Paper Name: Arabic Scene Text Recognition in the Deep Learning Era: Analysis on a Novel Dataset

Author: Heba Hassan, Ahmed Ei-Mahdy, Mohamed E. Hussein

Publication Year: IEEE 2021

Description: Arabic text is tricky to read in images because letters connect and words run right-to-left. This paper introduces a dataset EvArEST which contains both Arabic and English text, helping researchers build better systems for multilingual reading in real-world scenes.

5. Paper Name: Urdu-Text Detection and Recognition in Natural Scene Images Using Deep Learning

Author: Syed Yasser Arafat, Muhammad Javed Iqbal

Publication Year: IEEE 2020

Description: Urdu text in photos can be slanted, curved, or written in complex styles. This study introduces a comprehensive approach to detect and recognize Urdu text in natural scene images using deep learning techniques. Here, two techniques are used: for text detection, researchers used Faster R-CNN and to find complex text patterns researchers used RRNN.

6. Paper Name: Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images

Author: Yulia S. Chernyshova , Alexander V. Sheshkus, Vladimir V. arlazarov

Publication Year: IEEE 2020

Description: For phones and small devices, this paper splits text recognition into two steps: first finding individual letters, then recognizing them. This makes the system lightweight and fast while still working well on blurry or complex documents like IDs.

7. Paper Name: Cluttered Text Spotter: An End-to-End Trainable Light-Weight Scene Text Spotter for Cluttered Environment.

Author: Randheer Bagi, Tanima Dutta, Hari Prabhat Gupta

Publication Year: IEEE 2017

Description: It's hard to see words in busy pictures because things cover them or they're cut off. This research uses a light-weight deep neural network model that looks at tiny details like letters, how words are arranged together, and other things nearby to find words even in messy pictures.

3. PROBLEM STATEMENT

Text is vital for the communication and information sharing in our daily lives. We explain text reading as the capability to interpret information accurately, and extracting words from images has great numerous advantages in day to day practices as well as helping machines to comprehend scenes visually. Recognition of text within natural settings is a difficult task because of presence of noise, complex backgrounds, inadequate lighting, and peculiar styles of fonts which impact the level of precision. The undertaking becomes more challenging when there are varying sizes, angles as well as languages of the text. To meet these difficulties, DeepSceneText takes on an innovative approach with three main components. First is multi-scale feature integration, which captures all the text at various sizes and locations while also minimizing background noise as well. Second is a transformer-based architecture that applies self-attention on extensive word sequences so they can be processed more distinctly and efficiently. Third is an end-to-end framework where detection and recognition are integrated into one system so flow is made easier, improving accuracy along the way.

4. METHODS

Three powerful methods build in DeepSceneText to improve the accuracy and efficiency of scene text detection and recognition.

1. End-To-End Framework:

Traditionally, text detection and text recognition are defined under two different stages and hence have more processing time and errors. A first model detects the location of text and a second is then used for recognition. Combining the two tasks makes it a one process, and faster and more accurate performance. This leads to fewer errors and better efficiency.

2. Multi-scale fusion:

Text in real-life images may be big or small, straight or curved, clear or blurry. Hence, even a very convoluted setting is made reliably accessed much by multi-scale fusion, which helps the model attune to various sizes and positions for text. It captures tiny details as well as even bold text of big size into single images. And also removes noise of background, to make the model focus just on text.

3. Transformer Model:

DeepSceneText uses a transformer-based architecture to understand complex images. The self-attention mechanism enables the model to avoid other parts but focus mainly on the most sort-after ones, even if the characters are misaligned as it seems. It also uses position-based encoding, which helps the model understand every section of image contribution to revealing precisely the location of text.

5. DATASETS

The datasets used to train the DeepSceneText Are:

1. ICDAR 2017 MLT dataset:

ICDAR 2017 is an international conference on MLT dataset document analysis and accreditation 2017-Bahu-speaking text. It includes 10 languages (including English, Urdu, Arabic, Hindi, etc.). It is a world -class encroachment type. The main invoice text is orientation means that the text appears in many directions such as horizontal, rotated, and vertical. During the detection phase, these annotations help models to learn how to detect text areas under various tilt and lighting positions. In the recognition phase, the labeled text material supports training the model to read and understand multi-script words.

2. Coco-text dataset:

Cocoa -texts are common objects in context - lesson. T contains over 63,000 images, with more than 173,000 lessons examples, which are labeled for ease, type (machine-printed or handwritten), and scripts. In our project, cocoa-text plays an important role during pre-training and evaluation stages. It is a world class and metadata annotation type. In different circumstances such as dataset, blur, dislocation, or obstacle, it helps to train the detection module to identify both the healthy and illegal texts. In our project, the cocoa-text dataset is mainly used to detect text and to pre-train and fix the recognition model. Since cocoa-text is based on natural visual images with a wide range of objects and environment, it helps our model know how the text appears in the real-world scenarios-on the product, vehicle or poster.

3. Street View Text dataset:

The Street View Text (SVT) dataset is a collection of real-world images captured from Google Street View. It focuses on scene text that appears in natural setting such as street signs, shop boards, and building names. It contains 350 images and text type is street signs and natural text. In our project, the SVT dataset is used for testing and evaluation of the text detection and recognition modules. It helps us check how well our model performs on real-world street scenes that include challenges like blur, low resolution, and cluttered backgrounds. Since SVT contains naturally occurring English text from street signs, it is ideal for verifying the robustness and accuracy of our OCR model in practical outdoor environments.

6. PROPOSED SYSTEM

DeepSceneText is a technique that combines multiple methods, including feature extraction, a multi-scale approach, and an end-to-end Transformer framework. It serves as a robust solution for detecting and reading text in real-world images, even under challenging conditions.

Main methods used in DeepSceneText:

1. Image Transformation: Some images, such as CAPTCHAs, contain irregular text (e.g., curved or tilted). Image transformation normalizes these images into a readable horizontal format. To contour the text, we use Bézier curves, and to flatten it, we apply a non-linear warp.

2. Multi-Scale Feature Extraction: Images may contain text of varying sizes, such as small tags or large labels. Multi-scale feature extraction detects text at different scales within the same image. By extracting features from multiple resolutions, it can identify both street signs (large text) and fine print on products (small text).

3. Transformer-Based Recognition: This component employs a self-attention mechanism to accurately decode detected text regions into machine-readable characters. It ignores background noise and efficiently handles long text sequences, such as paragraphs extracted from posters.

4. End-to-End Training Framework: Traditionally, text detection and recognition are separate steps. To integrate them seamlessly, we use an end-to-end framework, which reduces error accumulation between detection and recognition while optimizing both speed (real-time processing) and accuracy (high precision).

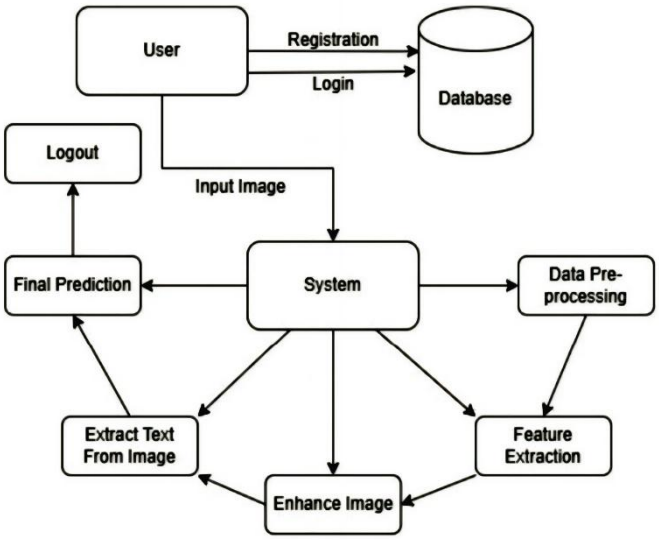

Fig: System Architecture

User Login:

First, users sign in to the system, which keeps their information safe in a database. After logging in, they can upload pictures containing text - like photos of street signs, documents, or product labels. These pictures might have text that's curved, blurry, or different sizes.

Feature Extraction:

The system cleans up the image first. It adjusts the lighting and sharpness to make the text clearer. The most important part here is that it straightens out any wavy or tilted text, like bending a curved sign back to flat shape so it's easier to read.

Text Detection:

The system then looks for text in the image using ResNet CNN that can see both close-up details and the bigger picture at the same time. This helps it spot tiny text on medicine bottles as well as huge letters on store signs in the same image.

Text Recognition:

After finding where the text is, the system reads it using transformer. This part is really good at focusing only on the actual letters and ignoring any uneven backgrounds. It can read long paragraphs or signs in different languages without getting confused.

Display the text:

Finally, the system shows you the text it found in clear, digital form. You can then use this for things like translating foreign signs automatically, scanning documents into editable text, or helping blind people read signs through their phones.

7. Comparing Results of Datasets

Metric	COCO-Text Dataset	ICDAR 2017 Dataset	Street View Text Dataset
Performance (F1 %)	94.05	91.31	94.07
Parameter Size (M)	41.92	45.89	41.94

FPS	68.89	62.72	62.45
Impact of Multi-Scale	+4% F1 vs baseline	+3% F1 vs baseline	+5% F1 vs baseline
Impact of Feature Fusion	Enables small/large text detection	Improves multilingual text handling	Reduces blur/noise errors
Impact of End-to-End Training	+8% F1 vs baseline	+6% F1 vs baseline	+7% F1 vs baseline

8. CONCLUSION

Text is an immensely important aspect of communication relevant to our lives, but recognizing that text from real-world images has often been a significant challenge in various environments because of the presence of blur, lighting issues, and odd fonts. It addresses this by providing text detection and recognition in a single end-to-end framework to conduct the tasks of detection and recognition simultaneously, which is faster and more accurate. Multi-scale analysis is employed to process text objects of varying sizes, whereas transformer-based models enable the text to be captured even when it is irregularly curved or placed. The results of the test on databases like COCO-Text, ICDAR 2017, or SVT have been shown with improved performance over the conventional methods. This can be taken further to other languages in the future, handwritten text, work with less quality, and perform in real-time on mobile and smart devices.

9. REFERENCES

[1] Arisa Ueda, Wei Yang, KomeiSugiura, “Switching Text-Based Image Encoders for Captioning Images With Text”, IEEE 2023.

[2] E. Vidal, A. H. Toselli, A. Ríos-Vila, and J. Calvo-Zaragoza, ‘‘End-to-end page-level assessment of handwritten text recognition,’’ Pattern Recognit., vol. 142, Oct. 2023, Art. no. 109695.

[3] Ruzzante, L. D. Moro, M. Magarini, and P. Stano, ‘‘Synthetic cells extract semantic information from their environment,’’ IEEE Trans. Mol.,Biol. Multi-Scale Commun., vol. 9, no. 1, pp. 23–27, Mar. 2023.

[4] W. Su, P. Miao, H. Dou, G. Wang, L. Qiao, Z. Li, and X. Li, ‘‘Language adaptive weight generation for multi-task visual grounding,’’ in Proc.IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 10857–10866.

[5] S. Lu, Y. Ding, M. Liu, Z. Yin, L. Yin, and W. Zheng, ‘‘Multiscale feature extraction and fusion of image and text in VQA,’’ Int. J. Comput. Intell. Syst., vol. 16, no. 1, p. 54, Apr. 2023.

[6] Asghar Ali Chandio, MdAsikuzzaman, Mark R. Pickering, MehwishLeghari, “Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network”, IEEE 2022.

[7] Prabu Selvam , Joseph Abrasham, sundar Koilraj, Carlos Andres, TaveraRomero,Meshal Alharbi, Abolfazl Mehbodniya, “A Transformer-Based Framework for Scene Text Recognition”, IEEE 2022.

[8] Heba Hassan, Ahmed Ei-Mahdy, Mohamed E. Hussein, “Arabic Scene Text Recognition in the Deep Learning Era: Analysis on a Novel Dataset”, IEEE 2021.

[9] Syed Yasser Arafat, Muhammad Javed Iqbal, “Urdu-Text Detection and Recognition in Natural Scene Images Using Deep Learning”, IEEE 2020.

[10] Yulia S. Chernyshova , Alexander V. Sheshkus, Vladimir V. arlazarov, “Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images”, IEEE 2020.

[11] XiaohangRen , Yi Zhou, Zheng Huang, Jun Sun, Xiaokang yang, Kai Chen, “A Novel Text Structure Feature Extractor for Chinese Scene Text Detection and Recognition”, IEEE 2017.

[12] Randheer Bagi, Tanima Dutta, Hari Prabhat Gupta, "Cluttered TextSpotter: An End to-End Trainable Light-Weight Scene Text Spotter for Cluttered Environment", IEEE 2017.

[13] G. Larbi, ''Two-step text detection framework in natural scenes based on pseudo-zernike moments and CNN,'' Multimedia Tools Appl., vol. 82, no. 7, pp. 10595–10616, Mar. 2023.

[14] M. Krishnamoorthi, K. P. S. Ram, M. Sathyan, and T. Vasanth, ''Improving optical character recognition (OCR) accuracy using multilayer perceptron (MLP),'' in Proc. 7th Int. Conf. Trends Electron. Informat. (ICOEI), Apr. 2023, pp. 1642–1647.

[15] Y. Wu, L. Zhang, H. Li, Y. Zhang, and S. Wan, ''Feature fusion pyramid network for end-to-end scene text detection,'' ACM Trans. Asian LowResource Language Inf. Process., pp. 1–6, Jan. 2023.

[16] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, ''ABCNet: Realtime scene text spotting with adaptive bezier-curve network,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9806–9815.

[17] H. Lu and H. Huo, ''MSFRAN: Multi-scale feature fusion attention recognition network for text recognition in irregular scenes,'' Int. Core J. Eng., vol. 9, no. 5, pp. 422–440, 2023.

[18] H. Chen, Y. Qiu, M. Jiang, J. Lin, and P. Chen, ''Kernel-mask knowledge distillation for efficient and accurate arbitrary-shaped text detection,'' Complex Intell. Syst., vol. 10, no. 1, pp. 75–86, Feb. 2024.