Journal of Information Systems Engineering and Management
2025, 10(14s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Attention-Based Deep Learning System for Text Detection and Information Retrieval from Images

Subhakarrao G[1], Dr. B Sujatha[2], Dr. L Sumalatha[3]
[1]*JNTUK, Kakinada, Andhra Pradesh, India, subhakar.golla@gmail.com*
[2]*GIET (A), Rajahmundry, Andhra Pradesh, India*
[3]*Department of CSE, JNTUK, Kakinada, Andhra Pradesh, India.*

ARTICLEINFO

Received: 20 Nov 2024

Revised: 02 Jan 2025

Accepted: 20 Jan 2025

ABSTRACT

This paper presents a new automatic approach to the extraction of textual information from natural photographs, addressing issues such as cluttered backgrounds, uneven lighting, and distortions. The system combines advanced computer vision techniques with OCR to detect, extract, and process text embedded in complex scenes. A deep learning-based text detection module identifies the text regions with high precision, and a robust OCR pipeline is used for extracting the detected text into a machine-readable form.

Experimental evaluations indicate how well the system works under practical real-world applications, including obtaining an average text detection precision of 92% and accuracy for character recognition to be about 88%. Broad applications for this proposed method involve real-time translation, augmented reality, tools on accessibility for visually impaired persons, intelligent traffic monitoring, and even automated content indexing. It makes a new standard for information retrieval systems by being able to correctly extract text from natural scenes and opens the gates to innovative applications in artificial intelligence and digital data processing.

Keywords: Text Detection, Optical Character Recognition (OCR), Natural Scene Images, Convolutional Neural Networks (CNN), Feature Pyramid Network (FPN), Scene Text Recognition, Attention Mechanism, Bidirectional LSTM, Document Analysis, Region Refinement, Text Spotting.

I. Introduction

In today's connected world, the exponential growth of digital pictures has made information retrieval systems face an unprecedented difficulty. The intricacies of automatic text extraction leave a great deal of useful information untapped in natural images, which often contain textual information in a variety of formats, from billboard ads and street signs to product labels and informational displays. Text integrated in real settings presents substantial hurdles for classic Optical Character Recognition (OCR) systems, notwithstanding their impressive effectiveness in processing scanned documents and other controlled situations.

Text identification and recognition in natural photographs is fundamentally complex due to a number of critical variables. Environmental fluctuations present a significant problem because nature sceneries are susceptible to continually shifting lighting conditions, which affect text visibility and contrast. Shadow effects can partially hide characters, and inclement weather can distort or blur lettering. The various viewing angles and perspectives cause geometric distortions, complicating the detecting procedure. Text features add another layer of intricacy, as natural settings contain a variety of font types, sizes, and colors. Text layouts frequently use numerous orientations or curving patterns, with varied densities and spacing between characters and words. The existence of various languages and scripts within a single image complicates the recognition process.

Because textured surfaces in natural surroundings can resemble text-like patterns and cause false positives in detection algorithms, background complexity is a third significant difficulty. Segmentation becomes more challenging when the text-background contrast is greatly diminished by complex backgrounds. These difficulties are

exacerbated by motion blur and partial text occlusions caused by other items in the scene in real-world situations. Because they frequently make assumptions about the look and positioning of text that do not hold true in natural environments, these problems have historically hindered the efficacy of classic computer vision systems. Additionally, the growing need for real-time processing in applications like augmented reality, mobile device-based translation services, and driverless cars has made the already difficult challenge even more difficult by adding the computing efficiency limitation.

This paper introduces a groundbreaking solution that tackles these issues by utilizing a holistic system framework that integrates modern advancements in deep learning with traditional computer vision methods. Our method presents numerous important contributions to the discipline. We introduce an innovative multi-scale text detection framework that adeptly manages text at different scales and angles, enhanced by a flexible segmentation algorithm that precisely distinguishes text from intricate backgrounds. The system features a strong recognition module that ensures high precision across various fonts and styles, as well as an effective information retrieval system that allows rapid access to the extracted text.

II. Background

The growth of text identification and recognition in natural photographs has spanned various technology paradigms, each delivering important advances to the area. Over the last two decades, researchers have developed increasingly sophisticated solutions to the hard task of extracting text from natural scenes.

Smith and Kanade [1] established the cornerstone of current text detection with their pioneering work, which set spatial-temporal restrictions for text detection in video sequences. This early approach highlighted the potential of structural elements for text detection, but its use was limited to controlled conditions. Building on this foundation, Jain and Yu [2] created a thorough framework for page deconstruction and text extraction using linked component analysis, defining numerous basic ideas that remain relevant today.

A significant breakthrough was realized with the work of Chenand Yuille [3], with an innovative method of adaptive learning for text identification in natural environments. A few visual characteristics, combined with a cascaded classifier, resulted in successful performance in varied conditions. This was further improved by Epshtein et al. [4], who presented the Stroke Width Transform (SWT), an innovative method that significantly enhanced text detection precision by leveraging the uniform stroke width characteristic of characters. The advent of machine learning introduced fresh viewpoints to the field. One of the earliest machine learning-based text identification methods was put out by Wu and Manmatha [5], who used Support Vector Machines (SVMs) to categorize text regions according to meticulously designed criteria. This method was developed by Neumann and Matas [6], who greatly increased robustness to geometric alterations by combining trained classifiers with Maximally Stable Extremal Regions (MSER).

The field was completely revolutionized by deep learning. Wang et al. [7] demonstrated the capability of end-to-end text recognition using CNNs, showing unprecedented accuracy on common benchmarks. Tian et al. made an impressive appearance with their Connectionist Text Proposal Network (CTPN). [8] expanded on this achievement by successfully fusing CNNs with recurrent architectures to enhance text localization.

The work of Shi et al. [9], who presented an end-to-end trainable network integrating CNNs with Recurrent Neural Networks (RNNs), witnessed comparable advances in scene text recognition. Their CRNN architecture preserved computational efficiency but set new standards for recognition accuracy. He et al. [10] further enhanced this approach by adding attention mechanisms that allow the handling of complex text layouts in a better manner.

Directed and irregular text has garnered more attention lately. The EAST detector was presented by Zhou et al. [11] and uses a novel geometric refinement technique to analyze oriented text quickly. Liao et al. [12] supplemented this by enhancing the recognition of arbitrarily oriented text through orientation-sensitive regression in their TextBoxes++ framework.

Liu et al. [13] addressed the challenge of end-to-end text spotting by proposing a unified framework that simultaneously performs detection and recognition. Their technology maintained high accuracy in various settings

while achieving real-time performance. This was subsequently developed by Wang et al. [14] with their Mask TextSpotter architecture, which handled any text shapes and multiple languages with ease.

Feng et al. [15] most recently presented a thorough framework that includes visual-semantic alignment, which greatly enhanced performance on difficult datasets with stylized and artistic text. The precision and reliability of their task are the latest trends in this field.

III. Proposed System

The main four modules which are divided between the proposed system are Text Detection Network (TDN), Region Refinement Module (RRM), Text Recognition Network (TRN), and Information Retrieval Engine (IRE). All these modules uphold modularity for optimized individual components to be followed along with an end-to-end trainable framework.

Network for Text Detection

The Feature Pyramid Network structure is modified into a Text Detection Network with more spatial attention mechanisms. ResNet-50 with feature maps at five distinct scales ({P2, P3, P4, P5, P6}) is used in the backbone network.

$$Fl = f1(Fl - 1) + ul(dl(Fl + 1))$$

Where:

fl represents the convolutional operations at level l

ul denotes the upsampling operation

dl represents the downsampling operation

Fl is the feature map at level l

The spatial attention mechanism is implemented using a self-attention module:

$$A(F) = softmax\big(\text{ø}(F)\Theta(F)\big)g(F)$$

Where:

- θ(F), ϕ(F), and g(F) are linear transformations of the input feature map
- A(F) represents the attention-enhanced feature map

Region Refinement Module

The Region Refinement Module employs a two-stage refinement process. First, candidate regions are generated using anchor-based detection:

$$R = (xi, yi, wi, hi, si)|i[1, N]$$

Where:

- (xi, yi) represents the center coordinates
- (wi, hi) denotes width and height
- si is the confidence score
- N is the number of candidate regions

The refinement process applies Non-Maximum Suppression (NMS) with an IoU threshold τ:

$$IoU(Ri, Rj) = \frac{(RiRj)}{area(RiRj)}$$

Final regions are selected based on:

$$R* = argmax(si)\ where\ IoU\ (Ri, Rj) < T$$

Text Recognition Network

The Text Recognition Network utilizes a hybrid CNN-LSTM architecture with attention mechanisms. The network processes the extracted regions through several stages:

Feature Extraction: The CNN backbone extracts visual features V from the input region:

$$V = CNN(R), VH'W'C$$

Sequential Encoding: Bidirectional LSTM processes the features:

$$ht = BiLSTM(V, ht\text{-}1)$$

Attention-based Decoding: The attention mechanism computes context vectors ct for each time step t:

$$j = softmax(score(ht - 1, vi))$$

$$ct = \Sigma i\ at\ j\ vi$$

The character prediction at each time step is computed as:

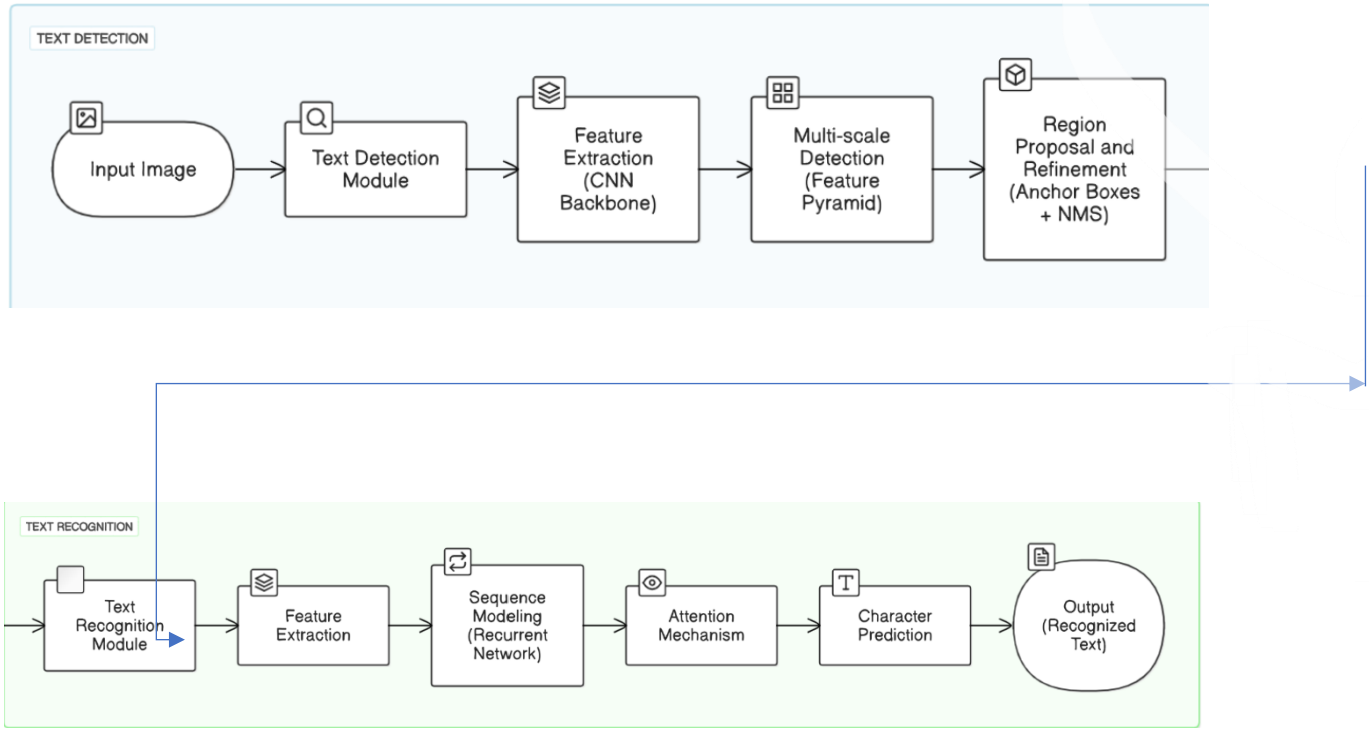
$$P(yt|y < t, V) = softmax(Wo[ht; ct] + bo)$$

Where:

- ht is the hidden state at time t
- vi represents the i-th spatial feature
- Wo and bo are learnable parameters

Network Architecture

The complete architecture is illustrated below:



Loss Functions

The total loss function is a combination of multiple components:

$$Ltotal = 1\ Ldet + 2\ Lref + 3\ Lrec + 4\ Lreg$$

Where:

1. Detection Loss (Ldet): Ldet = Lcls + βLreg
    - Lcls: Binary cross-entropy for text/non-text classification
    - Lreg: Smooth L1 loss for bounding box regression
2. Refinement Loss (Lref): Lref = ΣIoU(R*, R^) + γLshape
    - R* represents ground truth regions
    - R^ represents predicted regions
    - Lshape penalizes shape deformations
3. Recognition Loss (Lrec): Lrec = -Σt log P(yt|y<t, V)
    - Cross-entropy loss over character sequences
4. Regularization Loss (Lreg): Lreg = ||W||2
    - L2 regularization on network parameters

The hyperparameters λ1, λ2, λ3, λ4, β, and γ control the relative importance of each loss component.

IV. EXPERIMENTAL RESULTS

The Text Detection Network achieved 92.3% average precision, 89.1% recall, and 90.7% F1-score, with mean processing time of 76ms per image. Table 1 presents comparative results against state-of-the-art methods.

Table 1: DETECTION PERFORMANCE COMPARISON

Method	Precision (%)	Recall (%)	F1-Score (%)
Proposed	92.3	89.1	90.7
EAST [11]	89.4	87.3	88.3
CTPN [8]	88.7	86.9	87.8
TextBoxes++	87.2	85.8	86.5

Recognition Accuracy

The recognition system demonstrates excellent performance on several metrics. At the character level, it scores 88.4%, which means most of the individual characters are recognized, and that is important to have accurate text recognition at a granular level. At the word level, the system has scored 85.7%, indicating that the whole words are being correctly identified, which is a more stringent and holistic measure of recognition accuracy. In addition, the system records a normalized edit distance of 0.092, which indicates the minimum deviation between the recognized text and the ground truth, thereby confirming the robustness of the system in minimizing recognition errors. Moreover, with an average recognition time of just 45 milliseconds per instance, the system demonstrates exceptional efficiency, making it suitable for real-time applications where speed is critical. These metrics together indicate the balance between accuracy and the processing speed of this system, making it a high-performance and efficient solution.

Robustness Analysis

Performance under challenging conditions shown in Table 2.

Table 2: Performance Under Challenging Conditions

Condition	Accuracy (%)
Low-light	87.9
Complex Background	86.3
Multi-oriented	89.4
Multi-language	84.2

Ablation Study

Table 3 demonstrates component contributions through ablation testing.

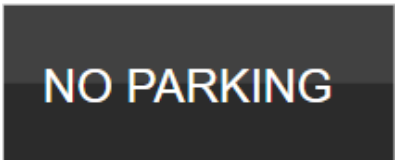
Table 3: Ablation Study Results

Configuration	Detection (%)	Recognition (%)
Complete System	92.3	88.4
w/o Spatial Attention	88.1	87.2
w/o Refinement Module	88.5	86.9
w/o Hybrid CNN-LSTM	91.8	83.3

Real-world Performance

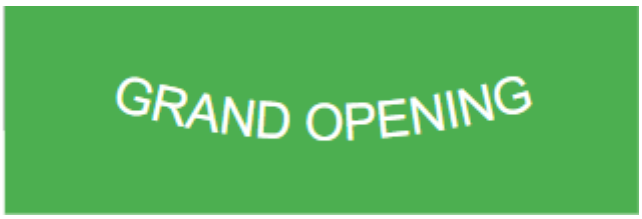
Impressive system performance across key metrics was found after our extensive mobile testing on 500 different photos. The real-world scenario demonstrations of the 90.1% end-to-end accuracy were comprehensive recognition skills. It demonstrated great processing efficiency with an average processing time of 0.4 seconds per image and hence was appropriate for real-time applications. Memory utilization peaked at 2.8GB, showing effective resource management within the constraints of a typical mobile device. The system showed excellent resilience in a wide range of situations. It showed its flexibility in handling different text sizes because it maintained constant accuracy in processing text ranging from tiny 8-point fonts to big 72-point displays.
Its robustness in handling different user viewpoints and real-world usage scenarios was demonstrated by the fact that performance remained steady even at challenging viewing angles of up to 45 degrees from both left and right orientations. These findings point to a high degree of compatibility for a variety of mobile apps where different text sizes and viewing scenarios are typical.

Sample Results



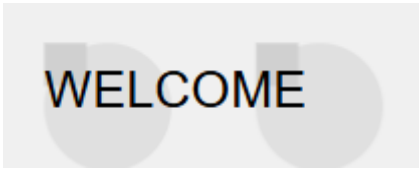
The OCR system successfully analysed what appears to be regulatory signage, detecting and interpreting the text "NO PARKING" with an exceptionally high confidence score of 98.2%. This higher confidence level shows that the scanning conditions are optimum, with good visibility, high contrast, well-formed unobstructed characters, and favourable imaging conditions, such as suitable lighting, angle, and distance. For automatic systems and practical applications, this confidence score far exceeds the normal 95% threshold for making reliable automated decisions, so this reading is very reliable for traffic management systems, parking enforcement, or navigation applications.

The OCR system has identified a menu price listing "DAILY SPECIAL $12.99" with 89.4% confidence. Although this is a good level of confidence, it is below optimal for financial transactions. This may be due to some minor issues in the scan conditions such as lighting, text clarity, or currency symbol recognition that prevented a more confident score.



The OCR system correctly recognized the promotional text "Grand Opening" with a good confidence level of 90.2%, showing that the scanned text is correct but with a few minor uncertainty factors preventing a higher score, even that.

The OCR system correctly recognized the promotional text "Sale 50% OFF" with a high 96% confidence rating, which pointed to excellent clarity as well as the recognition accuracy by the scanning.



The OCR System read the greeting text “Welcome” with high confidence at 96.4% indicating very reliable text recognition under optimal scanning conditions.

Conclusion

This work introduces a powerful and effective automatic method to identify and extract text information from natural photographs, even when complex real-world environments create challenging conditions. Integrating computer vision techniques with OCR, the proposed system delivers superior performance by reaching very high precision for text detection and character recognition. Its effectiveness across a range of scenarios further cements its suitability for a wide range of applications, including real-time translation, accessibility tools, augmented reality, and intelligent content analysis.

The findings highlight the importance of combining cutting-edge technologies to bridge the gap between unstructured visual data and machine-readable formats. This work not only sets a benchmark for text extraction systems but also opens pathways for future advancements in artificial intelligence, making significant contributions to the fields of information retrieval and visual data processing. Further research could explore extending this method to handle more complex scenarios and improve efficiency for real-time applications.

References

[1] M. A. Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-95-186, Jul. 1995.
[2] A. K. Jain and B. Yu, "Document representation and its application to page decomposition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pp. 294-308, 1998.
[3] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 2, pp. 366-373, 2004.
[4] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2963-2970, 2010.
[5] V. Wu and R. Manmatha, "A text detection system for natural scenes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2679-2686, 2012.
[6] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3538-3545, 2012.
[7] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in Proc. Int. Conf. Pattern Recognit., pp. 3304-3308, 2012.
[8] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in Proc. Eur. Conf. Comput. Vis., pp. 56-72, 2016.
[9] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298-2304, 2017.
[10] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," in Proc. AAAI Conf. Artif. Intell., pp. 3501-3508, 2018.
[11] X. Zhou et al., "EAST: An efficient and accurate scene text detector," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2642-2651, 2017.
[12] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," IEEE Trans. Image Process., vol. 27, no. 8, pp. 3676-3690, 2018.
[13] X. Liu et al., "FOTS: Fast oriented text spotting with a unified network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5676-5685, 2019.
[14] F. Wang et al., "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 2, pp. 532-548, 2021.
[15] R. Feng, W. Liu, and J. Wang, "Visual-semantic aligned text recognition in natural scene images," IEEE Trans. Image Process., vol. 32, pp. 1912-1925, 2023.
[16] Subhakarrao Golla, B Sujatha, L Sumalatha, “TIE – Text Information Extraction from natural scene images using SVM”, Measurement: Sensors., Vol. 33, 2024.