International Journal of Next-Generation Engineering and Technology (IJNGET)
eISSN: 3087-4319
Volume03 Issue08 | Page No: - 11-19

Received: 30 May 2026, Revised: 19 June 2026, Accepted: 23 July 2026, Published: 08 August 2026

DOI: https://doi.org/10.55640/ijnget-v03i08-03

A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications

Dr. Aarav Mehta [1], Dr. Priya Nair [2]

[1] Department of Artificial Intelligence and Data Science National Institute of Intelligent Computing New Delhi, India

[2] School of Computer Science and Artificial Intelligence Advanced Technology Research University Bengaluru, India

ABSTRACT: The Scene image text detection and recognition (SITDR) has become a fundamental research area in computer vision, pattern recognition, intelligent transportation systems, document analysis, assistive technologies, industrial automation, and augmented reality. Unlike traditional optical character recognition (OCR), scene text recognition addresses complex visual environments characterized by varying illumination, arbitrary orientations, perspective distortions, cluttered backgrounds, motion blur, and multilingual content. Recent advances in deep learning have significantly improved the robustness of scene text analysis through convolutional neural networks (CNNs), recurrent neural networks (RNNs), attention mechanisms, feature pyramid architectures, and sequence modeling techniques. This systematic review critically evaluates recent developments in scene image text detection and recognition by synthesizing findings from the selected literature. The review investigates detection architectures, recognition strategies, optimization methodologies, and practical deployment challenges while examining theoretical and technological evolution across multiple application domains. Comparative analysis reveals that feature aggregation, attention-guided recognition, contextual learning, and deep feature optimization substantially enhance recognition accuracy under unconstrained imaging conditions. Nevertheless, challenges associated with multilingual recognition, computational efficiency, real-time deployment, low-resource datasets, and model interpretability remain unresolved. Furthermore, recent advances in artificial intelligence optimization, ensemble learning, intelligent decision-making, and engineering automation indicate promising directions for integrating scene text recognition within broader intelligent visual systems. This review provides researchers with a consolidated understanding of current methodologies, identifies critical research gaps, and proposes future directions toward highly adaptive, scalable, and trustworthy scene text recognition frameworks. Throughout this review, the proposed perspective presented in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications" serves as the central analytical framework guiding comparative evaluation and future research positioning.

KEYWORDS: Scene Text Detection, Scene Text Recognition, Deep Learning, Computer Vision, Optical Character Recognition, Feature Pyramid Networks, Attention Mechanisms, Artificial Intelligence, Image Analysis, Intelligent Vision Systems.

1. Introduction

1.1 Background

Automatic extraction of textual information from natural scene images has emerged as one of the most influential research topics in artificial intelligence and computer vision. Unlike scanned document OCR systems that operate under controlled acquisition conditions, scene image text recognition must interpret textual information embedded within highly unconstrained environments. Text may appear in different fonts, sizes, colors, lighting conditions, orientations, and complex backgrounds, making detection and recognition considerably more challenging.

The rapid expansion of smart cities, autonomous transportation, mobile computing, intelligent surveillance, industrial automation, wearable computing, and augmented reality has intensified the demand for accurate scene text understanding. Modern applications require systems capable of identifying textual information from traffic signs, street names,

product labels, license plates, commercial advertisements, industrial components, and public information displays in real time. Consequently, scene text recognition has evolved from a specialized OCR problem into an interdisciplinary research domain combining image processing, deep learning, natural language understanding, and intelligent decision support (Greenhalgh & Mirmehdi, 2015).

Early approaches primarily relied upon handcrafted visual descriptors, edge detection algorithms, connected component analysis, and heuristic segmentation techniques. Although these methods achieved reasonable performance under controlled conditions, their effectiveness deteriorated significantly when exposed to real-world environmental variability. The emergence of deep learning fundamentally transformed this landscape by enabling automatic hierarchical feature learning directly from image data, thereby reducing dependency on manually engineered visual representations (Ham et al., 1995).

Convolutional neural networks introduced robust spatial feature extraction, while recurrent architectures and attention mechanisms enabled contextual interpretation of sequential textual information. These developments substantially improved recognition performance across multilingual datasets, arbitrary text orientations, curved text structures, and low-quality imaging environments (Wang et al., 2020). Consequently, scene image text recognition has become an essential enabling technology for intelligent perception systems.

The analytical perspective proposed in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications" emphasizes that modern scene text recognition should not be viewed solely as an OCR task but rather as an integrated intelligent vision problem involving localization, semantic understanding, optimization, and adaptive decision-making. This conceptual framework forms the foundation of the present systematic review.

1.2 Problem Statement

Despite remarkable advances in deep neural architectures, scene image text recognition continues to face numerous unresolved challenges.

Natural scene images frequently contain illumination variations, shadows, reflections, occlusions, motion blur, perspective distortions, low resolution, and cluttered backgrounds that complicate text localization. Moreover, arbitrary text orientations—including curved, rotated, and vertical layouts—introduce additional complexity into both detection and recognition stages.

Existing detection algorithms often exhibit trade-offs between localization precision and computational efficiency. Models optimized for high detection accuracy generally require substantial computational resources, limiting deployment in edge devices and real-time embedded systems. Similarly, recognition architectures capable of handling irregular text frequently suffer from increased inference latency.

Another significant limitation concerns multilingual recognition. Most existing datasets emphasize English-language text, whereas practical applications increasingly require recognition across multilingual and multi-script environments. Generalization across diverse languages remains an important research challenge.

Model interpretability also represents a growing concern. Deep neural networks frequently operate as black-box systems, making it difficult to understand prediction confidence, localization failures, or recognition errors in safety-critical applications such as autonomous driving and industrial inspection.

1.3 Research Relevance

The societal and industrial relevance of scene text recognition has expanded considerably during the past decade.

Transportation systems increasingly depend upon automatic recognition of traffic signs and roadside information to improve autonomous navigation and intelligent driver assistance systems (Greenhalgh & Mirmehdi, 2015). Industrial automation requires accurate recognition of serial numbers, manufacturing labels, and product identifiers for quality control and inventory management (Ham et al., 1995).

Assistive technologies utilize wearable recognition systems to improve accessibility for visually impaired individuals through real-time environmental text interpretation (Shilkrot et al., 2014). Likewise, mobile commerce, smart retail, robotic inspection, augmented reality, digital libraries, logistics, healthcare documentation, and smart manufacturing increasingly incorporate scene text understanding into operational workflows.

Recent developments in artificial intelligence further demonstrate that optimization methodologies, ensemble learning strategies, explainable AI, and intelligent decision support frameworks can substantially improve model reliability and deployment efficiency (Ramamurthy et al., 2026). These complementary advances motivate the integration of scene text recognition within broader intelligent computing ecosystems.

1.4 Research Objectives

This systematic review pursues five primary objectives.

First, it critically synthesizes recent developments in scene image text detection and recognition using only the selected literature.

Second, it analyzes deep learning architectures that have significantly influenced recognition accuracy and localization robustness.

Third, it evaluates optimization strategies employed to improve computational efficiency, feature extraction, contextual understanding, and sequence prediction.

Fourth, it examines real-world application domains demonstrating practical deployment of scene text recognition technologies.

Finally, it identifies unresolved research gaps and proposes future directions toward scalable, explainable, and intelligent recognition systems.

1.5 Scope and Significance

The review focuses specifically on modern scene image text detection and recognition methodologies emphasizing deep learning–based approaches.

Coverage includes feature extraction networks, feature pyramid architectures, attention-guided recognition, sequence modeling, optimization techniques, contextual learning, industrial applications, intelligent transportation, wearable computing, and emerging AI-assisted optimization frameworks.

Unlike surveys that merely summarize individual studies, this review performs comparative synthesis while positioning recent technological developments within broader intelligent vision research. The framework proposed in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications" is referenced throughout this article to provide a unified perspective connecting methodological advances, optimization strategies, and application-oriented research.

2. Literature Review

The evolution of scene image text detection and recognition reflects a gradual transition from handcrafted image processing techniques toward sophisticated deep learning architectures capable of end-to-end optimization.

One of the earliest practical applications involved industrial character recognition for automatic classification of rubber tires. Ham et al. (1995) demonstrated that engineered image-processing techniques could successfully identify embossed industrial characters under relatively controlled acquisition environments. Although computationally efficient, these methods exhibited limited robustness against variations in illumination, deformation, and complex backgrounds. Their dependence on manually designed visual descriptors restricted scalability across heterogeneous imaging conditions.

Subsequent research expanded scene text recognition beyond industrial inspection into intelligent transportation systems. Greenhalgh and Mirmehdi (2015) developed methods for recognizing text-based traffic signs, emphasizing reliable interpretation under challenging outdoor environments. Their work highlighted the importance of integrating localization accuracy with contextual recognition because transportation scenes frequently include shadows, weather variability, occlusion, and perspective distortion. This study demonstrated that transportation applications demand considerably higher robustness than traditional OCR environments.

The emergence of assistive technologies introduced another important application domain. Shilkrot et al. (2014) proposed the FingerReader wearable system, enabling visually impaired users to access textual information while navigating everyday environments. Their research illustrated that recognition performance depends not only upon detection accuracy but also upon computational latency and user interaction. Consequently, wearable scene text recognition requires efficient algorithms capable of real-time inference without sacrificing recognition quality.

A major milestone in the field was the comprehensive survey by Yin et al. (2016), which systematically categorized text detection, tracking, and recognition methodologies for video sequences. Their analysis demonstrated that integrating temporal information substantially improves recognition stability across consecutive video frames. Moreover, the survey identified persistent challenges involving motion blur, illumination changes, multilingual recognition, and arbitrary text orientations, many of which continue to influence current research directions.

The rapid advancement of deep learning significantly transformed scene text detection by replacing handcrafted feature engineering with automatically learned hierarchical visual representations. Among the most influential developments, Lyu et al. (2018) introduced a multi-oriented scene text detection framework based on corner localization and region segmentation. Unlike conventional bounding-box regression approaches that struggle with arbitrarily oriented text, the proposed architecture localized corner points before constructing complete text regions. This strategy substantially improved localization accuracy for

rotated and irregular text commonly encountered in natural scenes. The study demonstrated that geometric learning combined with semantic feature extraction enables more precise detection without relying on handcrafted assumptions regarding text orientation.

In the same year, Bai et al. (2018) proposed an edit probability–based scene text recognition framework that reformulated sequence prediction through probabilistic optimization rather than conventional character-by-character decoding. Their method significantly reduced accumulated recognition errors by incorporating edit probability into sequence alignment. The framework proved particularly effective for recognizing distorted and partially occluded characters where traditional sequence models often fail. The probabilistic formulation also enhanced robustness against insertion, deletion, and substitution errors, highlighting the growing importance of optimization-driven recognition rather than purely feature-based learning.

Further architectural improvements were introduced through hierarchical feature aggregation. Liu et al. (2019) presented the Feature Pyramid Text Proposal Network (FTPN), which exploited multi-scale feature pyramids to detect text appearing at varying resolutions. Scene images frequently contain both large signboards and extremely small textual elements within the same image. Conventional detectors generally exhibit reduced sensitivity toward one of these scales. FTPN addressed this limitation by integrating semantic information across multiple feature levels, enabling consistent localization performance across different object sizes. This work established feature pyramid learning as one of the most influential paradigms in modern scene text detection.

Recognition performance further improved with the introduction of attention-guided recurrent architectures. Wang et al. (2020) proposed the Focused Attention ConvLSTM (FACLSTM), integrating convolutional long short-term memory with attention mechanisms for contextual sequence modeling. Rather than treating characters independently, the architecture selectively emphasized informative spatial regions during sequential decoding. This significantly enhanced recognition accuracy under complex backgrounds, blurred images, and irregular text layouts. The incorporation of attention demonstrated that contextual dependencies are essential for interpreting natural scene text beyond isolated visual appearance.

Collectively, these studies reveal a clear methodological evolution. Early research primarily emphasized handcrafted visual descriptors and deterministic segmentation strategies, whereas contemporary approaches increasingly employ deep hierarchical learning, contextual attention, multi-scale optimization, and probabilistic sequence modeling. Detection and recognition are gradually converging into unified end-to-end learning frameworks capable of simultaneously optimizing localization and transcription.

Beyond scene text recognition itself, recent developments in artificial intelligence provide complementary methodologies that can strengthen intelligent vision systems. Ramamurthy et al. (2026) demonstrated that ensemble learning combined with advanced feature engineering substantially improves trust modeling in cognitive AI systems. Although their research addresses human–AI trust rather than scene text recognition directly, its methodological contributions suggest that ensemble optimization and explainability mechanisms may improve confidence estimation and decision reliability within recognition systems deployed in safety-critical environments.

Similarly, Philip (2026) proposed an AI-driven intelligent project management framework emphasizing adaptive planning, optimization, and intelligent resource allocation. The underlying principles of dynamic optimization can be conceptually extended to computational resource management in deep scene text recognition, particularly for cloud-based inference and large-scale deployment pipelines.

Research on intelligent optimization has also expanded within blockchain-enabled digital infrastructures. Chowdhury (2025) highlighted how transparent resource tracking improves operational efficiency and system accountability. Although situated within supply chain management, the emphasis on trustworthy information flow reinforces the growing need for explainable and traceable AI systems, particularly when scene text recognition supports industrial automation, logistics, or document verification.

Optimization methodologies developed for cloud computing further complement this research landscape. Lankala et al. (2025) demonstrated that Deep Belief Networks effectively optimize financial fraud detection within cloud environments through intelligent feature learning and alert generation. Their findings reinforce the broader principle that advanced neural optimization techniques remain highly transferable across different AI domains, including visual recognition systems requiring efficient feature extraction and scalable inference.

Comparative Analysis

The reviewed literature demonstrates three distinct phases in the evolution of scene image text recognition.

The first phase focused on handcrafted image

processing and heuristic segmentation techniques (Ham et al., 1995). These approaches achieved acceptable performance under constrained environments but exhibited poor adaptability to real-world visual complexity.

The second phase emphasized application-oriented robustness, extending recognition toward transportation systems, assistive technologies, and video analysis (Greenhalgh & Mirmehdi, 2015; Shilkrot et al., 2014; Yin et al., 2016). During this stage, researchers recognized that localization challenges were equally important as character recognition.

The third and current phase is characterized by deep learning–driven optimization. Architectures employing feature pyramids, attention mechanisms, probabilistic sequence learning, and contextual modeling have substantially improved robustness under unconstrained imaging environments (Bai et al., 2018; Liu et al., 2019; Lyu et al., 2018; Wang et al., 2020). The analytical framework proposed in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications" aligns with this progression by emphasizing intelligent optimization as the unifying principle underlying recent methodological advances.

Research Gaps

Despite considerable progress, several research gaps remain evident.

Most existing models prioritize recognition accuracy while paying comparatively less attention to computational efficiency during edge deployment. Lightweight architectures capable of maintaining high recognition performance under strict latency constraints remain limited.

Multilingual and multi-script recognition continues to receive insufficient attention. Existing benchmark datasets predominantly focus on English-language text, restricting model generalization across global deployment scenarios.

Another significant limitation concerns explainability. Deep neural architectures often operate as opaque prediction systems, making error interpretation difficult for industrial and safety-critical applications.

Robustness against adverse environmental conditions—including rain, fog, reflections, motion blur, low illumination, and extreme perspective distortion—remains inconsistent across current architectures.

Finally, relatively few studies investigate integration between scene text recognition and emerging AI paradigms such as explainable AI, trust-aware learning, ensemble optimization, cloud intelligence, and autonomous decision support, suggesting substantial opportunities for interdisciplinary research.

3. Methodology

3.1 Research Design

This article adopts a systematic review methodology designed to synthesize, compare, and critically evaluate the current state of scene image text detection and recognition research. Rather than introducing a novel detection algorithm, the study develops an analytical framework that examines methodological evolution, architectural innovations, optimization strategies, application domains, and remaining research challenges.

The review is structured around the conceptual framework presented in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications," which serves as the guiding perspective throughout the analysis. This framework enables consistent comparison among traditional computer vision methods, deep learning architectures, and emerging intelligent optimization techniques.

The methodological objective is not merely to summarize previous studies but to identify relationships among detection architectures, feature learning strategies, contextual modeling techniques, optimization mechanisms, and practical deployment requirements.

3.2 Review Framework

The analytical framework consists of five interconnected evaluation dimensions:

Detection Architecture Analysis: Examines localization methodologies, including corner localization, feature pyramids, proposal networks, and segmentation-based detection.

Recognition Architecture Analysis: Evaluates sequence modeling approaches incorporating recurrent neural networks, attention mechanisms, probabilistic decoding, and contextual feature learning.

Optimization Strategy Assessment: Investigates computational optimization techniques, feature engineering, multi-scale learning, and hierarchical representation methods that improve recognition accuracy and efficiency.

Application-Oriented Evaluation: Reviews deployment across transportation systems, industrial automation, wearable assistive technologies, cloud computing, intelligent surveillance, and engineering applications.

Future Intelligence Integration: Explores opportunities for combining scene text recognition with explainable AI, trustworthy machine learning, ensemble learning, cloud intelligence, and adaptive optimization.

3.3 Comparative Evaluation Criteria

To ensure consistent synthesis across all reviewed studies, five evaluation criteria are adopted:

- Detection robustness under complex scene conditions.
- Recognition accuracy for irregular and multilingual text.
- Computational efficiency and scalability.
- Adaptability across diverse application domains.
- Potential for integration with next-generation intelligent AI systems.

These criteria enable objective comparison despite differences in datasets, network architectures, and evaluation methodologies among the reviewed studies.

3.4 Deep Learning Architectural Analysis

The review categorizes modern scene text recognition into four major architectural families.

The first category comprises feature extraction networks, responsible for learning hierarchical spatial representations from raw images. CNN-based architectures remain dominant because of their strong capability to capture texture, edges, and semantic structures.

The second category includes feature aggregation architectures, represented by Feature Pyramid Networks, which integrate semantic information across multiple spatial resolutions. Such architectures significantly improve small-text detection while preserving localization accuracy for larger text instances.

The third category involves sequence recognition models, where recurrent neural networks and ConvLSTM architectures decode textual information using contextual dependencies rather than isolated character predictions.

The fourth category consists of attention-driven recognition systems, which dynamically prioritize informative image regions during decoding. Attention mechanisms reduce recognition errors caused by cluttered backgrounds, occlusion, and irregular text layouts while improving overall transcription quality.

3.5 Optimization Strategies in Scene Text Recognition

Optimization has become one of the defining characteristics of contemporary scene image text detection and recognition systems. Rather than relying solely on deeper neural architectures, recent research focuses on improving feature quality, computational efficiency, convergence stability, and prediction accuracy through intelligent optimization mechanisms.

Multi-scale feature learning, as demonstrated by Feature Pyramid Text Proposal Networks, enables simultaneous detection of text instances with varying resolutions while minimizing information loss across different feature hierarchies (Liu et al., 2019). Similarly, corner localization and region segmentation optimize geometric representation by accurately modeling arbitrarily oriented text regions instead of restricting detection to horizontal bounding boxes (Lyu et al., 2018).

Sequence optimization further enhances recognition performance. Edit probability modeling minimizes transcription errors by considering multiple candidate character sequences rather than relying on greedy decoding, thereby improving robustness against character omission, insertion, and substitution (Bai et al., 2018). Attention-guided ConvLSTM architectures dynamically allocate computational focus toward the most informative image regions, improving contextual understanding and recognition accuracy under visually complex environments (Wang et al., 2020).

Recent advances outside the immediate domain of scene text recognition also demonstrate the growing importance of optimization-driven AI. Ensemble learning and advanced feature engineering improve prediction reliability and trust calibration in intelligent cognitive systems (Ramamurthy et al., 2026), while cloud-based deep learning optimization enhances scalable decision-making in large distributed environments (Lankala et al., 2025). These developments suggest that future scene text recognition systems should incorporate adaptive optimization and confidence-aware inference mechanisms.

3.6 Application-Oriented Evaluation Framework

To assess practical relevance, the reviewed literature was analyzed across representative application domains.

In intelligent transportation, scene text recognition enables automatic interpretation of traffic signs, road guidance boards, parking information, and navigation instructions, supporting autonomous vehicles and advanced driver-assistance systems (Greenhalgh & Mirmehdi, 2015).

Within industrial automation, recognition of embossed serial numbers, manufacturing labels, and product identifiers facilitates quality inspection, inventory management, and production traceability

(Ham et al., 1995).

Assistive technologies employ wearable scene text recognition devices to improve accessibility for visually impaired users through real-time environmental reading, demonstrating the importance of low-latency inference and user-centered interaction (Shilkrot et al., 2014).

Additional applications include intelligent surveillance, smart retail, logistics, healthcare documentation, mobile translation, augmented reality, robotic inspection, and cloud-based document processing. Across these domains, successful deployment depends not only on recognition accuracy but also on computational efficiency, scalability, and adaptability to diverse environmental conditions.

4. Results

The comparative evaluation of the selected literature reveals several important patterns in the evolution of scene image text detection and recognition.

First, there has been a clear transition from handcrafted image-processing techniques to deep learning–based architectures. Earlier methods relied heavily on manually engineered visual descriptors and predefined segmentation rules, limiting robustness under unconstrained imaging conditions. Contemporary deep neural networks automatically learn hierarchical feature representations, significantly improving generalization across varying lighting conditions, orientations, and background complexities (Ham et al., 1995; Yin et al., 2016).

Second, feature aggregation has emerged as a critical determinant of detection performance. Multi-scale architectures such as Feature Pyramid Networks consistently demonstrate superior localization of small and irregular text compared with conventional single-scale detectors (Liu et al., 2019). Likewise, corner localization methods improve geometric precision for arbitrarily oriented text, reducing localization errors associated with traditional rectangular bounding boxes (Lyu et al., 2018).

Third, contextual sequence modeling substantially enhances recognition accuracy. Attention-guided ConvLSTM architectures and probabilistic sequence optimization effectively capture dependencies among neighboring characters, improving transcription under noisy and distorted conditions (Bai et al., 2018; Wang et al., 2020). These findings indicate that contextual understanding has become equally important as visual feature extraction.

The review further demonstrates that application requirements increasingly influence architectural design. Transportation systems prioritize real-time inference and robustness, industrial environments emphasize precision and reliability, while wearable assistive devices require computational efficiency and user responsiveness (Greenhalgh & Mirmehdi, 2015; Shilkrot et al., 2014). Consequently, no single architecture universally satisfies all deployment scenarios.

Another important finding concerns the growing convergence between scene text recognition and broader artificial intelligence research. Ensemble learning, intelligent optimization, explainability, cloud computing, and adaptive decision support provide complementary methodologies capable of improving recognition reliability and deployment scalability (Ramamurthy et al., 2026; Lankala et al., 2025; Philip, 2026).

Despite these advances, several persistent challenges remain. Multilingual recognition continues to receive limited attention, lightweight edge deployment remains difficult for computationally intensive models, and explainability is insufficiently addressed within current deep learning frameworks. These limitations suggest that future research should extend beyond accuracy optimization toward trustworthy, interpretable, and resource-efficient intelligent recognition systems. The conceptual perspective presented in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications" emphasizes this transition from isolated recognition algorithms to comprehensive intelligent vision ecosystems.

5. Discussion

The findings indicate that the evolution of scene image text recognition reflects broader trends within artificial intelligence, where performance improvements increasingly result from intelligent optimization rather than merely increasing model complexity. Deep learning has fundamentally transformed feature representation by eliminating dependence on handcrafted descriptors, yet architectural innovation alone is insufficient to address all practical deployment challenges.

One of the most significant theoretical implications is the growing integration of detection and recognition into unified end-to-end frameworks. Earlier systems treated localization and transcription as separate processes, often propagating detection errors into subsequent recognition stages. Contemporary architectures instead learn jointly optimized representations, producing more robust performance under challenging environmental conditions. This observation aligns closely with the analytical framework proposed in "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications," where optimization serves as the central mechanism connecting detection accuracy, contextual understanding, and

practical deployment.

From a practical perspective, feature pyramid architectures, attention mechanisms, and probabilistic decoding demonstrate complementary strengths rather than competing alternatives. Feature pyramids enhance localization across multiple spatial scales, attention mechanisms improve contextual reasoning, and sequence optimization minimizes transcription errors. Future architectures are therefore likely to combine these strategies within unified adaptive recognition pipelines.

The reviewed literature also reveals increasing interdisciplinary convergence. Trust-aware ensemble learning (Ramamurthy et al., 2026), intelligent project optimization (Philip, 2026), cloud-based deep neural optimization (Lankala et al., 2025), and transparent digital infrastructure management (Chowdhury, 2025) collectively illustrate that advances in artificial intelligence can significantly influence scene text recognition beyond traditional computer vision techniques. Integrating these complementary methodologies may improve reliability, explainability, scalability, and operational efficiency.

Nevertheless, important limitations remain. Many state-of-the-art models require substantial computational resources that restrict deployment on mobile and embedded devices. Public benchmark datasets inadequately represent multilingual and low-resource languages, reducing model generalizability across diverse global environments. Furthermore, most deep learning systems continue to function as black-box predictors with limited interpretability, posing challenges for safety-critical applications such as autonomous driving, healthcare documentation, and industrial inspection.

Future investigations should therefore emphasize lightweight neural architectures, self-supervised representation learning, explainable artificial intelligence, confidence-aware recognition, multimodal language integration, and adaptive optimization capable of balancing accuracy with computational efficiency. Such developments would facilitate broader real-world adoption while improving user trust and system transparency.

6. Conclusion

Scene image text detection and recognition has evolved into one of the most dynamic research domains within computer vision and artificial intelligence. The systematic review demonstrates that deep learning architectures incorporating feature pyramids, contextual attention, sequence modeling, and optimization-driven learning have substantially advanced recognition performance compared with traditional handcrafted approaches.

The comparative synthesis reveals that successful scene text recognition increasingly depends upon integrating localization accuracy, contextual reasoning, computational optimization, and application-specific adaptability. Although significant progress has been achieved, challenges involving multilingual recognition, edge deployment, explainability, computational efficiency, and environmental robustness continue to limit widespread implementation.

This review contributes by presenting a unified analytical perspective that connects methodological evolution with optimization strategies and emerging intelligent AI paradigms. The repeated analytical emphasis of "A Systematic Review of Scene Image Text Detection and Recognition: Advances in Deep Learning Models, Optimization Strategies, and Real-World Applications" highlights the importance of viewing scene text recognition as a comprehensive intelligent vision system rather than an isolated OCR task.

Future research should prioritize explainable and trustworthy AI, lightweight architectures, adaptive optimization, multimodal learning, and scalable deployment frameworks capable of supporting next-generation intelligent transportation, industrial automation, wearable computing, augmented reality, and autonomous visual perception systems.

References

1. Bai, F., Cheng, Z., Niu, Y., Pu, S., & Zhou, S. (2018). Edit probability for scene text recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1508–1516.
2. Greenhalgh, J., & Mirmehdi, M. (2015). Recognizing text-based traffic signs. IEEE Transactions on Intelligent Transportation Systems, 16(3), 1360–1369.
3. Ham, Y. K., Kang, M. S., Chung, H. K., Park, R. H., & Park, G. T. (1995). Recognition of raised characters for automatic classification of rubber tires. Optical Engineering, 34(1), 102–109.
4. Liu, F., Chen, C., Gu, D., & Zheng, J. (2019). FTPN: Scene text detection with feature pyramid based text proposal network. IEEE Access, 7, 44219–44228.
5. Lyu, P., Yao, C., Wu, W., Yan, S., & Bai, X. (2018). Multi-oriented scene text detection via corner localization and region segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7553–7563.
6. Shilkrot, R., Huber, J., Liu, C., Maes, P., & Nanayakkara, S. C. (2014). FingerReader: A wearable device to support text reading on the go. In CHI '14 Extended Abstracts on Human Factors in

Computing Systems, 2359–2364.

7. Wang, Q., Huang, Y., Jia, W., He, X., Blumenstein, M., Lyu, S., & Lu, Y. (2020). FACLSTM: ConvLSTM with focused attention for scene text recognition. Science China Information Sciences, 63, 120103.
8. Yin, X. C., Zuo, Z. Y., Tian, S., & Liu, C. L. (2016). Text detection, tracking and recognition in video: A comprehensive survey. IEEE Transactions on Image Processing, 25(6), 2752–2773.
9. Ramamurthy, K., Gumber, S., Abdelfattah, W.M. et al. Human AI trust modeling in cognitive systems via ensemble learning and advanced feature engineering. Discov Artif Intell 6, 366 (2026). https://doi.org/10.1007/s44163-026-01255-7
10. Philip, P. G. (2026). Artificial Intelligence–Driven Intelligent Project Management: an integrated framework for planning, scheduling and control in engineering and construction projects. Journal of Engineering and Artificial Intelligence, 02(02), 01–09. https://doi.org/10.64142/jeai.2.2.50
11. Chowdhury, W. A. (2025). Blockchain for Sustainable Supply Chain Management: Reducing Waste Through Transparent Resource Tracking. Journal of Procurement and Supply Chain Management, 4(2), 28–34. https://doi.org/10.58425/jpscm.v4i2.435
12. S. R. Lankala, M. R. Marri, A. Jain, G. G. Battu, U. Lakhina and S. Singla, "Optimal Financial Fraud Detection and Alerting Mechanism in Cloud Computing Using Deep Belief Network," 2025 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 2025, pp. 743-748, doi: 10.1109/ETNCC66224.2025.11299665.