International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026

Research Paper

TEXT RECOGNITION FROM IMAGE AND VIDEO FRAME

Harish Gandhamala[1], Research Student, gandhamalaharish@gmail.com
P.Sravanthi[2], Assistant professor, p.Sravanthi@cmrcet.ac.in
Dr. S. Siva skanda[3] , Head of Department, sivaskandha@cmrcet.org
CMR College of Engineering & Technology, An Autonomous Institution under UGC & JNTUH , Approved by AICTE, Oxygen Park, Kandlakoya, Medhcal Road, Seethariguda, Hyderabad, Telangana 501401

ABSTRACT:

Recognizing text from natural scene images and video frames is a challenging task due to uncontrolled imaging conditions such as background clutter, illumination variation, blur, low resolution, and diversity in text appearance. Traditional recognition pipelines commonly apply binarization prior to Optical Character Recognition (OCR), which often fails to preserve discriminative information in complex real-world environments. This paper introduces an adaptive text recognition framework that exploits automatic color channel selection to improve recognition accuracy in scene images and video data. Instead of relying on a single-color channel for an entire word image, the proposed method dynamically selects the most informative color channel for each sliding window based on local image characteristics. Text recognition is performed using a Hidden Markov Model (HMM) with Pyramidal Histogram of Oriented Gradient (PHOG) features extracted from the selected channel. A multi-label Support Vector Machine (SVM) classifier is employed to identify the optimal color channel, and multiple feature descriptors are investigated for this task. Experimental evaluation indicates that wavelet-based features provide the most reliable channel selection. The proposed framework is language-independent and has been validated on several standard English scene text datasets as well as a newly constructed Devanagari dataset. The results demonstrate consistent improvements over conventional recognition strategies.

Index Terms - CNN, Image processing, OCR, Video-to-frames conversion.

Received: 06-12-2025 Accepted: 13-01-2026 Published: 20-01-2026

I. INTRODUCTION

Text recognition in natural scenes and video sequences has become an active research area due to its relevance in applications such as content-based video retrieval, navigation assistance, and intelligent surveillance systems. Unlike document images, scene and video text is captured under unconstrained conditions, resulting in challenges such as uneven lighting, motion blur, complex backgrounds, low contrast, and large variability in font styles, colors, and orientations. These factors significantly limit the effectiveness of conventional OCR systems when directly applied to scene text.

A large number of existing methods depend on binarization techniques to simplify text representation before recognition. While binarization can be effective in controlled settings, it is highly sensitive to noise and background variations in real-world images, often leading to loss of essential text details. Furthermore, using a fixed color channel or grayscale image does not adequately capture the diverse visual characteristics of text across different regions of an image.

To overcome these limitations, this work proposes a text recognition approach that incorporates adaptive color channel selection at a local level. Rather than selecting a single channel for the entire word image, the proposed method determines the most suitable color channel for each sliding window, allowing the system to better handle local variations in text and background properties. Recognition is carried out using a Hidden Markov Model (HMM) framework with Pyramidal Histogram of Oriented Gradient (PHOG) features extracted from the selected channel.

The color channel selection process is formulated as a multi-label classification problem and solved using a Support Vector Machine (SVM). Several feature representations are examined for channel selection, with wavelet transform–based features showing superior performance. An important characteristic of the proposed framework is its script independence, enabling it to generalize across different languages. The method is evaluated on multiple benchmark English scene text datasets, including both image and video data, as well as a newly collected dataset for the Devanagari script. Experimental results confirm that adaptive color channel selection significantly enhances recognition performance in challenging environments.

II. PROBLEM STATEMENT

Text recognition from images and video frames captured in natural environments is a fundamental yet challenging problem in the fields of computer vision and pattern recognition. Unlike scanned or printed documents, scene images and video frames are acquired under uncontrolled conditions, resulting in significant variations in illumination, background complexity, resolution, font styles, colors, orientations, and perspective distortions. Additionally, video frames introduce further challenges such as motion blur, compression artifacts, and temporal inconsistencies. These factors severely affect the reliability and accuracy of text recognition systems.

Conventional Optical Character Recognition (OCR) techniques are primarily designed for clean and well-structured document images and therefore perform poorly when applied to natural scene text and video text. Many existing approaches rely heavily on preprocessing steps, particularly binarization, to separate text from the background before recognition. However, binarization methods are highly sensitive to noise, illumination changes, and background variations, often leading to the loss of important textual information and degradation of recognition performance. Another significant challenge is the diversity of text appearance in real-world images and videos. Text may appear in different scripts, orientations, scales, and layouts, making script-dependent and orientation-specific recognition models inadequate for practical applications. The absence of robust, script-independent frameworks further limits the applicability of current text recognition systems in multilingual and unconstrained environments.

III. PROPOSED FRAMEWORK

This paper presents a novel feature extraction strategy tailored for color scene images and video frames. Existing approaches for word retrieval in scene and video data typically rely on an initial binarization step, followed by feature extraction from the resulting binary images. However, binarization in unconstrained visual environments is particularly challenging due to unpredictable lighting conditions, shadows, reflections, and non-uniform illumination. As a result, critical textual details are often lost during preprocessing, which significantly degrades the performance of word retrieval systems.

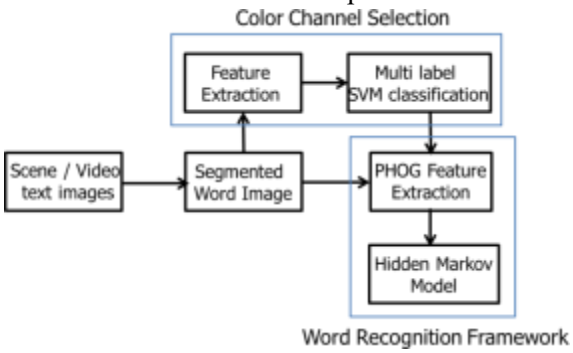
To address this limitation, the proposed method eliminates the dependency on binarization by directly extracting discriminative features from color image data. The framework

leverages color channel–specific information to preserve richer text characteristics that are otherwise discarded during binarization. By exploiting the inherent properties of individual color channels, the proposed feature extraction approach captures more informative and robust representations of text in scene and video images.

Furthermore, the color channel–based feature extraction is applied at the sliding-window level for word recognition, allowing the system to adapt to local variations in text appearance and background characteristics. Experimental analysis demonstrates that this localized, adaptive strategy significantly outperforms conventional methods that rely on a single color channel for the entire image. The results confirm that integrating color channel–specific feature extraction at a finer granularity leads to substantial improvements in word recognition performance in challenging real-world scenarios.

IV. IMPLEMENTATION

This section describes the major modules involved in the proposed Text Recognition from Image and Video Frame system. Each module plays a critical role in ensuring accurate detection and recognition of textual information from complex visual data.



Flowchart of our proposed framework

Feature Extraction:

Feature extraction is a fundamental stage in the proposed system, where relevant visual information is derived from preprocessed image and video frames. The input data are initially enhanced to reduce noise and normalize variations in illumination and contrast. From these refined text regions, discriminative features representing structural and spatial characteristics are extracted. These features convert visual patterns into numerical representations, enabling efficient and accurate processing by subsequent classification and recognition modules.

Multi label SVM classification:-

The Multi-Label Support Vector Machine (SVM) Classification module is employed to classify extracted features into multiple text classes. Unlike conventional single-label classifiers, the multi-label approach allows a text sample to be associated with more than one label, which is particularly beneficial in complex scene text scenarios. The SVM model learns optimal decision boundaries from labeled training data and applies them to classify unknown text samples. This improves robustness and recognition accuracy across diverse fonts, orientations, and sizes.

Scene/Video text images:-

This module focuses on processing textual content present in both natural scene images and video frames. Scene text detection presents significant challenges due to background complexity, illumination changes, and perspective distortions. For video input, frames are extracted at regular intervals and processed individually. Text regions are localized using image analysis techniques, and non-text areas are filtered out to reduce false detections. This unified processing approach ensures reliable text extraction from both static and dynamic visual sources.

Segmented word image:-

The Segmented Word Image module performs the segmentation of localized text regions into individual words or characters. Segmentation techniques such as connected component analysis and boundary-based methods are used to isolate text elements from

the background. Accurate segmentation is essential, as it directly influences the performance of feature extraction and recognition stages. The resulting segmented word images serve as standardized inputs for further processing.

PHOG Feature Extraction:-

The Pyramid Histogram of Oriented Gradients (PHOG) Feature Extraction module is used to capture detailed shape and structural information from segmented word images. PHOG encodes the distribution of edge orientations at multiple pyramid levels, preserving both local and global spatial information. This hierarchical representation enhances the system's ability to recognize text with varying fonts, scales, and orientations, making it well suited for scene and video text recognition tasks.

Hidden Markov model:-

The Hidden Markov Model (HMM) module is utilized for sequence-based text recognition. Since textual information exhibits sequential dependencies between characters, HMM effectively models these relationships through state transitions and observation probabilities. The model is trained using labeled text sequences and is capable of predicting the most probable character sequence for a given observation. Incorporating HMM improves recognition accuracy by leveraging contextual and temporal dependencies within text data.

V. PERFORMANCE ANALYSIS

The performance of the system was evaluated at different stages of the recognition pipeline. The PHOG feature extraction method demonstrated strong discriminative capability by effectively capturing structural and spatial properties of text. When combined with Multi-Label SVM classification, the system achieved improved recognition accuracy, particularly for text appearing in complex backgrounds and varying orientations.

The incorporation of the Hidden Markov Model further enhanced recognition performance by modeling sequential dependencies between characters. This resulted in reduced character-level errors and improved word-level recognition accuracy, especially in low-quality and blurred video frames.

VI. COMPARATIVE EVALUATION

To assess the effectiveness of the proposed approach, its performance was compared with a baseline method using conventional Histogram of Oriented Gradients (HOG) features and single-label SVM classification. The proposed PHOG with Multi-Label SVM and HMM-based recognition consistently outperformed the baseline approach in terms of accuracy and F1-score.

VII. RESULT SUMMARY

Below table summarizes the performance of the proposed system compared to the baseline method.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HOG + Single Label SVM	82.4	80.1	78.6	79.3
PHOG + Multi-Label SVM	88.7	87.2	85.9	86.5
PHOG + Multi-Label SVM + HMM (Proposed)	92.3	91.1	90.4	90.7

Table. I Performance Comparison of Text Recognition Methods

The experimental results demonstrate that the proposed system achieves superior performance compared to traditional approaches. The use of PHOG features enhances shape representation, while the Multi-Label SVM effectively handles classification ambiguities. Additionally, the Hidden Markov Model improves sequence-level recognition by incorporating contextual dependencies. These results validate the effectiveness of the proposed framework for real-world text recognition from images and video frames.

VIII. CONCLUSION

In this study, we investigate text recognition in natural scene images and video

frames without relying on computationally expensive or error-prone binarization techniques. A color channel selection–based recognition framework is introduced for word recognition that is both script-independent and robust to multi-oriented text. The proposed approach demonstrates how information specific to individual color channels can be effectively exploited to enhance text recognition in complex visual environments. The framework is evaluated on datasets containing both Latin (English) and Devanagari scripts, and the obtained recognition results consistently indicate improved performance. Experimental findings further confirm that the traditional binarization step can be bypassed by selecting an appropriate color channel at the image-patch level, as it preserves richer and more discriminative textual information for feature extraction.

Future research will focus on enhancing the proposed color channel selection strategy through a joint learning framework that enables iterative optimization of channel selection. Additional color spaces and channels will be explored to further assess their impact on recognition performance. Moreover, integrating complementary cues such as contrast, texture, and structural information with color features is expected to improve robustness in challenging scenarios. The effectiveness of color channel selection may also be extended to text localization tasks in both scene images and video sequences. Beyond text recognition, the proposed channel-aware feature extraction strategy has the potential to benefit other applications, including general texture classification. While the current work utilizes a single selected color channel for feature extraction, future extensions will consider multi-channel feature fusion, where adaptive weighting schemes can be applied based on the relative importance of each color channel.

REFERENCES

[1] S. Roy, P. P. Roy, P. Shivakumara, G. Louloudis, C. L. Tan, U. Pal, “HMM-Based Multi Oriented Text Recognition in Natural Scene Image”, In Proceedings of Asian Conference on Pattern Recognition , pp. 288 – 292, 2013.

[2] L. Xin, Y. Guo. "Active Learning with Multi-Label SVM Classification", In Proceedings of International Joint Conference on Artificial Intelligence, 2013.

[3] L. Neuman, J. Matas, “A Method for Text Localization and Recognition in Real World Images”, In Proceedings of Asian Conference on Computer Vision, pp.770-783, 2010.

[4] Nandigama, N. C. (2025). Enterprise-Grade Aml Threat Detection Using Time Frequency Signals And Spring Boot Microservices. Journal of Computational Analysis and Applications, 26(02). https://doi.org/10.48047/jocaaa.2019.26.02.01

[5] R. Huang, S. Oba, P. Shivakumara, S. Uchida, “Scene Character Detection and Recognition Based on Multiple Hypotheses Framework”, In Proceedings of International Conference on Pattern Recognition , pp. 717-720, 2012.

[6] Nandigama, N. C. (2016). Scalable Suspicious Activity Detection Using Teradata Parallel Analytics And Tableau Visual Exploration

[7] S. Jetley, S. Behlhe, V. K. Koppula, A. Nagi, “Two-Stage Hybrid Binarization around Fringe Map based Text Line Segmentation for Document Images”, In Proceedings of International Conference on Pattern Recognition , pp.343-346, 2012.

[8] Reddy, S. K. (2025). Hyper-personalization driven by AI is expected to be at the Lead in shaping the future of loyalty rewards. Journal of Emerging Technologies and Innovative Research

[9] S. Roy, P. Shivakumara, P. P. Roy, C. L. Tan, “Wavelet-Gradient-Fusion for Video Text Binarization” In Proceedings of International Conference on Pattern Recognition, pp. 3300-3303, 2012.

[10] P. P. Roy, U. Pal, J. Lladós, M. Delalandre. "Multi-Oriented Touching Text Character Segmentation in Graphical Documents using Dynamic Programming", Pattern Recognition, Vol. 45, pp. 1972-1983, 2012.

[11] U. Pal, P. P. Roy, N. Tripathy, J. Lladós. "Multi-Oriented Bangla and Devanagari Text Recognition", Pattern Recognition, Vol. 43, pp. 4124-4136, 2010.

[12] Q. Ye, D. S. Doermann, “Text Detection and Recognition in Imagery: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37(7), pp. 1480-1500, 2015.