International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

A Review on Text Detection and Recognition in Images and Videos Using Machine Learning and Deep Learning Techniques

M. B. Gohil
Department of Computer Science
Veer Narmad South Gujarat University
Surat, India

Dr. A. A. Desai
Department of Computer Science
Veer Narmad south Gujarat University
Surat, India

Abstract - Reliable text detection and recognition systems are needed in order to enable efficient indexing, searching, and semantic retrieval due to the increasing volume of digital video and image content. For extracting and recognising textual information from video frames and images, considerable research has focused on Optical Character Recognition (OCR) techniques since long. The aim of this review is to study the evolution of various methods for text detection and recognition, tracing the transition from traditional feature-driven approaches, such as the Haar Discrete Wavelet Transform (HDWT), to modern machine learning based architectures. The objective is to support ongoing research in regional language text detection and recognition.

Text analysis of the Gujarati script presents unique challenges, including its multi-zonal character structure and the absence of a shirorekha (header line), which distinguish it from scripts such as Devanagari. Additional complications emerge from elaborate character shapes, conjunct forms, diacritical markings, and considerable structural variety. Early study demonstrated the efficacy of wavelet-based methods for text localization in videos and images, setting the stage for later studies. Recently, sophisticated models employing Convolutional Recurrent Neural Networks (CRNN) and Long Short-Term Memory (LSTM) networks have achieved state-of-the-art performance, demonstrating enhanced resilience to intricate backgrounds, fluctuating lighting, and the cursive features of Gujarati text in dynamic visual environments. This study aims to advance the field of regional language text detection and recognition in images and videos.

Keywords: Deep Learning, Machine Learning, CNN, CRNN, LSTM, Transfer learning

I. INTRODUCTION

The quick expansion of online repositories holding videos and images has heightened the necessity for effective text detection and recognition methodologies to provide content-based indexing and retrieval and many more. Text within images and videos represents a significant source of information, manifesting either as explicitly embedded text, including captions and overlays, or as natural scene text obtained from real-world contexts. Recognising this information from video frames is especially difficult because of low spatial resolution, motion blur, intricate backgrounds, and fluctuating lighting conditions. This review specifically addresses regional languages, with a focus on Gujarati.

These problems become even bigger when working with local Indian scripts like Gujarati. The Indo-Aryan script Gujarati, which is derived from Devanagari, has three separate zones: upper, middle, and lower. Gujarati does not have a shirorekha (header line), which is normally employed to make text line and word segmentation easier, unlike Devanagari and Hindi [17]. Character boundaries are more difficult to identify and recognize in all three zones due to the presence of modifiers and diacritical marks. As a result, text identification and recognition in Gujarati images and videos are much more difficult than in scripts like Latin, necessitating the use of specific, reliable OCR techniques.

II. LITERATURE REVIEW

Early work on Gujarati OCR relied heavily on traditional, rule-based techniques that required manual intervention and were prone to high error rates. Initially, Gujarati OCR has received limited research attention [1, 3, 5, 6, 7, 8, 9, 10, 11, and 12]. According literature, the first ever work found on Gujarati Character Recognition was reported in 1999 by Antani and Agnihotri [1].

In early years, there were good reviews given by Dholakia et al. [2] in 2009. It was about analyzing Gujarati documents and OCR. They made a review of various feature extraction technologies such as wavelets, DCT and fringe maps.

In 2011, Maloo and Kale [3] also gave a review of Gujarati script recognition, handwritten character recognition steps and gave several feature extraction approaches used for Indian scripts.

In 2005, Dholkia et al. [4] presented accurate zone detection in images of printed Gujarati using a smearing algorithm and the slope of a line information.

In 2013, Patel C. N. and Desai A. A. [9] suggested a hybrid method for handwritten Gujarati OCR which use hybrid feature set derived using structural features and moments of the character. The Binary tree-classifier and k-NN are used for character identification purposes.

Over the last two decades, however, the integration of Artificial Intelligence has transformed Gujarati character recognition, marking a clear shift from conventional approaches to advanced deep learning models. The introduction of Machine Learning and Deep Learning techniques led to substantial improvements in performance. Researchers experimented with diverse feature extraction methods and robust classifiers to boost accuracy, with architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) becoming widely adopted for effective recognition.

In 2020, Ali Mirza et al. [13] proposed a framework that first detects text regions in video frames using fine-tuned CNN-based object detectors (Faster R-CNN, SSD, and EAST) to localize candidate text areas. Detected text lines are then classified by script using a CNN-based script identification module to separate English and Urdu text. Finally, recognized text is obtained via Google Tesseract OCR for English and a custom CNN–BiLSTM model (UrduNet) with CTC decoding for Urdu, enabling end-to-end multi-script text recognition.

In 2017, Aneeshan Sain et al. [14] used a method that detects candidate text regions using Fourier–Laplacian filtering and maximum-difference mapping, followed by skeletonization to handle multi-oriented and curved text structures. Skeletal analysis separates branches, fits polynomial curves, and reconstructs text regions by estimating text width and shape while eliminating non-text artifacts. Finally, PHOG features extracted along fitted curves are verified using an HMM-based classifier to remove false positives and confirm valid text regions.

In 2017, Ayan Kumar Bhunia et al. [15] created a framework that takes segmented word images from scene/video text and applies a novel color channel–specific feature extraction approach to avoid information loss caused by binarization under varying illumination. Multiple texture and image-property features are extracted from individual color channels, and an automatic multi-label SVM is used to select the most informative channel at the sliding-window level. Selected channel features are fed into a sliding-window PHOG descriptor and recognized using an HMM-based word recognition framework for robust scene and video text recognition.

In 2018, Dhara S. Joshi and Y. R. [16] proposed a system that acquires handwritten character images via offline (scanner/camera) or online (digital pen) methods and applies pre-processing steps including grayscale conversion, skew correction, noise filtering, and morphological operations to enhance image quality. Feature extraction is performed using thinning and skeletonization to obtain discriminative character characteristics for effective recognition. Classification is performed using the K-Nearest Neighbour (K-NN) algorithm, which assigns an input sample to the class of its closest neighbours in the feature space.

In 2018, J. M. Patel and A.A. Desai [17] proposed a system that first preprocesses input images by extracting luminance, reducing noise, and applying Haar Discrete Wavelet Transform (HDWT) to highlight potential Gujarati text features. Candidate text edges are detected using Sobel operators applied to wavelet detail components, followed by morphological dilation and connected-component analysis with geometric filtering to localize true text regions. Finally, the detected text regions are extracted, segmented, and binarized using Otsu thresholding with polarity adjustment to clearly separate text from the background.

In 2020, Jyoti Pareek et al. [18] proposed a handwritten character recognition system that first pre-processes input images through scanning, resizing, noise removal, binarization, and skew correction to standardize and clean the data. Pre-processed images are then segmented into lines, words, and individual characters using connected-component labelling, histogram projection profiles, text-blob detection, and contour-based methods. Finally, the segmented characters are classified using deep learning models, specifically MLP and CNN architectures, with data augmentation to improve recognition accuracy.

In 2019, K. S. Raghunandan et al. [19] used a method that detects text by applying bit-plane slicing on gray images, selecting the most informative plane using Iterative Nearest Neighbour Symmetry (INNS), and identifying representative text components through Mutual Nearest Neighbour Pair (MNNP) analysis based on gradient directions. Missing characters are restored using edge-based outward gradient propagation, enabling robust multi-oriented text detection under noise, blur, and low resolution. For recognition, an automatic window selection mechanism combined with contour let-wavelet-based statistical, texture, and run-length features is employed, and text is finally recognized using Hidden Markov Models (HMM).

In 2017, Kiran Agre et al. [20] developed a system that first converts input videos into frames at regular or user-selected time intervals to avoid redundant text extraction. Text regions are detected in each frame using the MSER algorithm, refined using stroke width analysis, merged into words, and recognized using OCR techniques. The recognized text from all frames is sequentially stored in a text file, enabling compact storage and efficient retrieval of video content.

In 2022, Krishn Limbachiya et al. [21] proposed an approach that uses pre-trained convolutional neural networks (VGGNet, InceptionV3, DenseNet, NASNet, and MobileNet) with transfer learning to recognize Gujarati handwritten characters. A shallow classifier with 512 hidden neurons and a softmax output layer for 54 classes is added on top of the pre-trained CNNs, with dropout applied to prevent overfitting. The models are fine-tuned on the Gujarati dataset to extract hierarchical features and achieve accurate classification with reduced training data and computational effort.

In 2024, L. Rasikannan et al. [22] proposed an approach that first pre-processes video frames through segmentation, grayscale conversion, noise reduction, and contrast enhancement to improve text visibility. Text regions are then detected using OCR with thresholding, contiguous object detection, and morphological operations to localize candidate text areas. Finally, the extracted text regions are recognized using a pre-trained CRNN model, which analyses features and spatial relationships to convert visual text into digital text, and outputs integrated results across all frames.

In 2017, Nidhin Raju and Dr. Anita H. B. [23] collected news videos from YouTube and converted them into frames at 3-second intervals, pre-processed using grayscale conversion and Otsu's thresholding, and divided them into fixed-size blocks for analysis. Thirteen global and local features are extracted from binary and grayscale images using statistical measures along with DCT and FFT transformations. The extracted features are classified using machine learning classifiers such as Simple Logistic, Random Forest, and J48 to evaluate and compare text recognition performance.

In 2016, S.Nithyadheviet et al. [24] developed a system that preprocesses input images or video frames by converting to grayscale, removing noise, and binarizing, followed by text detection using MSER, wavelet subband analysis, and clustering to extract candidate text regions. Text recognition is performed using character structure descriptors and SVM-based classification to accurately recognize multilingual and multi-font text. The recognized text is converted into speech using a text-to-speech module, enabling effective multimedia content understanding and retrieval.

In 2017, Shu Tian et al. [25] used a method that first performs multi-orientation scene text detection and recognition on individual video frames using multi-channel and multi-scale learning to generate robust text candidates. Multiple tracking strategies—tracking-by-detection, spatio-temporal context learning, and template matching—are jointly applied to predict text positions across consecutive frames. All detection and tracking results are integrated into a weighted graph and optimized globally using dynamic programming to obtain the most reliable text trajectories and final text locations.

In 2016, Too Kipyego Boazal and Prabhakar C. J. [26] Used an approach that first estimates stereo disparity from rectified video frames and detects candidate planar surfaces using gradient analysis, followed by plane fitting with PCA and RANSAC to model text-supporting regions. Planar and non-planar regions are segmented using Markov Random Field labelling with graph cuts, reducing background complexity by isolating planar text blocks. Text regions are extracted from the planar surfaces using Fourier–Laplacian filtering, maximum gradient difference analysis, and k-means clustering to accurately separate text from non-text areas.

In 2016, V.N. Manjunath and M. S. Pavithra [27] used a method that first extracts texture and edge features using a single-level 2D Discrete Wavelet Transform and Gabor filtering, followed by k-means clustering to separate background, foreground, and true text pixels. Morphological operations and a linked-list–based grouping strategy are applied to form connected components and generate candidate text line sequences. Finally, wavelet entropy is computed for each component sequence to accurately identify true multilingual text regions while eliminating false positives.

In 2016, Vijeta Khare et al. [28] proposed a methodology that automatically estimates the window size using stroke width and gradient information from Sobel and Canny edge images to identify candidate text pixels. Temporal moments and an iterative k-means clustering approach are employed to separate static caption text from dynamic scene text and background using frame-to-frame deviations. False positives are reduced using gradient-direction analysis, and boundary-growing is applied to recover complete text lines from video frames.

In 2018, Wei LU et al. [29] decoded the video first into frames, and then candidate text regions were detected using a continuous corner response feature map enhanced by morphological processing and adaptive thresholding. Candidate text lines are accurately localized using contour-based projection analysis or an FCM-based text layer separation method when text lines overlap or have similar lengths. False text lines are eliminated using transfer-learning-based deep CNN classifiers, and the verified text lines are further refined through binarization and morphological restoration for OCR readiness.

In 2016, Yaqi Wang et al. [30] used a method that first generates text candidates using Stroke Width Transform (SWT) with edge detection, ray shooting, and color consistency constraints to extract connected components. On-text candidates are filtered using rule-based constraints and neural network classifiers (MLP/CNN) that categorize components into radicals, single characters, and multi-character blocks. Finally, valid text components are grouped into complete text lines based on geometric, stroke, and color similarity criteria.

In 2024, Zhanzhan Cheng et al. [31] used a method that employs a spatial–temporal video text detector that enhances single-frame text detection by aggregating information across consecutive frames using feature warping, spatial matching, and temporal aggregation. Detected text regions are passed to a unified text recommender that jointly performs quality scoring, metric-learning–based tracking, and attention-based text recognition in an end-to-end trainable framework. The entire system is optimized using multi-task learning with detection, tracking, quality estimation, and recognition losses, followed by inference that selects the highest-quality text from tracked text streams.

III. RELATED LIMITATIONS

Despite these advancements, several limitations remain:

- Data Scarcity: Deep learning models are data-dependent. Compared to languages like English or Chinese, Regional Languages like Gujarati are lacking of the proper large dataset.
- Small Modifiers: In Gujarati, the anusvar (dot) is easily mistaken for noise due to its small size and simple shape, leading to detection errors [22].
- Video Artifacts: Occasionally, video-specific distortions can significantly impair character strokes, often resulting in incorrect boundary identification.
- Lack of structure: The size of the text, the orientation of the text, and the text position is unknown.
- Computational Expense: Models that deliver high performance, such as CRNN and end-to-end spotters, continue to be resource-intensive for real-time processing unless substantial hardware optimization is implemented. [31].

IV. CONCLUSION

This survey aims to provide a comprehensive review of text detection and recognition from images and videos, including the various methods, current challenges, and potential directions. Gujarati text extraction has developed from the frequency-domain analysis of Haar wavelets to the comprehensive modelling capabilities of machine learning. Although wavelets were initially used to handle the special geometry of the script, hybrid CNN-LSTM architectures provide the resilience required to analyse multi-zonal text in dynamic video environments. In order to increase the recognition accuracy of low-resource languages in digital archiving and real-time accessibility technologies, future research should focus on the creation of extensive regional datasets and transfer learning techniques.

REFERENCES

[1] Antani Sameer and Agnihotri Lalitha, "Gujarati Character Recognition." Fifth International Conference on Document Analysis and Recognition (ICDAR'99), pp. 418-421, 1999.

[2] Dholakia Jignesh, Negi Atul and Mohan S. Rama, "Progress in Gujarati document processing and character recognition." In: Govindaraju V., Setlur S. (eds) Guide to OCR for Indic Scripts, Advances in Pattern Recognition, Springer, London, pp. 73-95, 2009

[3] Maloo Mamta and Kale K. V., "Gujarati script recognition: a review." International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 4, pp. 480-489, 2011.

[4] Dholakia Jignesh, Negi Atul and Mohan S. Rama, "Zone identification in the printed Gujarati text." Proceedings of the eight international conference on document analysis and recognition (ICDAR’05), Vol. 1, pp. 272-276, 2005.

[5] Patel Chhaya N. and Desai Apurva A, "Segmentation of text lines into words for Gujarati handwritten text." In international conference on signal and image processing, IEEE Xplore, pp. 130-134, 2010.

[6] Desai Apurva A, "Handwritten Gujarati numeral optical character recognition using Hybrid feature extraction technique." In Proceeding of international conference on image processing, Computer vision and pattern recognition, (IPCV’10), Vol. 2, pp. 733–739, 2010.

[7] Desai Apurva A, "Gujarati handwritten numeral optical character recognition through neural network." Pattern Recognition, Vol. 43, Issue 7, pp. 2582–2589, 2010.

[8] Desai Apurva A, "Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space." CSI transactions on ICT, Vol. 2, Issue 4, pp. 235-241, 2015.

[9] Patel Chhaya N. and Desai Apurva A, "Gujarati handwritten character recognition using hybrid method based on binary tree-classifier and knearest neighbour." International Journal of Engineering Research Technology (IJERT), Vol. 2, Issue 6, pp. 2337-2345, 2013.

[10] Patel Jagin M. and Desai Apurv A., "A comparison of four edge detection methods for identifying Gujarati Numerals from images." VNSGU Journal of Science & Technology, Vol. 3, Issue 2, pp. 113-124, 2012.

[11] Chaudhari Shailesh A. and Gulati Ravi M., "Script identification from bilingual Gujarati-English documents." International Journal of Computer Applications (0975 – 8887) Vol. 93, Issue17, pp. 35-40, 2014

[12] Maloo Mamta and Kale K V, "Support vector machine based Gujarati numeral recognition." International Journal on Computer Science and Engineering, Vol. 3, Issue 7, pp. 2595–2600, 2011.

[13] Mirza, A., Zeshan, O., Atif, M., & Siddiqi, I. (2020). Detection and recognition of cursive text from video frames. EURASIP Journal on Image and Video Processing, 2020(1), 34.

[14] Sain, A., Bhunia, A. K., Roy, P. P., & Pal, U. (2018). Multi-oriented text detection and verification in video frames and scene images. Neurocomputing, 275, 1531-1549.

[15] Bhunia, A. K., Kumar, G., Roy, P. P., Balasubramanian, R., & Pal, U. (2018). Text recognition in scene image and video frame using color channel selection. Multimedia tools and applications, 77(7), 8551-8578.

[16] Joshi, D. S., & Risodkar, Y. R. (2018, February). Deep learning based Gujarati handwritten character recognition. In 2018 International conference on advances in communication and computing technology (ICACCT) (pp. 563-566). IEEE.

[17] Patel, J., & Desai, A. (2018). Gujarati Text Localization, Extraction and Binarization from Images. International Journal of Computer Sciences and Engineering, 6(8), 714-724.

[18] Pareek, J., Singhania, D., Kumari, R. R., & Purohit, S. (2020). Gujarati handwritten character recognition from text images. Procedia Computer Science, 171, 514-523.

[19] Raghunandan, K. S., Shivakumara, P., Roy, S., Kumar, G. H., Pal, U., & Lu, T. (2018). Multi-script-oriented text detection and recognition in video/scene/born digital images. IEEE transactions on circuits and systems for video technology, 29(4), 1145-1162.

[20] Agre, K., Chheda, A., Gaonkar, S., & Patil, M. (2017). Text Recognition and Extraction from Video. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICIATE–2017, 5(01).

[21] Limbachiya, K., Sharma, A., Thakkar, P., & Adhyaru, D. (2022). Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks. Sādhanā, 47(2), 102.

[22] L. Rasikannan, K. G. (2024). Text Extraction from Video using Deep Learning. International Research Journal of Engineering and Technology (IRJET), 947-951.

[23] Raju, N., & Anita, H. B. (2017). Text extraction from video images. Int J Appl Eng Res, 12(24), 14750-14754.

[24] Nithyadhevi, S., Venkatesan, R., Mahendran, B., & Nijasudeen, M. M. (2016). Novel Text Extraction from Video Image. International Journal of Computer Trends and Technology (IJCTT) – Volume 33 Number 1, 43-46.

[25] Tian, S., Pei, W. Y., Zuo, Z. Y., & Yin, X. C. (2016, July). Scene Text Detection in Video by Learning Locally and Globally. In IJCAI (pp. 2647-2653).

[26] Too, B. K., & Prabhakar, C. J. (2016). Extraction of scene text information from video.

[27] Aradhya, V. M., & Pavithra, M. S. (2016). A comprehensive of transforms, Gabor filter and k-means clustering for text detection in images and video. Applied Computing and Informatics, 12(2), 109-116.

[28] Khare, V., Shivakumara, P., Paramesran, R., & Blumenstein, M. (2017). Arbitrarily-oriented multi-lingual text detection in video. Multimedia Tools and Applications, 76(15), 16625-16655.

[29] Lu, W., Sun, H., Chu, J., Huang, X., & Yu, J. (2018). A novel approach for video text detection and recognition based on a corner response feature map and transferred deep convolutional neural network. IEEE Access, 6, 40198-40211.

[30] Wang, Y., Peng, L., & Wang, S. (2016). A multi-stage method for Chinese text detection in news videos. Procedia Computer Science, 96, 1409-1417.

[31] Cheng, Z., Lu, J., Zou, B., Qiao, L., Xu, Y., Pu, S., & Zhou, S. (2020). Free: A fast and robust end-to-end video text spotter. IEEE Transactions on Image Processing, 30, 822-837.