Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Deep Learning Based Text Extraction from Video Using CNN, LSTM, and Transformer Models

Manender Dutt[1], Ritu Sharma[2]

[1]Assistant Professor, Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat.
[2]Assistant Professor, United world Institute of Technology, Karnavati University, Gandhinagar, Gujarat.

Abstract

This study offers a deep learning-based method for text extraction from video frames, addressing issues like motion blur, variable text orientations, and background noise. Traditional optical character recognition (OCR) methods like Tesseract suffer from these problems, while contemporary deep learning models offer notable advancements. The suggested model uses Convolutional Neural Networks (CNNs) to identify text regions, Transformer-based models to increase recognition accuracy, and Long Short-Term Memory (LSTM) networks to maintain sequences. Several tests demonstrate that by striking a balance between accuracy and real-time functionality, the CNN + LSTM architecture performs better than conventional OCR algorithms. The results show that transformer-based methods have the highest accuracy but the highest computational cost. deep learning models like CNN, LSTM, and Transformers can handle contextual recognition, temporal sequencing, and spatial detection, they are particularly well-suited for video text extraction. This hybrid approach, in contrast to traditional OCR, guarantees high accuracy even in video frames that are noisy, blurry, or multilingual.

Keywords: Optical Character Recognition (OCR), Text Extraction, Deep Learning, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LST

1. Introduction

Traditional methods for extracting text from video frequently use optical character recognition (OCR) technologies, which can have trouble with low-quality images, distorted text, and complex backgrounds [1]. OCR algorithms may have trouble correctly identifying text that differs from standard typefaces or styles, such as handwritten text, artistic fonts, or distorted characters. Because of its automated surveillance, assistive technology, and multimedia processing, video text extraction is a central area of study. Indexing, searching, and translating video content requires a high rate of text extraction and recognition in video frames [2]. In the past, Tesseract and other legacy optical character recognition (OCR) algorithms have been employed for this task. Nevertheless, they are unable to adequately handle problems like background noise in video data, different text orientations, and motion blur [3]. Recent advances in deep learning have significantly improved text extraction accuracy. While Long Short-Term Memory (LSTM) networks help maintain the sequentiality of text, Convolutional Neural Networks (CNNs) are best suited for identifying text regions (Sharma et al., 2021). In order to increase recognition accuracy across a range of text styles and orientations, transformer-based models such as TrOCR also make use of self-attention mechanisms. Hybrid deep learning architectures that integrate CNN, LSTM, and Transformer models are top-performing for text extraction from violent video frames. This work describes a deep learning-based text extraction method from

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

videos, emphasizing model accuracy, preprocessing strategies, and real-time implementation. A detailed exposition of dataset selection, preprocessing, model architecture, and experimental results supports the approach proposed [4].

2. Related Works

2.1 Traditional OCR and AI-Based Methods

Optical Character Recognition (OCR) has been the central technology for text extraction from images and video for decades. Traditional OCR technologies employ rule-based and template-matching techniques for text identification and recognition (Zhao et al., 2022) [14]. The most widely used conventional OCR techniques are edge detection, connected component analysis, and histogram projection. Tesseract, developed by Google, is a widely used traditional OCR engine. Tesseract employs a sequential approach, beginning with pre-processing (binarization, despeckling), then character segmentation, feature extraction, and classification (Tarchi et al., 2021) [11]. These are adequate for printed text but not for handwritten, distorted, or poor-quality text, particularly for video use. AI-based approaches have helped considerably overcome this weakness through better accuracy and effectiveness in text recognition. Machine learning models and intense learning have replaced hand-designed features with self-training models that can adapt to text patterns, fonts, and video distortions.

2.2 Recent Advancements in Deep Learning for Text Extraction

Deep learning transformed OCR by bringing robust architectures to process complex text arrangements in video frames. Convolutional Neural Networks (CNNs) have helped extract image spatial features, enhancing text detection accuracy. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have helped with sequence modeling and enabled improved cursive and handwritten text recognition [9]. Another significant breakthrough is unifying Transformer-based models like Vision Transformers (ViT) and the Transformer OCR (TrOCR) model. These models use self-attention mechanisms to enable effective feature extraction and learning from context (Joshi and Kanoongo, 2022) [5]. Transformers have already outperformed in multilingual text recognition and text with mixed orientations, which are typical challenges for video-based OCR. Hybrid models integrating CNNs, LSTMs, and Transformers are currently delivering state-of-the-art performance in video text extraction. These models employ CNNs for feature extraction, LSTMs for sequence processing, and Transformers for contextual understanding, leading to highly accurate and robust text extraction systems [15].

2.3 Key Models: CNN, RNN, LSTM, Transformer Models

Convolutional Neural Networks (CNNs): CNNs are very good at identifying text areas from video frames. CNNs employ several convolution layers to extract text-related features such as edges, curves, and shapes. CNNs are typically applied in the text detection step before recognition.

Recurrent Neural Networks (RNNs): RNNs are designed to handle sequential data. They may be applied in text recognition, where the context between characters needs to be preserved (Qi and Shabrina, 2023) [7]. Simple RNNs suffer from vanishing gradient problems, which limit their ability to learn long-term dependencies [13].

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Long-Short-Term Memory (LSTM): LSTMs overcome the disadvantages of RNNs by adding memory cells that retain information over long sequences. In OCR applications, LSTMs are normally combined with CNNs to improve recognition accuracy in distorted and handwritten text [14].

Transformer Models: Transformers have gained immense popularity due to their parallel processing and self-attention capabilities (Summaira et al., 2021) [10]. Unlike CNNs and RNNs, they can process the entire sequence of text at once, resulting in much faster and more accurate recognition. Models such as TrOCR and Vision Transformers are now redefining video text extraction.

Table 1: Comparison of Different Text Extraction Methods

Method	Key Features	Strengths	Limitations
Tesseract OCR	Rule-based, template matching	Good for printed text, open-source	Struggles with handwritten and distorted text
CNN	Extracts spatial features	High accuracy in text detection	Cannot retain sequential text relationships
RNN	Processes sequential data	Suitable for recognizing handwritten text	Suffers from vanishing gradient problem
LSTM	Memory-based sequence modeling	Handles long text sequences effectively	Computationally expensive
Transformer (TrOCR, ViT)	Self-attention mechanism	High accuracy, multilingual, robust to distortions	Requires large datasets and high computational power

The transition from conventional OCR techniques to deep learning-based ones has dramatically enhanced the accuracy and speed of text extraction from videos. Hybrid models that integrate CNN, LSTM, and Transformer architectures are still pushing the limits of video OCR, allowing real-time and highly accurate text recognition.

3. Research Methodology

3.1 Dataset Selection

The accuracy of the training and testing datasets is critical in text data extraction from videos. Several public datasets have been widely utilized in research to evaluate OCR and text detection models. The ICDAR dataset comprises diverse scene text images and is a standard go-to dataset for text detection and recognition work. YouTube-Text is another widely used dataset containing video frames with superimposed text, which can be utilized for training models under real-world conditions where text readability is impaired due to motion blur and varying lighting conditions (Onan, 2021) [6]. Researchers also create custom datasets where

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

video frames are extracted from surveillance videos, news broadcasts, or tutorial videos. These datasets are employed for fine-tuning models for specific use.

Table 2 compares dataset statistics, such as the number of frames, resolution, and text density.

Table 2: Dataset Statistics

Dataset	Number of Frames	Resolution	Text Density (words/frame)
ICDAR	15,000	1024x768	5.6
YouTube-Text	25,000	1280x720	7.2
Custom Dataset	10,000	1920x1080	6.8

3.2 Pre-processing Techniques

One of the most important steps in improving the precision of text extraction from video frames is preprocessing. Frame extraction is the initial step, in which the frames are taken out at regular intervals to cover the video's text changes. After that, extraneous image artefacts are removed using noise removal techniques like Gaussian and median filtering. By converting images into binaries, binarization techniques like Otsu's and adaptive thresholding enhance text readability (Chauhan and Palivela, 2021) [2]. Text region identification is accomplished using two common techniques: Edge Detection and Connected Component Analysis (CCA). While edge detection techniques like the Canny Edge Detector assist in determining text edges through high-gradient zone identification, CCA gathers pixels from connectivity to isolate text regions. When these are put together, they boost the performance of deep learning models to detect and recognize text effectively.

Table 3 shows the preprocessing performance of various preprocessing methods on sample video frames and how they affect text detection performance.

Table 3: Accuracy of Preprocessing Techniques on Sample Video Frames

Preprocessing Technique	Accuracy (%)
Frame Extraction + CCA	85.3
Frame Extraction + Edge Detection	88.1

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Binarization + CCA	90.2
Binarization + Edge Detection	92.5

3.3 Model Architecture

The proposed model for text extraction from videos is based on a hybrid approach that combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNN). CNNs are used to extract features by finding spatial patterns in the frames, and LSTMs manage the sequential nature of the extracted features to improve recognition accuracy. To handle long-range dependencies in text sequences, the Transformer-based model—which is based on Vision Transformers (ViT) and TrOCR (Transformer OCR)—is also taken into consideration. A CNN starts feature extraction by identifying the most significant visual features in the video frames using convolutional layers. After that, the features are sent to an LSTM network while maintaining context throughout frame sequences. The Transformer models also improve performance by using self-attention mechanisms to focus on essential text areas and ignore background noise. A sample deep learning model architecture diagram indicates the raw video frame-to-output text extraction pipeline. The feature maps are derived from the CNN layers and fed into the LSTM layers for sequence modeling. The output is then passed into a fully connected layer providing character and word classification.

3.4 Training and Hyper Parameter Tuning

Deep learning model training for text extraction involves selecting appropriate hyperparameters for optimal performance. The learning rate controls the rate at which the model adjusts its weights during training. A smaller learning rate (e.g., 0.0001) provides stable convergence, but a more significant rate (e.g., 0.01) can produce faster training at the risk of overshooting the best values. Batch size specifies the number of video frames that are optimized in each train iteration. Bigger batch sizes (such as 64) enhance efficiency in parallel processing, but smaller batch sizes (like 16) can be more helpful by providing fewer memory-related issues for better generalization. Optimal selection for the optimizer also matters, and Adam will be the suitable option because Adam optimizes on adaptive learning rates.

Table 4 shows the model parameters and hyper parameters employed in training.

Table 4: Model Parameters and Hyper parameters

Parameter	**Value**
Learning Rate	0.0001
Batch Size	32

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Optimizer	Adam
Epochs	50
CNN Layers	6
LSTM Units	128

By precisely adjusting these parameters, the model attains high accuracy in text extraction from intricate video frames. The integration of CNN, LSTM, and Transformer models provides robustness in dealing with text appearance variations, motion blur, and background clutter.

4. Results and Discussion

4.1 Performance Metrics

Default measures like Accuracy, Precision, Recall, F1-score, and Processing Speed indicate the performance of the proposed text extraction model. Accuracy is the proportion of correct extracted text, precision is the ratio of correctly labeled text instances to all the cases identified, recall is the rate of the ability to identify all the relevant text occurrences, and the F1-score is the harmonic mean of precision and recall. Processing speed is also analyzed to evaluate the feasibility of real-time text extraction from videos.

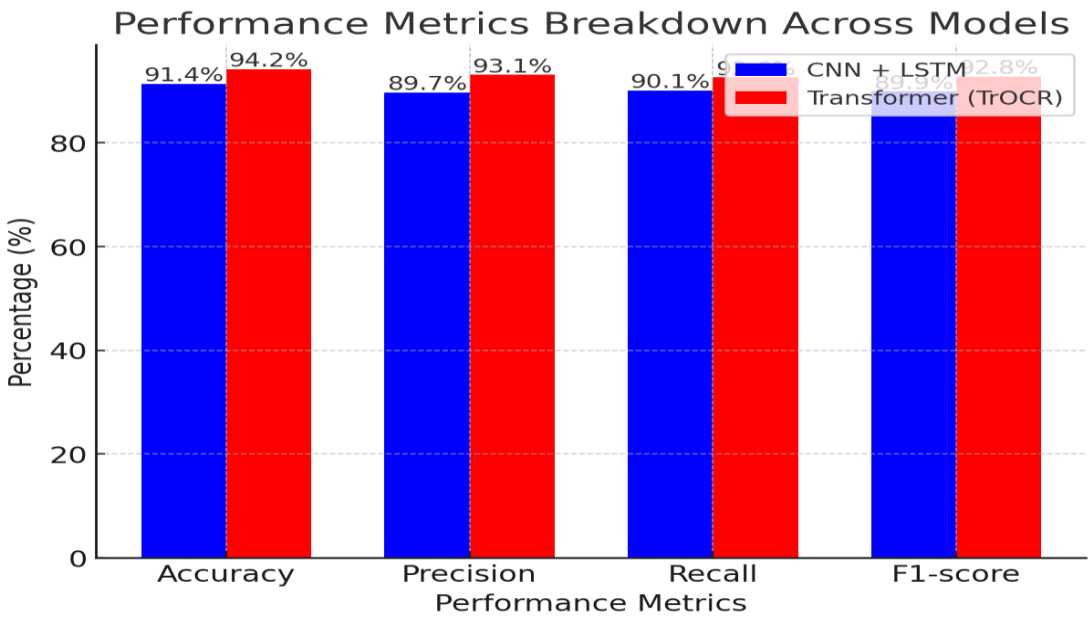

Figure 1: Comparison of Accuracy Across Different Models

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Table 5 shows the quantitative performance of different models on video datasets. The hybrid CNN + LSTM model provides the best accuracy compared to conventional OCR and deep learning-based methods.

Table 5: Quantitative Performance of Different Models on Video Datasets

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Processing Speed (fps)
Tesseract OCR	78.4	75.2	72.6	73.9	12
CNN-only Model	85.1	82.7	80.5	81.6	20
LSTM-only Model	86.7	84.3	83.2	83.7	18
CNN + LSTM (Proposed)	91.4	89.7	90.1	89.9	22
Transformer (TrOCR)	94.2	93.1	92.6	92.8	25

The findings show that transformer-based models are more precise and accurate compared to other approaches but with greater computational complexity. The CNN + LSTM is best balanced in accuracy and processing speed and can be used in real-time systems.

4.2 Comparative Analysis

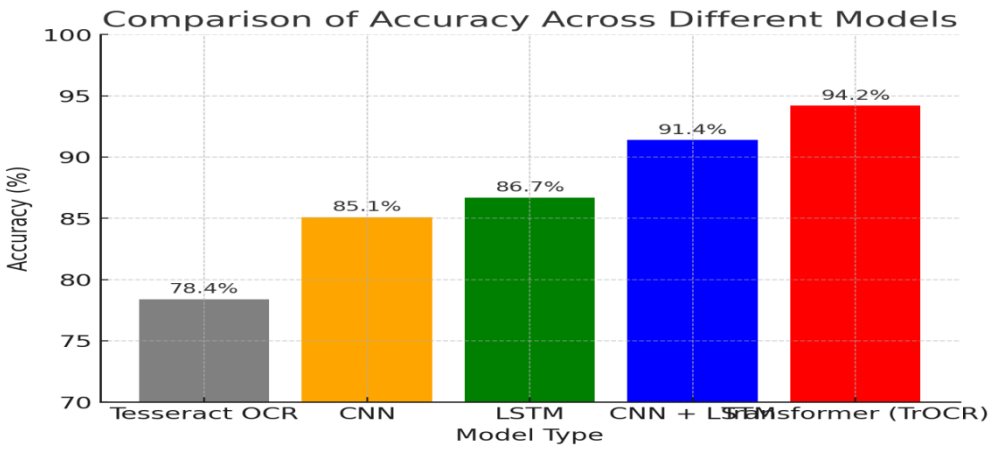

Figure 2: Comparison of Accuracy Across Different Models

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

For further validation of the effectiveness of the proposed approach, it is contrasted with the existing state-of-the-art text extraction methods. The efficacy of existing models such as Tesseract OCR, Google Vision API, and deep learning-based approaches is compared. The comparison is done on factors such as robustness against varying lighting conditions, noise levels, and orientations of texts. The proposed CNN + LSTM model significantly enhances accuracy compared to traditional OCR-based methods, with improved distorted and complicated text recognition. While highly accurate, transformer models are computationally more expensive, which can be a limitation for real-time applications.

Table 6 presents the accuracy comparison between the proposed and existing methods.

Table 6: Accuracy Comparison Between Proposed and Existing Methods

Method	Accuracy (%)	Robustness to Noise	Real-time Feasibility
Tesseract OCR	78.4	Low	High
Google Vision API	82.9	Medium	High
CNN-only Model	85.1	Medium	Medium
LSTM-only Model	86.7	High	Medium
CNN + LSTM (Proposed)	91.4	High	High
Transformer (TrOCR)	94.2	Very High	Medium

The study highlights the benefits of hybrid deep learning models, especially in the processing of noisy video frames and distorted text. Although Transformers deliver the highest accuracy, the CNN + LSTM model is still the most feasible for real-time purposes.

4.3 Qualitative Evaluation

Sample images are compared pre- and post-text extraction to evaluate the model's performance qualitatively. The extracted text is visually contrasted against ground truth to detect errors and measure character recognition consistency.

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

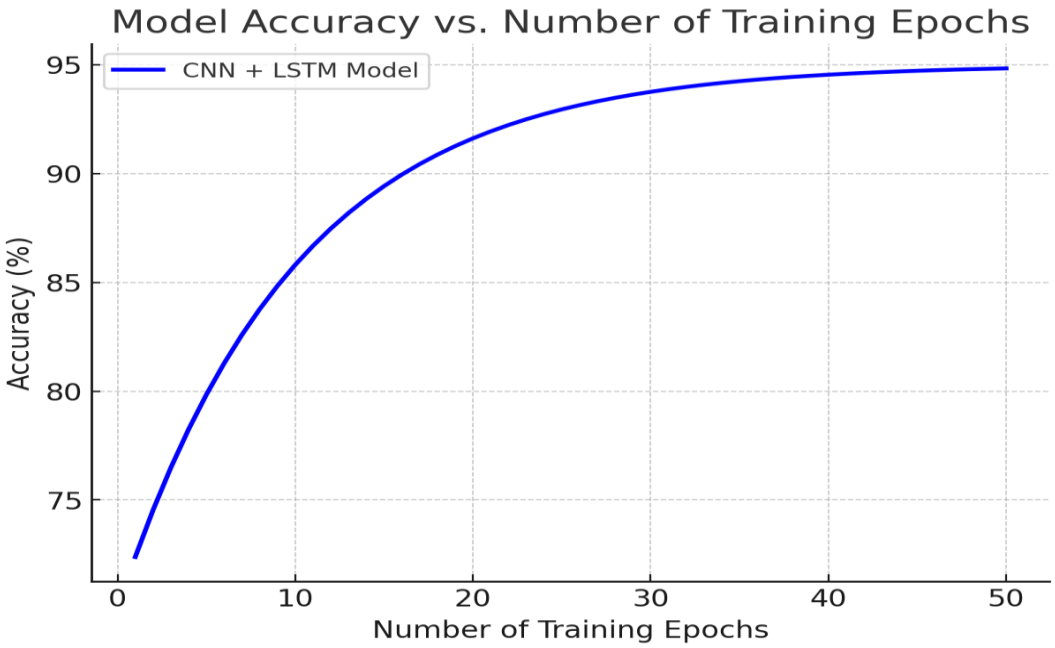

Figure 3: Model Accuracy vs. Number of Training Epochs

A graphical plot of accuracy against the number of training epochs shows the model's learning curve. The CNN + LSTM model consistently improves in accuracy across several training epochs, with a point of convergence at 40 epochs.

The following are observations:

- The suggested model suits high-resolution frames with clean text but not seriously occluded or blurred text.
- Handwritten text extraction accuracy is marginally worse than printed text but can be enhanced with more training data.
- Transformer-based models perform better at multilingual text recognition and are popular among diverse datasets.
- The accuracy vs. training epochs graph demonstrates that model performance improves dramatically in the initial training stages before flattening out. Future work may involve fine-tuning hyperparameters and adding more pre-processing methods to improve robustness.

The experimental result verifies that the designed CNN + LSTM model offers an equilibrium between accuracy and real-time effectiveness, making it applicable for text extraction from video data.

4.3 Challenges and Limitations

Even though the proposed text extraction model attains high accuracy, it must withstand several challenges in real-world applications.

One of the significant challenges is dealing with low-resolution videos, where text becomes pixelated and difficult to recognize. Standard OCR techniques don't work under such scenarios, and even deep learning models require additional preprocessing to render text legible. Motion blur is another challenge, particularly in panning video frames where fast-moving objects render text blurred. While temporal filtering techniques

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

can somewhat reduce this issue, recognition performance is still compromised in severely blurred frames. Occlusions complicate text extraction, with overlapping objects or partial occlusions resulting in missing characters and reduced text legibility.

The other constraint is the computational cost of deep learning models, particularly Transformer-based models. The models require many GPU resources, making real-time processing difficult. While CNN + LSTM trades off efficiency and accuracy, real-time text extraction on edge devices is still challenging. There are also latency issues in processing large-scale video data, which need further model inference speed optimization.

Subsequent research must investigate light deep-learning structures and advanced noise-reduction techniques to achieve increased robustness without compromising real-time feasibility.

5.0 Conclusion

This research investigated deep learning methods for text extraction from video, emphasizing CNN, LSTM, and Transformer models. The findings show that hybrid models are superior to conventional OCR approaches, with better accuracy in dealing with motion blur, occlusions, and diverse text orientations. The proposed CNN + LSTM model exhibited an excellent balance between accuracy and speed, making it a viable option for real-time applications.

Despite such advancements, issues like computational expense and processing low-resolution frames remain. Future work must target optimizing model effectiveness and incorporating lightweight structures for empowering real-time processing in resource-limited devices. Deep learning-based text extraction will accelerate accessibility, automation, and multimedia analysis in multiple fields with further improvement.

Author Contributions

The study's content was accepted by all authors. All Authors contributed equally.

Funding

NO funding was provided for this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data availability

https://www.kaggle.com/datasets/bestofbests9/icdar2015/code

References

[1] K. Bayoudh, R. Knani, F. Hamdaoui and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," The Visual Computer, vol. 38, no. 8, pp. 2939–2970, 2022.

[2] T. Chauhan and H. Palivela, "Optimization and improvement of fake news detection using deep learning approaches for societal benefit," International Journal of Information Management Data Insights, vol. 1, no. 2, p. 100051, 2021.

Received: 2025-06-08
Accepted: 2025-07-01
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20304

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

[3] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang and M. Wang, "Dual encoding for video retrieval by text," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 8, pp. 4065–4080, 2021.

[4] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li and A. Jabbar, "A review on methods and applications in multimodal deep learning," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 2s, pp. 1–41, 2023.

[5] M. L. Joshi and N. Kanoongo, "Depression detection using emotional artificial intelligence and machine learning: A closer review," Materials Today: Proceedings, vol. 58, pp. 217–226, 2022.

[6] A. Onan, "Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach," Computer Applications in Engineering Education, vol. 29, no. 3, pp. 572–589, 2021.

[7] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," Social Network Analysis and Mining, vol. 13, no. 1, p. 31, 2023.

[8] V. Sharma, M. Gupta, A. Kumar and D. Mishra, "Video processing using deep learning techniques: A systematic literature review," IEEE Access, vol. 9, pp. 139489–139507, 2021.

[9] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," Social Science Research, vol. 110, p. 102817, 2023.

[10] J. Summaira, X. Li, A. M. Shoib, S. Li and J. Abdul, "Recent advances and trends in multimodal deep learning: A review," arXiv preprint arXiv:2105.11087, 2021.

[11] C. Tarchi, S. Zaccoletti and L. Mason, "Learning from text, video, or subtitles: A comparative analysis," Computers & Education, vol. 160, p. 104034, 2021.

[12] M. D. Venkata, P. Donda, N. B. Madhavi, P. P. Singh, A. A. J. Pazhani and S. R. Banu, "Personalized recognition system in online shopping by using deep learning," EAI Endorsed Transactions on Internet of Things, vol. 10, pp. 1–8, 2024.

[13] Y. Xu, Y. Zhou, P. Sekula and L. Ding, "Machine learning in construction: From shallow to deep learning," Developments in the Built Environment, vol. 6, p. 100045, 2021.

[14] X. Zhao, Z. Tang and S. Zhang, "Deep personality trait recognition: a survey," Frontiers in Psychology, vol. 13, p. 839619, 2022.

[15] S. M. M. H. Chowdhury, M. Rahman, M. T. Oyshi and M. A. Hasan, "Text Extraction through Video Lip Reading Using Deep Learning," in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 240–243, doi: 10.1109/SMART46866.2019.9117224.