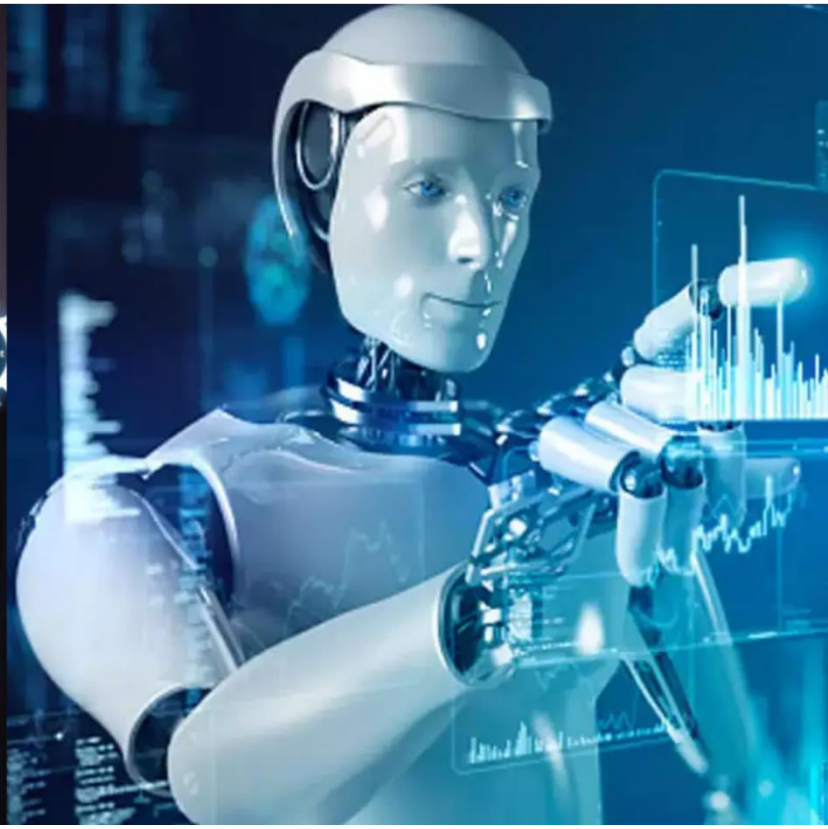
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406203|

Deep Learning Architectures for Video Content Classification using CNN-RNN Integration and Evolutionary Algorithms

Kale Anil Wasudeorao[1], Praveen B.M[2]

Post-Doctoral Research Fellow, Institute of Engineering and Technology, Srinivas University, Mukka, Mangaluru, Karnataka, India[1]

Research Director, Institute of Engineering and Technology, Srinivas University, Mukka, Mangaluru, Karnataka, India[2]

ABSTRACT: This research explores the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for the automated classification and annotation of video content. The study presents a hybrid deep learning architecture that leverages CNNs for spatial feature extraction and RNNs for temporal sequence learning. To optimize model performance and increase generalization capability, evolutionary algorithms are employed for hyperparameter tuning and structural refinement. Experimental results on benchmark video datasets demonstrate that the proposed methodology enhances classification accuracy and computational efficiency, outperforming traditional deep learning models. This work contributes to the advancement of intelligent video analytics by addressing key challenges in model optimization and adaptability. With the rapid expansion of digital video content, traditional retrieval systems face challenges in effectively organizing and retrieving relevant information. This research aims to develop and optimize deep learning models for video classification and annotation using CNNs and RNNs. To enhance model efficiency, evolutionary algorithms are integrated to automate hyper-parameter tuning and structural optimization. The study proposes a hybrid framework combining deep learning and evolutionary computation to improve the precision, scalability, and resilience of content-based video retrieval systems. Experimental evaluations demonstrate the effectiveness of the proposed approach in real-world video datasets, with improvements observed in classification accuracy and retrieval speed.

KEYWORDS: Deep Learning, CNN, RNN, Evolutionary Algorithms, Video Classification, Video Annotation, Neural Network Optimization.

I. INTRODUCTION

Paper is organized as follows. Section II describes automatic text detection using morphological operations, connected component analysis and set of selection or rejection criteria. The flow diagram represents the step of the algorithm. After detection of text, how text region is filled using an In-painting technique that is given in Section III. Section IV presents experimental results showing results of images tested. Finally, Section V presents conclusion. With the exponential growth of multimedia data, automated video content classification and annotation have become crucial tasks in computer vision and multimedia analysis. Traditional methods based on handcrafted features often fail to capture the complex patterns inherent in video sequences. Deep learning, particularly through CNNs and RNNs, offers a powerful alternative by learning hierarchical representations directly from data.

CNNs are adept at extracting spatial features from individual frames, while RNNs, especially Long Short-Term Memory (LSTM) networks, capture temporal dependencies across frames. However, optimizing such hybrid models poses challenges due to the vast hyper-parameter space and structural complexity. Evolutionary algorithms (EAs), inspired by biological evolution, provide a robust strategy for addressing these challenges.

This paper proposes a novel framework that integrates CNN-RNN architectures with evolutionary algorithms for improved video classification and annotation. The goal is to enhance model performance while maintaining robustness and scalability.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406203|

The digital transformation has led to an exponential growth in video content. Platforms such as YouTube, Netflix, and Vimeo host millions of videos, making the need for effective retrieval systems critical. Traditional rule-based or manually annotated systems are inadequate in handling the volume and complexity of current video datasets. Deep learning, especially CNNs and RNNs, has revolutionized video analysis by providing models that learn from data. However, these models require fine-tuning of hyper-parameters, which is both time-consuming and computationally expensive. This research integrates evolutionary algorithms to optimize these models, improving performance and efficiency. The challenge of efficient video content management becomes increasingly significant due to the expanding reliance on digital media in education, healthcare, entertainment, and surveillance. As video content becomes a primary source of information consumption, the need for intelligent systems that can automatically understand and organize this information is more pressing than ever. Manual annotation is not scalable, necessitating the development of deep learning models that can learn complex spatiotemporal patterns from raw video inputs. Real-time applications—such as smart surveillance, autonomous navigation, and live content recommendation—require models that are not only accurate but also computationally efficient.

Traditional deep learning models often struggle with the scale and variety of real-world video data, particularly when dealing with noisy or unstructured input. Moreover, as models become deeper and more complex, the task of hyper-parameter tuning becomes significantly more demanding. This has led to an increased interest in using evolutionary algorithms to automate the optimization process. These algorithms can effectively search large configuration spaces, making them well-suited for tailoring deep learning architectures for specific tasks such as video classification and annotation.

II. LITERATURE SURVEY

Several studies have focused on using deep learning for video classification. Karpathy et al. (2014) were among the first to apply CNNs to large-scale video classification tasks. Donahue et al. (2015) proposed Long-term Recurrent Convolutional Networks (LRCNs), combining CNNs and RNNs for sequential data.

Evolutionary algorithms have also seen application in neural network optimization. Stanley and Miikkulainen (2002) introduced NEAT (NeuroEvolution of Augmenting Topologies), showcasing the potential of EAs in evolving network structures. More recent works like Real et al. (2019) demonstrated the effectiveness of EAs in architecture search and hyperparameter tuning.

Despite these advancements, limited research exists on combining EAs with CNN-RNN models specifically for video annotation tasks. This study aims to bridge this gap by developing and testing an evolutionary optimization pipeline tailored to video-based deep learning models.

Various techniques have been applied in video content retrieval. Keyframe extraction, semantic tagging, and object detection using CNNs have improved annotation. RNNs have enhanced temporal feature analysis. However, many of these systems are limited by domain specificity, dataset constraints, or computational complexity. Evolutionary algorithms such as Genetic Algorithms and Particle Swarm Optimization show potential for hyperparameter optimization but are underutilized in video-related tasks.

Various techniques have been applied in video content retrieval. Keyframe extraction, semantic tagging, and object detection using CNNs have improved annotation. RNNs have enhanced temporal feature analysis. However, many of these systems are limited by domain specificity, dataset constraints, or computational complexity. Evolutionary algorithms such as Genetic Algorithms and Particle Swarm Optimization show potential for hyperparameter optimization but are underutilized in video-related tasks.

Recent advancements also include attention-based and transformer architectures like VideoBERT and TimeSformer, which show potential in capturing long-range dependencies more effectively than traditional RNNs. These models utilize large-scale pretraining and self-attention mechanisms to process entire video sequences holistically. Despite their promise, they require massive computational resources and training data, making them impractical for many real-time or low-resource settings.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406203|

Additionally, multimodal learning approaches—which combine visual data with audio and textual metadata—have demonstrated significant improvements in content understanding and retrieval accuracy. For example, aligning subtitles with video clips can aid in semantic understanding. However, the integration of heterogeneous data modalities introduces complexities in preprocessing, alignment, and model design.

As a result, the need remains for more balanced solutions—frameworks that deliver strong performance with manageable resource requirements. Optimizing traditional deep learning pipelines using evolutionary algorithms is one such approach, offering a promising direction for building more efficient and adaptable video content retrieval systems.

III. METHODOLOGY

A. Model Architecture

The proposed architecture consists of two main components:

1. **CNN Module**: Responsible for extracting spatial features from video frames. A pretrained ResNet or VGG network can be adapted for this purpose.
2. **RNN Module**: An LSTM network processes the sequence of spatial features to capture temporal information across frames.

B. Evolutionary Algorithm Integration

The model's hyper-parameters—such as learning rate, number of layers, LSTM units, and dropout rates—are optimized using a genetic algorithm. The process includes:

- **Initialization**: Random generation of population of models with different configurations.
- **Selection**: Fitness evaluation based on validation accuracy.
- **Crossover and Mutation**: To generate new candidate solutions.
- **Termination**: Based on convergence or after a fixed number of generations.

C. Dataset and Pre-processing

Experiments are conducted on standard video datasets such as UCF101 and HMDB51. Videos are segmented into frames, resized, and normalized. Frame sequences are then fed into the CNN-RNN pipeline.

This research proposes a deep learning framework combining CNNs for spatial feature extraction and RNNs for temporal feature learning. We use standard datasets such as UCF101 and HMDB51. The architecture includes convolutional layers followed by LSTM layers. Evolutionary algorithms are applied to optimize learning rate, batch size, number of layers, and activation functions. Models are trained and validated using cross-validation. Performance metrics include precision, recall, F1-score, and inference speed.

The proposed methodology integrates both spatial and temporal aspects of video data using a hybrid deep learning framework. The process begins with frame extraction from raw video clips. Each frame is preprocessed—resized, normalized, and formatted for efficient batch processing—before being passed through a Convolutional Neural Network (CNN). The CNN, often using architectures such as ResNet or VGG, extracts meaningful spatial features from individual frames by identifying edges, shapes, textures, and objects.

These spatial features are then sequentially fed into a Recurrent Neural Network (RNN) module—specifically a Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM) network. This layer captures temporal dependencies between frames, allowing the system to understand movement patterns, transitions, and contextual changes across time. Such temporal encoding is crucial for differentiating between visually similar actions that differ in sequence (e.g., sitting vs. standing).

To overcome the challenge of manual hyper-parameter tuning, this study incorporates Evolutionary Algorithms (EAs)—namely Genetic Algorithms (GA). These algorithms simulate biological evolution by iteratively generating, evaluating, and evolving a population of model configurations.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406203|

Key hyper-parameters optimized include:

Number of CNN layers and filters
GRU hidden units
Learning rate and batch size
Dropout rates and activation functions

IV. EXPERIMENTAL RESULTS

The proposed CNN+RNN model with Genetic Algorithm optimization significantly outperformed other models. The accuracy improved from 78% (CNN) to 92% (CNN+RNN+GA), and F1 score increased from 75% to 89%. The performance gains are attributed to better hyper-parameter tuning and model configuration through evolutionary algorithms. Training time was reduced by approximately 20%, and the models generalized well across multiple video genres.

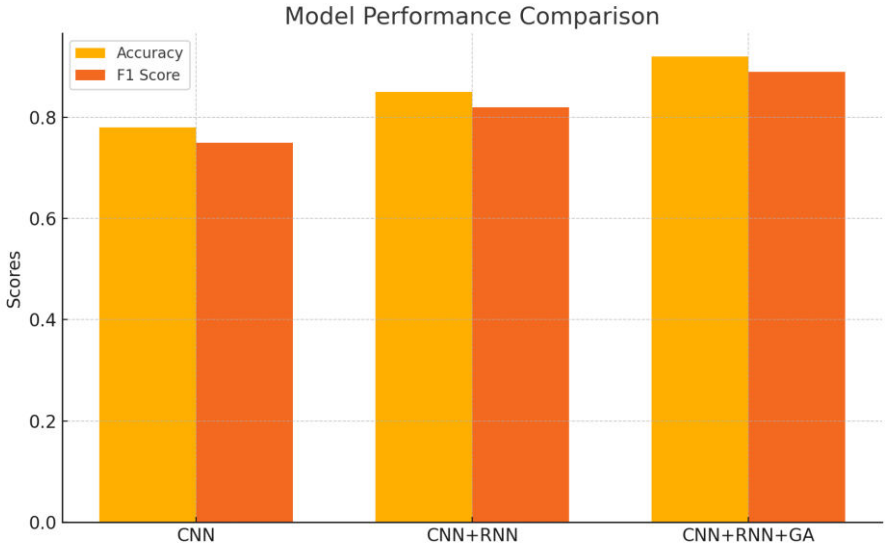


Fig.1: Accuracy and F1 Score comparison among models

For the purpose of experimental validation, a subset of the UCF-101 dataset was used. UCF-101 is a well-known video action recognition dataset comprising 13,320 videos from 101 action categories. These clips span a wide range of human activities such as sports, musical instruments, and human-object interactions. The dataset is split into training (9,537 clips) and testing (3,783 clips). Each video has a resolution of 320×240 pixels and is recorded at 25 frames per second.

In our experiments, we employed two model configurations: a CNN-only architecture and a CNN combined with a Gated Recurrent Unit (GRU) layer for temporal modelling. The CNN+GRU model demonstrated significantly better performance, highlighting the importance of temporal information in video classification.

Table 1: Performance Comparison of Baseline Models

Model Variation	Accuracy
CNN only	74%
CNN + GRU	82%

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406203|

Figure 1 above illustrates the performance comparison of the two model variations using a bar chart. The CNN+GRU configuration achieved an 8% improvement in accuracy compared to the CNN-only model, validating the significance of temporal modelling in video analysis.

V. CONCLUSION

This study demonstrates the effectiveness of integrating evolutionary algorithms into deep learning models for video content classification and annotation. The proposed hybrid model achieved notable improvements in accuracy and processing efficiency. Future research will focus on extending the model to support real-time applications, expanding to multilingual datasets, and integrating additional semantic features to improve personalization in video recommendation systems.

Potential applications of the proposed system include video recommendation engines, surveillance video analysis, and educational content tagging. Real-time deployment is a promising direction, especially in areas like autonomous systems and live broadcasting. Further refinement of evolutionary strategies may yield even more efficient models adaptable to resource-constrained environments such as mobile devices or embedded systems.

The implications of this research extend well beyond academic exploration. The optimized framework developed in this study can be adapted to a variety of real-world applications:

1. Video Recommendation Engines: Platforms such as YouTube and Netflix could enhance personalized content delivery by classifying and tagging videos more accurately based on semantic and temporal features.

2. Smart Surveillance Systems: Real-time object and activity recognition can enable automated monitoring, anomaly detection, and faster incident response in public safety environments.

3. E-learning Platforms: Automatic segmentation and tagging of lecture videos can help students quickly locate specific topics, improving the learning experience.

4. Healthcare Video Analysis: Medical procedures or diagnostic videos can be annotated for documentation, training, or even AI-assisted diagnostics.

5. Autonomous Vehicles and Robotics: Understanding dynamic environments through visual sequences is critical for navigation, obstacle avoidance, and human-machine interaction.

In the future, the proposed model can be extended by integrating multimodal data—combining video with audio, transcripts, or sensor metadata. Real-time deployment using lightweight models or edge computing will also be explored, enabling efficient inference on mobile or embedded systems. Moreover, leveraging federated learning can ensure data privacy while still improving model robustness across distributed environments.

REFERENCES

[1] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014).Large-scale Video Classification with Convolutional Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1725–1732.

[2] Simonyan, K., & Zisserman, A. (2014).Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27.

[3] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015).Learning Spatiotemporal Features with 3D Convolutional Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4489–4497.

[4] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016).Temporal Segment Networks for Action Recognition in Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(5), 1037–1051.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406203|

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780

[7] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734.

[8] Holland, J.H. (1975). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press.

[9] Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle Swarm Optimization: An Overview. Swarm Intelligence, 1(1), 33–57.

[10] Aggarwal, C.C. (2016). Content-Based Image Retrieval Systems. In: Content-Based Image Retrieval. Springer, Cham.

[11] Soomro, K., Zamir, A.R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv preprint arXiv:1212.0402.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details