WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Video Classification Using Convolutional Neural Networks (CNN)

Ayushman Thakur

Maharaja Agrasen Institute of Technology

ABSTRACT :

Video classification is a significant challenge in computer vision, involving the automatic categorization of video content into predefined labels. With the rise in video data across platforms, there is a growing demand for efficient and accurate classification methods. This research explores the application of Convolutional Neural Networks (CNNs) for video classification tasks. It addresses the limitations of traditional frame-by-frame approaches and presents CNNs as a robust solution due to their ability to extract spatial features from video frames. The study adopts a deep learning framework using preprocessed video data, applying CNN models to capture spatial patterns and employing temporal pooling or integration methods for sequence learning. Experimental results on benchmark datasets such as UCF101 demonstrate improved accuracy and performance. The paper concludes by highlighting CNNs' potential in real-time video analysis and suggests future directions, including the integration of Recurrent Neural Networks (RNNs) for capturing temporal dynamics more effectively.

Introduction

The explosion of video content on platforms like YouTube and surveillance systems has made video classification a vital tool in modern artificial intelligence applications. Understanding and organizing video data is essential for tasks such as content recommendation, anomaly detection, and autonomous driving. This study focuses on Convolutional Neural Networks (CNNs), a deep learning approach traditionally used for image classification, adapted to handle video sequences.

The main research problem addressed is: How effectively can CNNs be applied to classify video content based on spatial information extracted from frames? The objectives are to implement a CNN-based model, evaluate its performance on standard video datasets, and identify limitations. This paper is structured as follows: the next section reviews relevant literature, followed by the methodology used, results, discussion, and finally, the conclusion.

Literature Review

Previous research in video classification has employed handcrafted feature techniques such as Histogram of Oriented Gradients (HOG) and SIFT, but these methods lack robustness in complex scenarios. The advent of deep learning has shifted the focus to CNNs and 3D CNNs, which automatically learn spatial features. Karpathy et al. (2014) introduced a CNN model for video classification with modest accuracy improvements. Simonyan and Zisserman proposed the Two-Stream CNN architecture that incorporates both spatial and temporal features.

However, many existing models either lack temporal awareness or require high computational resources. This research aims to bridge these gaps by evaluating standard CNNs' effectiveness and discussing possible integrations with RNNs or LSTM for future improvements.

Methodology

The study utilizes a quantitative research design. The UCF101 dataset, which includes 13,320 videos from 101 action categories, is used for experimentation.

Data Collection: Preprocessed video data, where videos are split into frames and resized uniformly.

CNN Architecture: A simple CNN model with layers including convolution, pooling, ReLU activations, and fully connected layers.

Data Analysis: Performance evaluated using accuracy, precision, recall, and F1-score.

Limitations: The model only captures spatial features and does not account for frame sequence, limiting temporal understanding.

Results

The CNN model achieved an accuracy of 78% on the UCF101 dataset using only spatial features.

Metric	Value
Accuracy	78%

Precision | 76%
Recall | 75%
F1-score | 75.5%
Figures below illustrate the confusion matrix and sample predictions. The results indicate promising performance for basic video classification using CNNs alone.

Discussion

The findings suggest that CNNs can effectively classify videos when focusing solely on spatial features. Compared to traditional methods, the CNN model outperformed in terms of generalization and feature learning. However, the lack of temporal modeling poses a significant limitation.
When compared to Two-Stream or 3D CNNs, the single-frame CNN is computationally lighter but less accurate. Integrating RNNs or LSTM units could improve temporal context understanding. Real-world applications such as gesture recognition or surveillance would benefit from more advanced temporal architectures.

Conclusion

This research demonstrates that CNNs, though primarily designed for image tasks, can be adapted for video classification by treating video frames as a sequence of images. The study highlights the strengths of CNNs in spatial feature extraction while acknowledging their limitations in capturing motion dynamics. Future work could incorporate temporal models like LSTM or attention mechanisms to enhance performance further. These findings provide a foundation for developing more efficient and scalable video classification systems.

REFERENCES :

1. Karpathy, A., Toderici, G., Shetty, S., et al. (2014). Large-scale Video Classification with Convolutional Neural Networks. CVPR.
2. Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. NIPS.
3. Tran, D., Bourdev, L., Fergus, R., et al. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. ICCV.
4. UCF101 Dataset: https://www.crcv.ucf.edu/data/UCF101.php