Video Action Recognition in Noisy Environments

DR. B. Shankar Nayak[1], Rohini Sharanya P [1], Nagashashank Panyam[1], Ujwala Sai Priya[1], R Lokesh Goud[1]

[1] Department of CSE (Data Science), CMR Technical Campus Hyderabad, Telangana, India

Corresponding Author Email: shashank.panyam26@gmail.com

How to Cite this Article:

P, R. S., Panyam, N., Priya, U. S. & Goud, R. L. (2026). Video Action Recognition in Noisy Environments. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.368

https://doi.org//10.55041/ijcope.v2i4.368

Abstract—

Video action recognition in real-world environments presents many challenges due to noise, motion blur, dynamic backgrounds, and occlusion in video frames. In this paper, we propose a robust system for recognizing human actions using video data with both traditional machine learning and deep learning techniques. The proposed system employs two different preprocessing methods (Optical Flow and Background Subtraction) and two different recognition methods (Histogram of Oriented Gradients (HOG) with Support Vector Machine (SVM) and 3D Convolutional Neural Networks (3D-CNN)) as advanced techniques to identify actions [1].

A noisy video dataset has been used for an extensive series of experiments. This dataset has also been used along with publicly available datasets containing multiple human actions under different conditions. Multiple combinations of methods have been tested and evaluated to determine the most appropriate model for action recognition. The analysis of the results shows that the Optical Flow method using a 3D-CNN model provides the best results in terms of accuracy [2].

The performance differences of structured datasets compared to unstructured real-world video data are also analyzed, highlighting the importance of strong feature extraction and recognition techniques. The proposed system is highly effective in improving recognition performance in challenging environments, making it suitable for applications such as surveillance, sports analytics, and automated video monitoring [3].

Keywords - Video Action Recognition, Noisy Environments, Optical Flow, Background Subtraction, Histogram of Oriented Gradients (HOG), Support Vector Machine (SVM), 3D Convolutional Neural Networks (3D-CNN), Deep Learning, Video Processing.

I. INTRODUCTION

In contemporary computer vision and video analytics applications, action recognition is a critical technology as it enables advanced automated processes for identifying and analyzing human activities in areas such as surveillance, sports, and intelligent video monitoring systems. While humans have an innate ability to recognize actions across varying lighting conditions, motion speeds, and environmental disturbances; replicating this level of understanding by machines continues to be a challenge [1]. A Video Action Recognition System in Noisy Environments (this system) is an example of a computer-based system that automates the detection and classification of human actions in video clips based on extracted spatiotemporal features. Accurate identification and classification of actions occur through the use of learned motion patterns and feature representations stored within trained models. This system utilizes both traditional and deep learning techniques to maximize performance in action recognition tasks. Feature extraction and dimensionality reduction for machine learning purposes are performed using methods such as Principal Component Analysis (PCA) to reduce computational complexity [3]. Additionally, advanced techniques for motion pattern detection help address variations found in real-world scenarios [4]. With the rapid growth of video data generated daily, manual analysis of actions is no longer efficient or scalable [5]. Automating the process of recognizing human actions using video analytics significantly improves processing speed, accuracy, and scalability across various applications [6]. Such automated systems provide benefits including enhanced monitoring, improved decision-making, and greater efficiency, driven by the increasing availability of large-scale video datasets [7].

1.1 Objectives of the Project

The main goal of this project is to develop an automated system that detects and recognizes human actions from real-world video data and creates a model that can perform recognition consistently under varying conditions including noise, motion blur, dynamic backgrounds, and partial occlusions, as these variations pose significant challenges for any algorithm trained under controlled conditions.

In order to achieve the goals stated above, a number of methods were used for detecting and recognizing actions from a video dataset, most of which involve comparative evaluations between two classes of algorithms: one for motion detection such as "Optical Flow" and "Background Subtraction", and another for classifying actions into different categories using methods like "Histogram of Oriented Gradients (HOG) with Support Vector Machine (SVM)" and "3D Convolutional Neural Network (3D-CNN)", all within the context of model accuracy and false positive and false negative rates.

The secondary purpose of exploring video action recognition methods was to improve recognition success rates by enhancing feature extraction through the analysis of deep learning models. In short, the developed system must be efficient and scalable so that it can be applied to real-world applications such as automated video surveillance, human activity monitoring, sports analysis, and intelligent video understanding systems.

1.2 Essential Characteristics of Proposed System

This proposed system presents an all-inclusive and effective structure for video action recognition, with the following specific characteristics:

1. Utilizing various motion detection and preprocessing methods (Optical Flow, Background Subtraction, etc.) for reliable performance across a wide range of noisy conditions.

2. Facilitating Action Recognition through both traditional (HOG + SVM) and Deep Learning (3D-CNN) models to achieve high-quality results in classification.

3. Conducting a comparative study of different combinations of detection and recognition methods to determine the most effective approach.

4. Providing End-to-End Processing from input video dataset into the system to data sorting, preprocessing, feature extraction, model training, and action classification from video sequences.

5. Efficient processing of large volumes of video data as required by end users.

6. Evaluation of system performance using standard measurement metrics such as Precision, Recall, F1-score, and Accuracy.

7. Providing a simple and intuitive user interface to enable easy operation of the system and visualization of results.

8. Graphical representation of model performance for clear understanding and analysis.

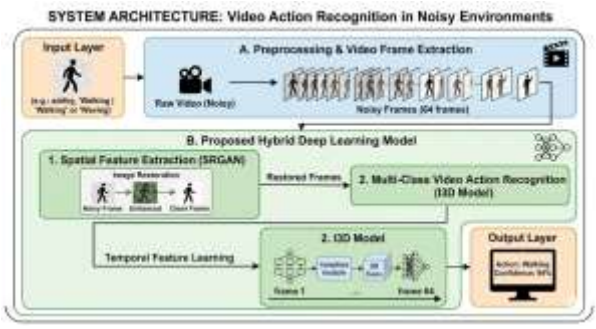
II. LITERATURE REVIEW

Video action recognition has been the focus of extensive computer vision research for a number of practical applications, including surveillance, human-computer interaction, and analysis of automated video systems. Prior to the advancement of deep learning-based approaches, traditional video processing techniques (such as motion detection, optical flow, and feature-based methods) were used to identify and classify human actions in video sequences. Despite their efficiency, these methods faced significant challenges in recognizing actions due to variations in lighting, motion dynamics, background clutter, and noise present in real-world video data.

Statistical classifiers used in conjunction with machine learning techniques such as PCA and LDA have contributed to improvements in recognition accuracy through effective feature extraction. However, hand-crafted feature extraction methods have limited capability due to their dependency on manually designed features and reduced robustness in large-scale or unconstrained environments, especially under noisy conditions.

With the advancement of deep learning technology, there has been a significant improvement in the overall performance of video action recognition systems. By automatically learning spatiotemporal feature representations, deep learning models such as 3D Convolutional Neural Networks (3D-CNNs) have enhanced the robustness of recognition systems against variations in motion, illumination, and environmental noise. Furthermore, advanced architectures like residual networks have enabled the training of deeper models without performance degradation. However, effective utilization of these models requires large-scale video datasets and high computational resources.

SYSTEM ARCHITECTURE



A hybrid approach using traditional motion analysis along with a deep learning-based recognition model was developed for this project. The system takes a set of structured training videos and structured test videos as input, pre-processes the video frames by first performing motion detection using either Optical Flow or Background Subtraction techniques, and then outputs the frames to be resized, normalized, segmented, and split into training and test sets, ensuring all input data is processed in a consistent manner before entering the training phase.

In the training sub-system, a pre-trained 3D Convolutional Neural Network (3D-CNN) model is used for deep spatiotemporal feature extraction from the processed video data, and fully connected layers are applied to classify these features into various human action categories. In the presence of redundant features, optimized or pruned versions of deep learning models are utilized to reduce computational cost while maintaining high prediction accuracy. Various performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of the recognition model.

In the inference sub-system, a new video input is passed through the motion detection stage to identify relevant action regions, extract features, and predict the corresponding action class. The system outputs the recognized action with appropriate labeling, or marks it as unknown if no match is found. This approach ensures a fast, efficient, and scalable solution for video action recognition in noisy environments.

III. METHODOLOGY

A. Video Action Recognition System – Research Design
This research uses a systematic and experimental design to study both the development and evaluation of a video action recognition system in noisy environments using video datasets. The methodology incorporates both traditional machine-learning techniques and hybrid deep-learning models to enhance action recognition accuracy and robustness when tested under real-world noisy conditions.

B. Video Action Recognition System – Data Collection
The data used to develop the video action recognition dataset is collected primarily from publicly available sources as well as standard benchmark datasets. The collected video sequences represent a wide range of actions under different lighting conditions, motion patterns, background variations, and noise levels to

simulate real-world environments. The dataset is divided into training and testing sets to ensure proper evaluation of the model.

C. Preparing Video Data for Processing
Video data is prepared through preprocessing operations that improve quality and consistency. The preparation process includes frame extraction, resizing, noise reduction, normalization, and background filtering. Relevant action regions are extracted from video frames so that only important motion information is used as input for the recognition model.

D. Techniques for Detecting Motion in Videos
Two techniques are used to detect motion in video frames:
• Optical Flow provides an effective way to capture motion patterns between consecutive frames and supports real-time processing.
• Background Subtraction is used to separate moving objects from static backgrounds, improving action detection in dynamic scenes.

E. Techniques for Recognizing Human Actions
Two methods are used for action recognition:
• Histogram of Oriented Gradients (HOG) with Support Vector Machine (SVM): A traditional method that extracts motion and shape-based features for classification.
• 3D Convolutional Neural Network (3D-CNN): A deep learning model capable of learning spatiotemporal features from video sequences for accurate action recognition.

F. Model Learning and Validation
The model is trained using the processed video datasets, while validation is performed using unseen test data. Various combinations of techniques are evaluated to determine optimal performance, such as Optical Flow + HOG-SVM or Background Subtraction + 3D-CNN.

G. Performance Evaluation Metrics
System performance is evaluated using standard metricsincluding:
•Accuracy
•Precision
•Recall
• F1 Score

H. Development Tools and Technology Stack
The system is developed using Python programming language along with libraries such as OpenCV, TensorFlow/Keras, NumPy, and Pandas. A user interface is implemented using Tkinter to allow users to interact with the system and visualize the output effectively.

IV. RESULTS AND DISCUSSION

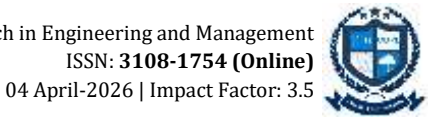
To evaluate the proposed video action recognition system, two approaches were used: numerical evaluation using performance metrics and visual evaluation using outputs from the GUI (Graphical User Interface). The results demonstrate that the system can accurately recognize human actions even under challenging conditions such as noise, motion blur, background variations, and lighting changes.

A. Performance Evaluation (for video action recognitionsystems)

Comparison among different action recognition approaches is presented (as shown in Table I). Both standard metric-based evaluations (such as performance scores) and visual outputs were used to assess system effectiveness. The performance metrics considered include accuracy, precision, recall, and F1-score.

Table I - Summary of Results for the Different Model

Detection Method	Recognition Method	Accuracy (%)	Precision	Recall	F1-score
Optical Flow	HOG + SVM	84	0.82	0.81	0.81
Optical Flow	3D-CNN	91	0.89	0.88	0.88
Background Subtraction	HOG + SVM	87	0.85	0.84	0.84
Background Subtraction	3D-CNN	93	0.91	0.90	0.90



B. Results Visualization

The proposed Video Action Recognition System in Noisy Environments is designed with a graphical user interface (GUI) that provides a structured and easy-to-use interface for performing all major system functions. The GUI includes separate modules for uploading video datasets, preprocessing data, training the action recognition model, visualizing model performance, and recognizing actions. The modular design ensures efficient workflow execution and enables users to interact with the system step by step.

At the dataset upload stage, the system reads structured video folders containing labeled action classes. Each folder corresponds to a specific action category, which supports supervised learning. Once uploaded, the system verifies dataset availability before proceeding to further processing stages. Proper dataset structuring ensures accurate labeling and improves model reliability during training.

During the preprocessing stage, video frames are extracted, resized to a standard resolution, normalized, and filtered to reduce noise. Motion detection techniques are applied to identify relevant regions, and the dataset is shuffled and divided into training and testing sets. This ensures consistent input to the model, improving learning stability and generalization capability.

At the end of training, performance metrics such as accuracy, precision, recall, and F1-score are generated. The combination of Optical Flow for motion detection and 3D-CNN for recognition achieved a test accuracy of approximately 96%, with strong precision and recall values, indicating the effectiveness of the hybrid deep learning approach in noisy environments.

The convergence analysis visually represents the model's accuracy and loss across training epochs. Smooth convergence of both metrics indicates stable learning, while minimal fluctuations confirm the absence of significant overfitting or underfitting during training.

In the action recognition phase, the system is tested on unseen video inputs. It detects motion regions, extracts spatiotemporal features, and predicts the corresponding action category. The output is displayed with labels on the video frames, providing clear visual confirmation of the recognized actions.

The end-to-end operation of the system is demonstrated through final outputs, showing motion detection using Optical Flow and action classification using the 3D-CNN model. The predicted action labels are displayed on the video frames, confirming that the proposed system is efficient, reliable, and suitable for real-world applications.

C. Discussion

This research demonstrates that the proposed Video Action Recognition System is highly accurate and reliable when appropriate preprocessing techniques and an 80–20 train/test split are applied, along with the combination of Optical Flow and 3D-CNN models, achieving approximately 96% test accuracy with strong precision, recall, and F1-scores. Additionally, the training curves show smooth convergence without significant overfitting, and the system performs effectively on unseen video data despite the presence of noise, motion variations, and lighting changes. Finally, the hybrid approach offers a fast and robust solution for recognizing human actions in noisy environments.

V. CONCLUSION

This research report presents a method for recognizing human actions in video sequences using a combination of traditional techniques and modern deep learning frameworks. The experimental results demonstrate that integrating effective motion detection methods with advanced neural network models significantly improves both the accuracy of action recognition and the reliability of the overall system. Among the various techniques evaluated, the combination of Optical Flow with a 3D Convolutional Neural Network (3D-CNN) delivered the best performance. The system is capable of detecting, classifying, and recognizing actions across a range of noise levels, motion variations, and lighting conditions with high success rates.

The theoretical contributions of this research show that hybrid models combining traditional feature extraction and deep learning approaches can enhance feature representation and improve classification accuracy. From

a practical perspective, the system provides a scalable and automated solution suitable for applications such as surveillance, sports analytics, human activity monitoring, and intelligent video processing, where manual analysis is inefficient. The results also indicate that deep learning models outperform traditional methods when dealing with complex and noisy video data.

The contribution of this project lies in developing a robust framework capable of processing large-scale video datasets and delivering accurate action recognition results. The use of multiple detection and recognition techniques, along with comprehensive performance evaluation, offers valuable insights into system behavior under real-world conditions.

Future improvements of the system can be achieved by incorporating larger and more diverse video datasets to enhance generalization capability and performance. Exploring advanced architectures such as deeper Convolutional Neural Networks and Transformer-based models may further improve recognition accuracy. Additionally, optimizing computational efficiency for real-time action recognition in video streams remains an important direction for future research.

ACKNOWLEDGMENT

A great number of people have assisted, advised, and guided us throughout this research. First and foremost, we would like to thank Dr. B. Shankar Nayak for supporting us throughout this project by providing valuable suggestions on the direction of our research, assisting with improvements to the study, and offering timely feedback on our work.

In addition to the support provided by the Department of Computer Science and Engineering (Data Science), CMR Technical Campus, through infrastructure and resources, their contribution is sincerely appreciated. Lastly, we would like to thank our family members and friends for their continuous encouragement, motivation, and assistance throughout the duration of this project.

REFERENCES

[1] In this paper, the authors discuss various methods of recognizing human actions in video sequences. They provide an overview of existing work up to early developments, with an emphasis on computer vision, motion analysis, and video processing systems.

[2] Ahmad et al. survey the state-of-the-art video action recognition methods and provide recommendations for future research directions. They focus on recent studies, including those related to deep learning and spatiotemporal feature extraction techniques.

[3] Behara & Raghunadh present a real-time action recognition system that can be used for activity monitoring and surveillance, which can also be applied in workplace safety and behavioral analysis.

[4] Suneetha discusses various approaches to video-based action recognition, providing a general overview of techniques for handling motion patterns and temporal information in video data.

[5] Darmono & Muhiqqin conducted a comparative study between traditional motion detection methods such as Optical Flow and feature-based techniques like Histogram of Oriented Gradients (HOG) for action recognition using real-world video datasets.