International Journal of Science, Strategic Management and Technology
Volume **02** Issue **04** April-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Noise-Resilient Video Action Recognition

K. Venu[1], Ch. Padmasree[2], G. Rasagna[3], G. Madhav Goud[4]

[1] Assistant Professor,[234] UG Scholars
Department of Computer Science and Engineering (AI & ML)
Nalla Narasimha Reddy Education Society's Group of Institutions
Hyderabad, Telangana, India
[1] venukairoju@gmail.com,[2] padmasreechunchu9@gmail.com,[3] rasagnyagaddam@gmail.com,[4] madhavgandhamalla@gmail.com

https://doi.org/10.55041/ijsmt.v2i4.053

Cite this Article: Padmasree[2],, C., Rasagna[3],, G. & Goud, G. M. (2026). Noise-Resilient Video Action Recognition. International Journal of Science, Strategic Management and Technology, *02*(04). https://doi.org/10.55041/ijsmt.v2i4.053

***Abstract*—**Video action recognition is a significant research domain in the field of computer vision, which has various practical applications, such as surveillance systems, sports analysis, health monitoring, intelligent video analysis, etc. However, the videos obtained from the environment are noisy, meaning that the videos may be blurred, the lighting conditions may be poor, the objects may be occluded, the background may be cluttered, etc. As a result, the accuracy of the video action recognition system is compromised. In this paper, a video action recognition system is proposed, which is more efficient in handling noisy videos. To achieve this, the video is processed, and the frames are obtained. Then, the obtained frames are enhanced using the Super Resolution Generative Adversarial Network, after which the InceptionV3 convolutional neural network is employed for the action classification. By enhancing the video frames, the accuracy of the video action recognition system is improved. Experiment results prove that the proposed system is more efficient compared to the other systems.

***Keywords*—**Video Action Recognition, Deep Learning, Computer Vision, SRGAN, InceptionV3, Human Activity Recognition, Video Processing

I. INTRODUCTION

Human action recognition from video sequences has become a key research area in computer vision and artificial intelligence. This process enables automatic detection and categorization of human activities through video analysis. The technology serves an essential function in intelligent surveillance systems, healthcare monitoring, robotics, human-computer interaction, and sports performance analysis.

The increasing amount of video content produced by smartphones and CCTV cameras and online platforms has created a requirement for automated systems that can analyze video content with high efficiency. The process of manually assessing extensive video data sets requires too much time and becomes unworkable. Intelligent video analysis systems need to exist because they must automatically identify and track human movements.

People still encounter difficulties when they attempt to identify actions from real-world videos even though deep learning methods have advanced significantly. Real-world recordings include noise which results from motion blur and poor lighting and occlusion and camera shake. The visual quality of video frames decreases because these elements apply noise, which prevents machine learning models from obtaining useful features.

Traditional action recognition approaches relied on handcrafted features such as Histogram of Oriented

Gradients (HOG), optical flow, and trajectory-based features. The methods failed to capture all the details about complex spatial and temporal video patterns.

Recent innovations in deep learning methods have produced substantial enhancements for action recognition systems. Researchers achieved success in image classification through the use of Convolutional Neural Networks (CNNs) together with InceptionV3 deep learning models. The models experience difficulties when they receive input frames that contain either noise or low-resolution content.

Super-resolution techniques can enhance video frame quality through their application to solve this problem. Super-Resolution Generative Adversarial Networks (SRGAN) excel at creating high-resolution images from low-resolution input materials.

Our research presents a video action recognition system which maintains effectiveness against noise through its integration of SRGAN frame enhancement and InceptionV3 for action detection. The system extracts frames from videos which it then enhances before using these enhanced frames to classify actions.

The main contributions of this work are:

• SRGAN technology enables the enhancement of noisy video frame material.
• The InceptionV3 deep learning model enables action classification.
• The system achieves better recognition results when handling noisy video datasets.

II. LITERATURE SURVEY

[1]. ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton have proposed a deep CNN model in their research paper titled "ImageNet Classification with Deep Convolutional Neural Networks." The authors have proposed an architecture for CNN that has improved the accuracy level in image classification with the help of the ImageNet dataset. The proposed CNN architecture uses multiple layers of convolution and pooling with fully connected layers to perform image classification. The proposed CNN architecture is effective in image classification. However, the proposed CNN architecture requires large datasets and high computational power.

[2]. 3D Convolutional Neural Networks for Human Action Recognition

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu proposed a framework for human action recognition with the help of 3D CNNs in their research paper titled "3D Convolutional Neural Networks for Human Action Recognition." The authors have proposed an architecture for CNNs by introducing 3D convolution filters to the network. The proposed CNN architecture is effective in recognizing actions with the help of videos. However, the proposed CNN architecture has relatively high computational complexity.

[3]. Two-Stream Convolutional Networks for Action Recognition in Videos

Karen Simonyan and Andrew Zisserman proposed a two-stream deep learning architecture for video action recognition in their research paper titled "Two-Stream Convolutional Networks for Action Recognition in Videos." This architecture improves the accuracy of action recognition in videos. However, it has the disadvantage of increasing computational cost in the system by using optical flow.

[4]. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

Christian Ledig and his research team proposed the SRGAN model in their research paper titled "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." This SRGAN model improves the image resolution by reconstructing high-resolution images. However, it has the disadvantage of increasing computational cost in the system by using GAN.

[5]. Rethinking the Inception Architecture for Computer Vision

Christian Szegedy and his research team proposed the InceptionV3 architecture in their research paper titled "Rethinking the Inception Architecture for Computer Vision." The proposed model is based on the use of convolutional filters of varying sizes for efficient classification. However, the training of the proposed

model requires a large dataset and time for training the model.

[6]. Large-Scale Video Classification with Convolutional Neural Networks

Andrej Karpathy and his research team proposed a large-scale video classification approach in their research paper titled "Large-Scale Video Classification with Convolutional Neural Networks." The authors proposed the use of convolutional neural networks for efficient classification of large-scale videos. The proposed approach is based on the efficient training of convolutional networks for the classification of large-scale videos. However, the proposed approach requires a large-scale dataset for efficient training of the proposed model.

[7]. Temporal Segment Networks for Action Recognition in Videos

Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao proposed the Temporal Segment Networks in their research paper titled "Temporal Segment Networks for Action Recognition in Videos." The authors proposed a framework that segments videos into various segments. They then extracted features from each video segment. This is to ensure that the action recognition performance is improved by incorporating both short-term and long-term information. However, it is important to note that the proposed method has to ensure that the video segments used are optimal.

[8]. SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He proposed the SlowFast network architecture in their research paper titled "SlowFast Networks for Video Recognition." The authors proposed the SlowFast network architecture. This architecture is used in video recognition. It has the ability to recognize various human actions. However, it is important to note that the proposed architecture has an increased computational complexity. This is because it has two separate pathways.

[9]. Video Action Recognition Using Deep Learning

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah introduced the UCF101 dataset for human action recognition in their research paper titled "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild." The dataset includes a large number of realistic video frames representing various human actions. This dataset is considered to be one of the most popular benchmarking datasets for video action recognition. However, the dataset might contain variations in lighting conditions that could affect the overall performance.

[10]. Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun introduced the Residual Network (ResNet) for deep learning in their research paper titled "Deep Residual Learning for Image Recognition." The researchers introduced the concept of residual connections that help deep learning networks learn more effectively without the occurrence of the vanishing gradient problem. This concept enhances the feature extraction abilities of deep learning networks for image and video recognition. However, deeper networks require more computational power for the overall process.

III. PROPOSED METHODOLOGY

The proposed system is designed to recognize human actions in noisy video environments by using a mix of super-resolution techniques and deep learning approaches. The main purpose of this proposed system is to improve the accuracy of action recognition by enhancing the quality of video frames.

The proposed method for recognizing actions in a video involves several stages, including frame extraction, frame enhancement, feature extraction, and action classification. Each of these stages is significant in contributing to the efficiency of the proposed action recognition system.

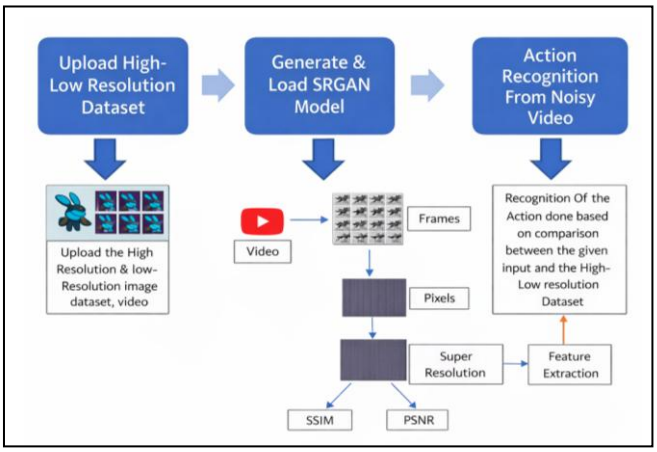

Figure. 1. Proposed Methodology for Action Recognition

A. Video Input Module

The first step of the system is the input of the video file. The input video may contain some level of noise because of the low lighting conditions, movement of the actors, background, and movement of the camera. The overall quality of the frames in the input video is reduced, and the accuracy of the action recognition system is compromised.

The system takes input in various formats such as MP4, AVI, and MOV. The input video is further processed frame by frame for further processing.

B. Frame Extraction

After the input of the video, the frames are extracted from the input video using various video processing techniques. The frames are extracted using the OpenCV library. The frames are the individual images of the input video at specific time intervals.

The frames are extracted from the input video because the input video needs to be processed using the image processing capabilities of the machine learning algorithm. The individual frames are extracted from the input video for the classification of the action.

C. Frame Enhancement using SRGAN

The extracted frames may contain noise or low resolution, which may affect the classification model. To solve this problem, a Super-Resolution Generative Adversarial Network (SRGAN) is used to enhance the quality of the video frames.

SRGAN consists of two neural networks:

- Generator Network
- Discriminator Network

The generator network takes low-resolution images and converts them to high-resolution images. The discriminator network checks if the generated images have realistic features or not.

Using this technique, the SRGAN network can learn to produce high-quality images with rich features. The generated images can be used for action recognition.

D. Feature Extraction using InceptionV3

After the video frames have been enhanced, they can be passed to the InceptionV3 deep learning network. The InceptionV3 network is a convolutional neural network architecture used for efficient image classification.

The network consists of multiple convolution filters of various sizes within a single layer. This allows the network to detect features of various sizes. The features extracted by the network contain important visual features of the video frames.These features can be passed to the classification network.

E. Action Classification

In this step, the extracted features are used to identify the action present in the video sequence. The classification layer is used to predict the action category based on the features extracted by the InceptionV3 model.

The actions identified by this model are:

- Riding horse
- Marching
- Javelin throw
- Baseball hitting

The output from the individual frames is used to generate the final action performed in the video sequence.

F. Output Generation

The final step in this action recognition system is to generate the output by displaying the action label and the processed video frames. The output is also used to generate annotated frames with the identified action.

This helps the user understand the actions performed in the video sequence.

IV. SYSTEM ARCHITECTURE

The system architecture of the proposed noise-resilient video action recognition system comprises various interconnected modules that collectively work for the processing of the video data and the recognition of the human actions. The system architecture is developed for the improvement of the accuracy of the recognition of the human actions by enhancing the quality of the video frames.

A. Video Input Module

The first module in the system is the video input module. This module is used to input the video into the system. The system supports various video formats like MP4, AVI, MOV, etc. However, the input video might contain noise caused by various environmental factors like low lighting, motion blurring, etc.

Once the video is input into the system, it is sent to the preprocessing stage where the video is extracted into frames.

B. Preprocessing and Frame Extraction

In the preprocessing stage, the input video is segmented into various frames using video processing techniques. Frame extraction is carried out using OpenCV library functions.

Each frame in the video represents an image captured during the video recording. The extracted frames are then resized and normalized before being sent to the enhancement stage. Frame extraction is an important stage in the system because the input video is extracted into frames before being sent to the enhancement stage. This is because deep learning models can only process image data.

C. Frame Enhancement Module

The extracted frames may have noise or low resolution, which may affect the classification model. To solve this issue, the system makes use of a Super-Resolution Generative Adversarial Network (SRGAN) to enhance the quality of the extracted frames.

The SRGAN consists of the following neural networks:

• Generator Network

• Discriminator Network

The Generator network works to produce high-quality images from low-resolution images. The Discriminator network works to check if the generated images are realistic. By using these two networks, the SRGAN works to produce high-quality images that contain important details.

The enhanced frames obtained from the SRGAN model will have better visual features for the action recognition model.

D. Feature Extraction Module

After the frame enhancement module, the frames are passed to the feature extraction module. The feature extraction module makes use of the InceptionV3 convolution neural network to obtain important features from the frames.

The InceptionV3 model has several convolutional layers that detect edges, textures, shapes, etc. The convolutional layers use various sizes of filters within a single layer to detect features of various sizes within an image.

These features obtained from the InceptionV3 model contain important visual features for recognizing human actions.

E. Action Classification Module

In this module, the classification of the features extracted by the InceptionV3 model takes place. The classification of the action in the video frames occurs in this module. The classification function used here is softmax.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **04** April-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

The actions that can be classified by the system include:

• Riding a horse

• Marching

• Baseball hitting

The output of each frame of the video is used to classify the action that was performed.

F. Output Module

The final module of the action recognition system consists of displaying the output of the system. The output of the system includes the action that was performed, as well as the video frames processed by the system.

The output of the system can be a video frame with the action classified.

V. EXPERIMENTAL ANALYSIS

To execute the proposed system, the user has to follow the following steps:

1. Open the web application interface of the proposed system, i.e., the Noise Resilient Video Action Recognition System, as shown in Fig. 2.

2. Through the interface, the user can upload the video that has to be processed.

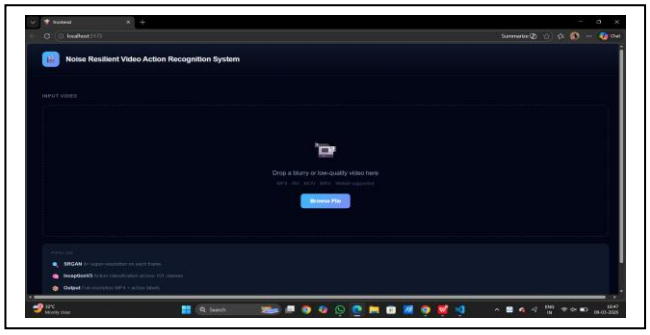

Figure. 2: System Interface for Uploading Input Video

In the above interface, the user can upload a low-resolution video by either dragging and dropping the video file or by

clicking the 'Browse File' button to upload the video from the local system.

Once the video file has been selected, the system will load the input video for processing, as shown in Fig. 3.

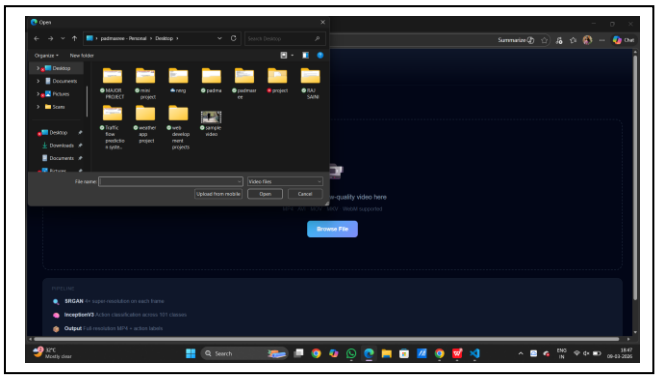

Figure. 3: Uploading Input Video for Processing

After the video has been uploaded for processing, the system will start processing the video frames. The video frames will be processed using the SRGAN (Super Resolution Generative Adversarial Network) model to enhance the quality of the video frames.

As the video frames are processed, the system will display the video frames' enhancement and recognition progress, as shown in Fig. 4.

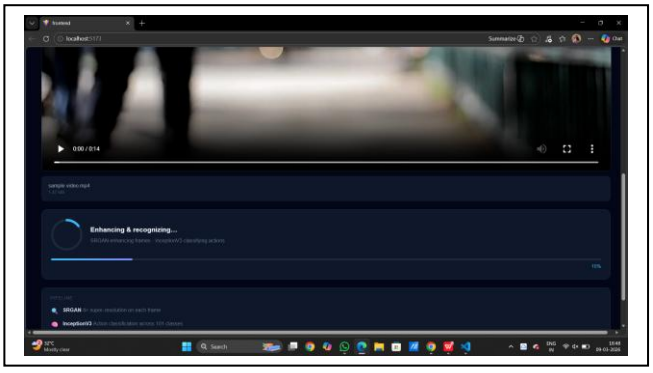

Figure. 4: Processing Video Using SRGAN and InceptionV3

In this stage, the system will enhance the quality of the video frames using the SRGAN super resolution techniques. The video frames will be passed to the

InceptionV3 deep learning model for classification after they have been enhanced.

Once the processing is done, the system shows the recognized action along with the confidence level, as depicted in Fig. 5.

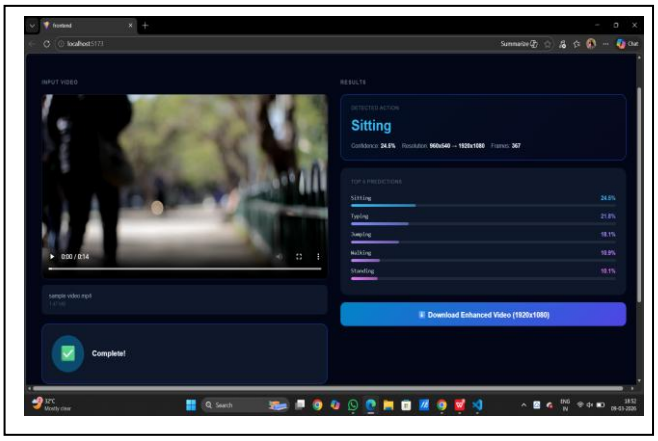

Figure. 5: Action Recognition Result

In the above action recognition result, the system recognizes the detected action as "Sitting" with a confidence level of 24.5%. Additionally, the system shows the video resolution before and after enhancement and the number of processed frames.

The system shows the Top 5 predicted actions, which are "Sitting," "Typing," "Jumping," "Walking," and "Standing," along with the respective probability scores.

The enhanced video obtained from the SRGAN model can be downloaded in high resolution (1920 x 1080) for further analysis.

VI. RESULTS

The results obtained from the proposed Noise Resilient Video Action Recognition system show that the proposed system is effective in combining SRGAN-based super resolution enhancement with deep learning-based action recognition. This is achieved by using the SRGAN model to enhance the low-quality video frames, followed by action recognition using the InceptionV3 convolutional neural network.

The experimental results show that the proposed method is effective in improving the quality of video frames before action recognition. This is achieved by providing improved spatial details in the video frames.

Figures 6-8 show the results obtained from the proposed system in recognizing various actions in videos. From the results, it is clear that the proposed system is effective in recognizing various actions in videos. This is achieved by comparing the original low-quality video frame with the SRGAN-enhanced video frame.

Figure. 6. Action being recognized as Feeding Horse

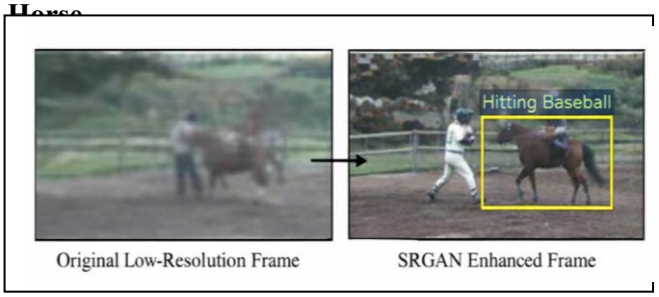

Figure. 7. Action being recognized as Hitting Baseball

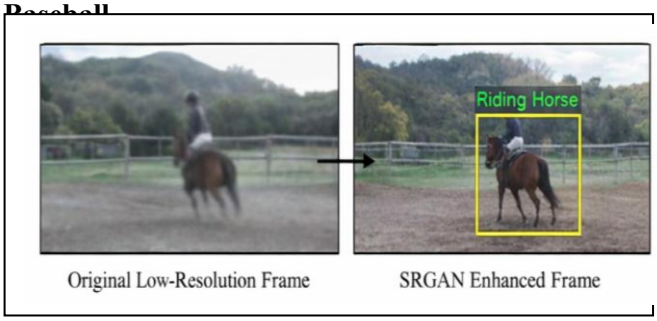

Figure. 8. Action being recognized as Riding Horse.

VII. CONCLUSION

In this paper, a noise-resilient video action recognition system was proposed using the SRGAN model and the InceptionV3 deep learning model. The video input is processed using the SRGAN model to enhance the video frames for better feature extraction.

The SRGAN model was used to enhance the quality of the low-resolution video frames. This enables the deep learning model to perform better feature extraction for video action recognition. The experimental results prove that the video action recognition system can successfully recognize different human actions.

From the results, it is evident that the video action recognition systems perform better when the super-resolution techniques are applied to the video frames.

VIII. FUTURE WORK

Although the proposed system demonstrates promising results, there are certain areas that can be explored for further improvement in the system. One such area is the development of real-time action recognition systems using optimized deep learning approaches.

Future research can also focus on the development of the system using the transformer approach for video understanding, as the approach has been promising for the latest computer vision-related problems. Moreover, the system can be trained using larger datasets for improving the generalization capacity of the system.

Further improvement in the system can also include optimizing the system for real-time processing and developing the system for real-time surveillance and smart monitoring applications.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton,
"ImageNet Classification with Deep Convolutional Neural Networks,"
NIPS, 2012.

[2] S. Ji, W. Xu, M. Yang, and K. Yu,
"3D Convolutional Neural Networks for Human Action Recognition,"
IEEE TPAMI, 2013.

[3] K. Simonyan and A. Zisserman,
"Two-Stream Convolutional Networks for Action Recognition in Videos,"
NIPS, 2014.

[4] C. Ledig et al.,
"Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,"
CVPR, 2017.

[5] C. Szegedy et al.,
"Rethinking the Inception Architecture for Computer Vision,"
CVPR, 2016.

[6] A. Karpathy et al.,
"Large-Scale Video Classification with Convolutional Neural Networks,"
CVPR, 2014.