Interpretable Human Action Recognition: A CNN-GRU Approach with Grad-CAM Insights

Md. Sabir Hossain[1], Mufti Mahmud[1,3], and Md. Mahfuzur Rahman[1,2]

[1]Information and Computer Science Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, [2]Interdisciplinary Research Center for Intelligent Secure Systems, KFUPM, Saudi Arabia,[3]SDAIA-KFUPM Joint Research Center for AI and Interdisciplinary Research Center for Bio Systems and Machines, KFUPM, Saudi Arabia.

Submission Number: 21

Abstract

This research introduces a CNN-GRU-based Human Action Recognition (HAR) framework combined with Grad-CAM for post-hoc interpretability. The system is trained on a 10-class subset of UCF101 and reaches 96.5% accuracy, outperforming deep CNN baselines while offering transparency in its decision-making.

Key Highlights

- Efficiently captures both spatial and temporal features from video sequences, achieving robust performance across diverse human actions.
- Frame-level visual explanations via Grad-CAM enhance transparency, aiding in understanding and trust in model predictions.
- Achieves 96.5% accuracy on a 10-class subset of UCF101, outperforming standard CNNs and matching state-of-the-art, without requiring complex multi-stream inputs.

Methodology Overview

- 10 frames sampled per UCF101 video
- CNN extracts spatial features per frame
- Conv3D and GRUs model spatio-temporal features
- Grad-CAM highlights critical attention areas

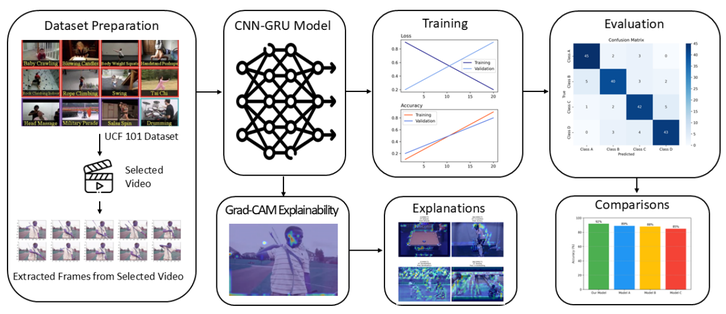

Fig. 1: Overview of the CNN-GRU pipeline with Grad-CAM explainability.

CNN-GRU Architecture

- 3x Conv2D + BatchNorm + Pooling
- Conv3D for spatio-temporal features
- Two GRUs (32, 50 units)
- Dense + Softmax output

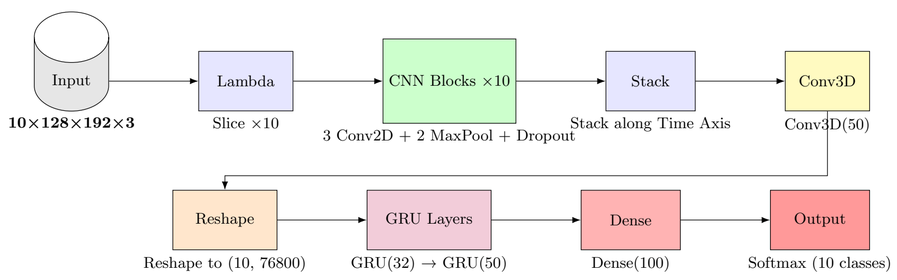

Fig. 2: CNN-GRU architecture for spatio-temporal action recognition.

Grad-CAM Working Pipeline

- Gradient-based heatmaps identify class-specific cues

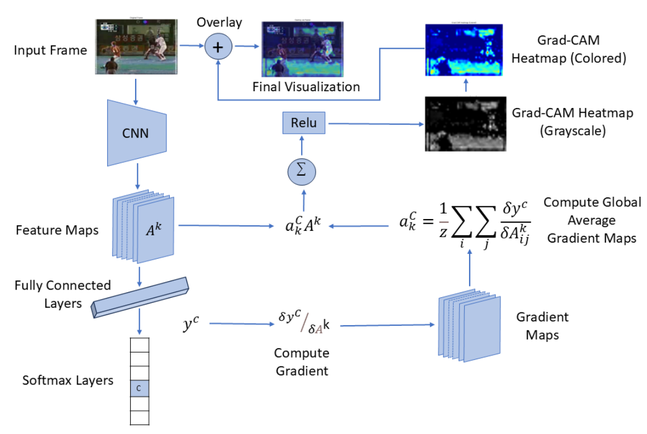

Fig. 3: Grad-CAM pipeline highlighting class-relevant regions.

Experimental Setup and Results

- Dataset: 10 UCF101 classes (100 videos each)
- Accuracy: 96.5%
- Hardware: Tesla T4 GPUs, Keras/TensorFlow

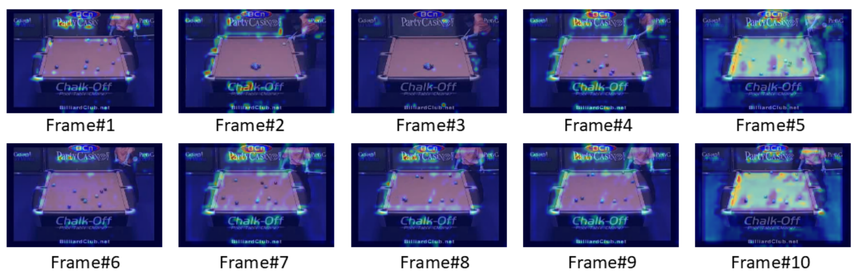

Fig. 4: Grad-CAM heatmaps for correctly classified *Billiards*.

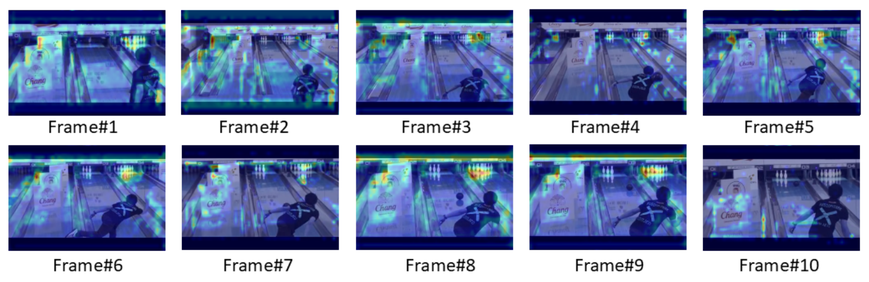

Fig. 5: Grad-CAM heatmaps of misclassified *Bowling*.

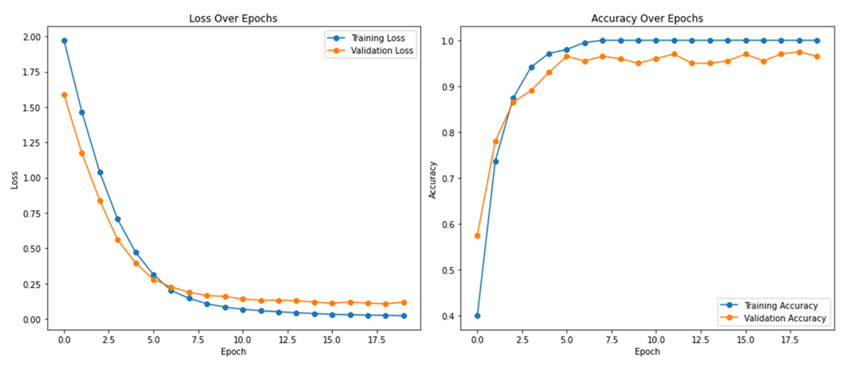

Fig. 6: Training and validation loss and accuracy.

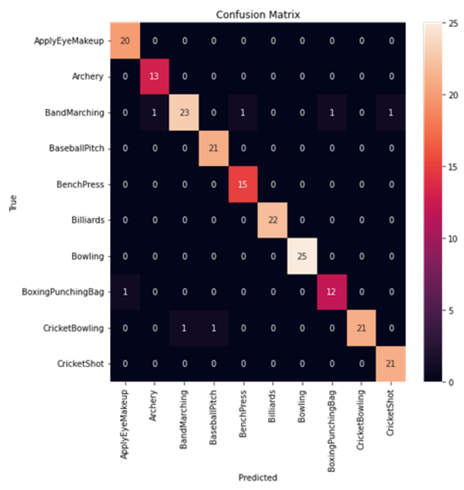

Fig. 7: Confusion matrix for 10-class UCF101 HAR.

Comparative Analysis

Comparison with baselines

Model	Acc	Prec	Rec	F1
CNN-GRU	0.97	0.97	0.97	0.96
Xception	0.94	0.95	0.94	0.94
DenseNet121	0.92	0.95	0.92	0.93
InceptionV3	0.90	0.91	0.90	0.89
MobileNet	0.88	0.91	0.88	0.88
ResNet50	0.33	0.30	0.33	0.24
VGG16	0.07	0.01	0.07	0.01
VGG19	0.07	0.00	0.07	0.01

Comparison with UCF101 SOTA

Model	Acc (%)
CNN-GRU (Ours)	**96.50**
A2-Net (ResNet-50)	96.40
I3D-LSTM	95.10
TS-LSTM	94.10
Two-stream+LSTM	88.60
HalluciNet	79.83

References

1. Abdellatef, E., Al-Makhlasawy, R. M., and Shalaby, W. A. *Detection of human activities using multi-layer convolutional neural network.* Scientific Reports, 15(1):7004, 2025.
2. Alam, M. T., Acquaah, Y. T., and Roy, K. *Image-based human action recognition with transfer learning using Grad-CAM for visualization.* In IFIP Int. Conf. on Artificial Intelligence Applications and Innovations, pp. 117–130. Springer, 2024.
3. Aquino, G., Costa, M. G. F., and Filho, C. F. F. C. *Explaining and visualizing embeddings of one-dimensional convolutional models in human activity recognition tasks.* Sensors, 23(9):4409, 2023.