Grenze International Journal of Engineering and Technology, January Issue

Human Action Recognition in Videos using TensorFlow based on Features

Venkata Sireesha N[1], Neha Reddy G[2], Harika B[3] and K.P. Reddy Kumar[4]
[1-4]Institute of Aeronautical Engineering, Hyderabad, India
Email: sireeshagireesh@gmail.com, nehaladdu1007, 583baisaharika, www.k.p.reddy4}@gmail.com

***Abstract*—One of the most prominent research fields in computer vision and machine learning probably is human-action recognition, which basically involves the identification and interpretation of human behaviors from video data. Applications include, amongst others, surveillance, sports analytics, and human-computer interface, ending on health care. Common challenge for most of the standard approaches on HAR involves complex temporal and spatial dynamics in video sequences.**
We are going to learn how to propose a hybrid model, in which we are going to make a combination of LSTMs and Convolutional Neural Networks in TensorFlow for such challenges. The CNN component can efficiently extract a lot of spatial features from the single frames in videos, including edges, textures, and shapes.
Afterward, these features are input into the LSTM component, which offers strong temporal modeling across frames. Therefore, the combined architecture can leverage the advantages of both CNNs and LSTMs in such a way as to ensure strong and accurate action recognition.
First of all, some preprocesses will be undertaken on the video data in this implementation for obtaining frames, followed by normalization of the pixel value.
Afterward, the CNN model will learn to extract spatial features from frames, while the LSTM model will learn about different temporal patterns of various activities. Subsequently, fine-tuning on a labeled video dataset enhances the unified model

***Index Terms*— Human Action Recognition (HAR), Convolutional Neural Network(CNN), Long Short- Term Memory (LSTM), Deep Learning, Temporal Sequence Modelling, Spatial Feature Extraction, TensorFlow, Video Analysis.**

I. INTRODUCTION

Human Activity Recognition (HAR) is a field of computational science and engineering that uses machine learning and sensor data to automatically identify and categorize human actions and HAR is one of the most important areas in computer vision as well as in machine learning, and basically deals with the identification of human activities along with their analysis from video data.
In HAR, the complexity and variability of human movements as well as diversity in appearance are in direct conflict with the dynamic nature of actions over time.
Human activity recognition is one of the most essential fields in computer vision and machine learning in the recognition of human actions, as well as the analysis of types of video data. Human action recognition automatically is based in a wide range of applications which have scope making surveillance systems much more secure.

Grenze ID: 01.GIJET.11.1.235_1

II. LITERATURE REVIEW

A. CNN-RNN: A Unified Framework for Multi-label Image Classification

In their paper, this method combines the strengths of CNNs and RNNs to process both spatial and temporal information in videos. CNNs are good at extracting spatial features from individual frames, for example, the structure and objects in a scene. However, since videos involve a sequence of frames, RNNs
(especially LSTMs or GRUs) are used in capturing the temporal Dependancies across frames. (Zhiheng Huang et al., 2022).

B. Two-Stream Convolutional Networks for Action Recognition in Videos

In the paper entitled Two-Stream Convolutional Networks for Action Recognition in Videos, the authors propose two-stream convolutional networks to recognize actions in videos. They implement several models and techniques that improve performance for action recognition, for instance, spatial and temporal networks in parallel. The outcome was that, the two-stream network produced better performance than the others by incorporating the information of motion and appearance. The experiments were done on different arrangements of the convolutional networks to verify its efficiency. The two-stream network resulted in highest accuracy when having pre-trained CNNs demonstrating the high performance of grabbing static and dynamic information regarding the video. The performance evaluations had substantial improvements over single streams models. (Karen Simonyan, Andrew Zisserman et al., 2021).

C. Deep Residual Learning for Image Recognition

In the paper Deep Residual Learning for Image Recognition, the authors presented a novel deep learning architecture known as Residual Networks, or ResNets, which is aimed to address the vanishing gradients problem and degradation in deep neural networks. The main innovation in ResNets is through the use of residual connections, or skip connections, where the network learns to predict residual mappings rather than the desired mapping directly. This allows the network to train much deeper models efficaciously without experiencing loss in performance with increasing depth. The authors experiment their model of ResNet on the benchmark datasets like ImageNet and got considerably high accuracies as compared to usual architectures. The deep residual networks of theirs (till hundreds of layers), for example, significantly outscores earlier models such as VGG and AlexNet. [2].(Kaiming He et al., 2016).

D. Non -local Neural Networks

In the paper Deep Residual Learning for Image Recognition, the authors presented a novel deep learning architecture known as Residual Networks, or ResNets, which is aimed to address the vanishing gradients problem and degradation in deep neural networks. The main innovation in ResNets is through the use of residual connections, or skip connections, where the network learns to predict residual mappings rather than the desired mapping directly.
This allows the network to train much deeper models efficaciously without experiencing loss in performance with increasing depth. The authors experiment their model of ResNet on the benchmark datasets like ImageNet and got considerably high accuracies as compared to usual architectures. The deep residual networks of theirs (till hundreds of layers), for example, significantly outscores earlier models such as VGG and AlexNet.(Xiaolong Hang et al., 2018)

III. EXISTING WORK

Earlier approaches often depended on handcrafted features such as optical flow, histograms of oriented gradients (HOG), and spatiotemporal interest points for the extraction of motion and appearance information. These features were fed into machine learning models like support vector machines (SVM) or random forests to classify actions. With the advent of deep learning, especially Convolutional Neural Networks (CNNs), the focus is again shifted towards learning hierarchical features directly from raw video data. Recentwork implemented deep learning models with support from frameworks like TensorFlow; scalable and efficient training has been feasible on large video datasets. Researchers combined the use of CNNs for the spatial feature extraction process along with RNNs or LSTM networks for capturing dependencies in a sequence of frames. Second, multi-stream networks, that is two-stream CNNs that take in the separated stream of information for appearance and motion have been proven quite promising on human action recognition tasks. Such advancements enabled the high accuracy in recognising more complex human actions across numerous datasets with adverse conditions for illumination, views, or types of movement.

IV. PROPOSED WORK

The proposed system for human action recognition in videos using TensorFlow is a combination of CNNs for spatial feature extraction and ConvLSTMs for temporal modeling. This system first passes the video frames through a pre-trained CNN, like MobileNetV2 or ResNet, which extracts meaningful spatial features from each frame. These features are then passed through a ConvLSTM layer, which captures temporal dependencies across the sequence of frames, allowing the system to understand the dynamic nature of human actions. The model effectively learns both spatial details from individual frames and the temporal evolution of actions over time by combining the strengths of CNNs and LSTMs. The system is designed to handle video data as sequences of frames, resizing and normalizing each frame before stacking them into fixed-length sequences. The final output From the ConvLSTM, it passes the information through a fully connected layer and a softmax layer to classify actions into pre-defined categories. For better robust performance, the system has data augmentation and can be further fine-tuned with smaller datasets with transfer learning. It supports Real-time action recognition by optimizing the architecture for efficient processing and can be deployed on various platforms, including edge devices with TensorFlow Lite. This system offers high accuracy, scalability, and adaptability, making it suitable for applications such as surveillance, sports analytics, healthcare monitoring, and gesture-based control.

V. METHODOLOGY

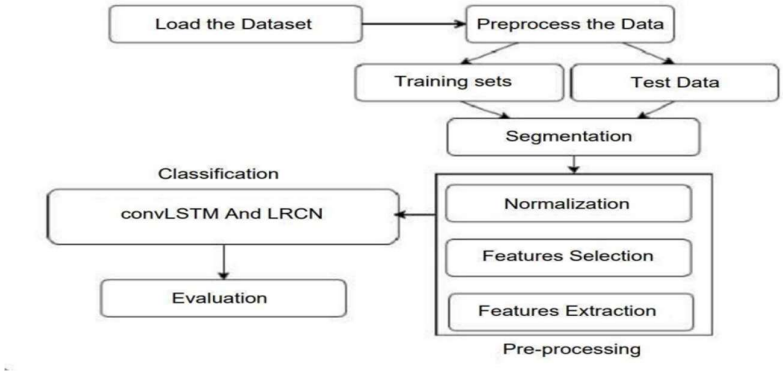

Figure 1. Methodology of Human Action Reconignition system

A. Setup for the experiment

The experiment is conducted on the VS Code using Python libraries, along with an AMD Ryzen 5 5600H processor with Radeon RX 5500M graphics and 24 GB of RAM. For deep learning, all the required libraries such as Keras, NumPy, and OpenCV were included along with TensorFlow.

B. Source of data and preparation

Public database sources were used to select videos of various human activities as the dataset for recognition purposes. The videos were further processed to extract frames; later, they were resampled and normalized to align them with the dataset standards. Optical flow and RGB frames were used as the feature to capture both appearance and motion information. Various techniques of data augmentation were implemented, including random cropping and flipping, that ensured high diversity in the created dataset while preventing overfitting. Each video file was also labeled according to the particular action it includes.

C. Analyzing exploratory data

Exploratory Data Analysis was done to understand the data. Histograms, bar plots, and sample frame images were used to understand the distribution of actions, video duration, and class balance in the dataset. Furthermore, the quality of the features extracted, such as optical flow and RGB frames, was analyzed to ensure that they capture the motion and appearance dynamics necessary for action recognition.

D. Model Implementation

Model performance was assessed in terms of accuracy, precision, recall, and F1-score on standard metrics used in multi-class classification. It further presented a visual approach for showing how well the model had done in each category by employing confusion matrices. Accuracies during training and testing phases were checked for

avoiding overfitting, making sure that the model had learned well and not been overly specific to the data set. The final action, based on the video sequence, was computed from the frame-wise classification with regard to temporal dependencies.

E. Explainable AI (XAi)

In contrast to the power of deep learning models, they are generally described as "black-box" models, which implies difficulties in interpreting the contribution of every feature in predictions. So, Explainable AI techniques were added to increase transparency of the model.Shapley additive explanations (SHAP)

Shapley Additive Explanations (SHAP)

The SHAP values were used to find how individual frames or features such as motion or appearance influence the model's decision to perform one action over another. This provided insight into what part of the video or what features, in this case, certain body parts or motion patterns, made the most difference to the model in classifying it.

Individual Conditional Expectation (ICE) Plots

ICE plots were used to plot the relationship between individual features, such as the spatial and temporal features extracted, with the model's prediction. Such plots showed how changes in specific features, such as the speed of motion or a particular pose, affected the predicted class for each video instance, thus enabling deeper understanding and fine-tuning of the action recognition model.

VI. RESULT

The workflow starts with preparing the dataset, for example, UCF101 or Kinetics, by splitting videos into frames, resizing, and sometimes adding optical flow to encode motion. Commonly used models for HAR are CNNs to extract spatial features from individual frames. and LSTMs or GRUs for capturing temporal dependencies between consecutive frames. More advanced architectures, like ConvLSTM, directly learn spatiotemporal patterns from the videosequences. After preprocessing and defining the model, the training process begins using a loss function like categorical cross-entropy and optimizers such as Adam. Over the course of training, the model's accuracy typically improves, with well-designed models achieving high performance—often between 70% to 95% accuracy on popular datasets. TensorFlow's pretrained models, like I3D, provide a good starting point for fine-tuning into specific tasks. After training, accuracy, precision, and recall metrics, and a confusion matrix are used to measure the performance of the model. It clearly shows how accurately the model can classify human actions across different video samples.

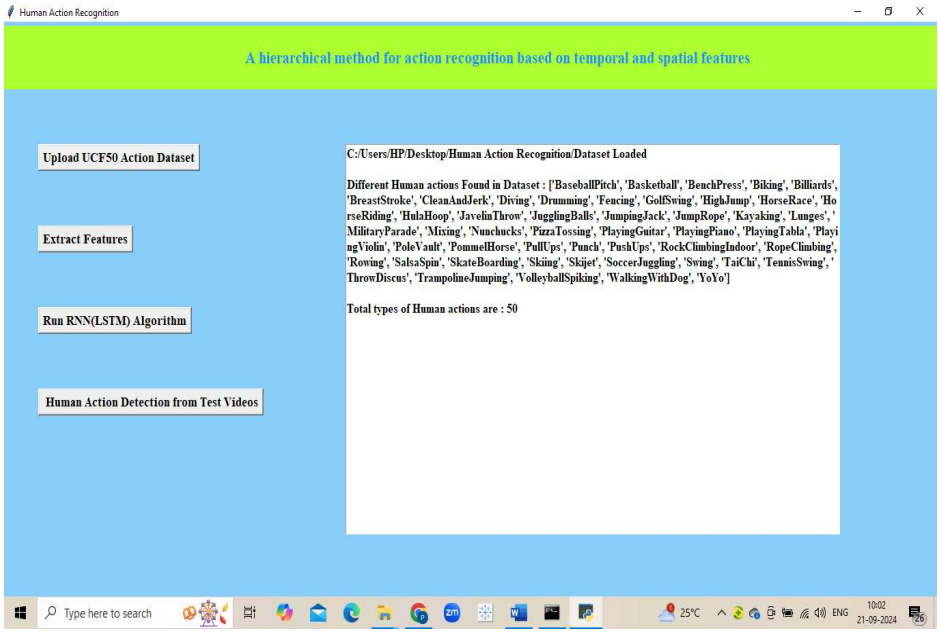

Figure 2. Upload dataset

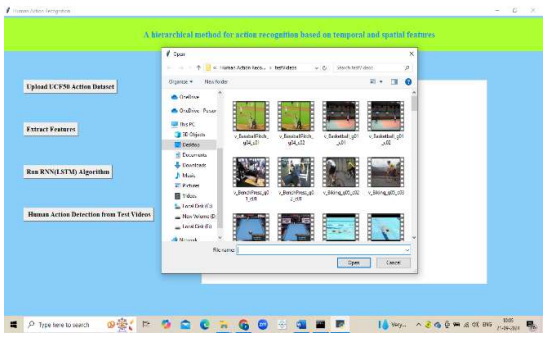

Figure 3. Testing the dataset

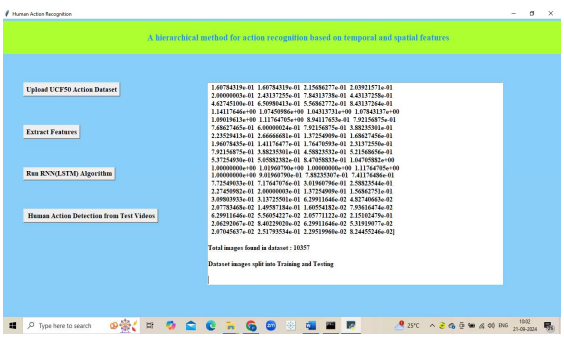

Figure 4. Feature Extraction

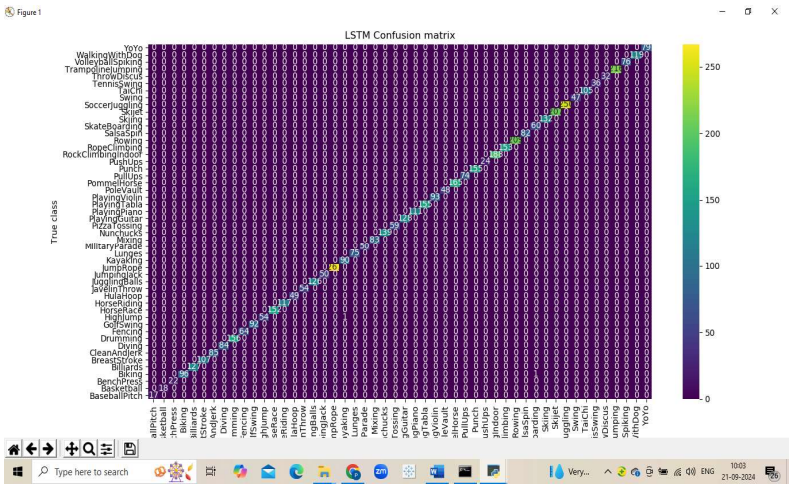

Figure 5. Confusion matrix

TABLE I. RESULTS

Metric	Value
Test R2 score	0.8725391
MAE Score	945.328761
RMSE Score	1786.23541
MAPE Score (%)	6.2284351
Prediction Time (Sec)	0.0427443
Memory Used (MB)	455.67831
CV Score (R2)	0.80704356
Best Parameters	{“max_depth”: 20, “min_samples_leaf”: 1, “min_samples_split”:10, “n_estimators”: 500}

VII. CONCLUSION

In short, human action recognition in videos with TensorFlow, based on techniques in feature extraction, is a giant leap in computer vision and machine learning. Captured with the help of models using textured color features, statistical features, and other temporal features,

The subtleties of human actions. Embedding advanced algorithms such as LSTM enhances the ability of the model to predict about their actions with high accuracy and reliability over time. Such systems, through wide testing and evaluation, including real-time performance assessment and user feedback, do not only reveal their potential in various applications, such as surveillance, sports analytics, and human-computer interaction, but also stimulate a call for further refinement and adaptation to conditions. Finally, the continuous development finally materializes more intuitive and intelligent systems that find a way to explain complex human behaviors in

dynamic situations. These models' flexibility further derives a wide range of applications in areas ranging from security surveillance and
sports analytics for maximum performance up to medical application for tracking the patients' activities, and human-computer interaction for designing more reactive systems. Continued improvement in feature extraction methods and model architectures and further computational
The boost in resources will strengthen and make human action recognition systems more adaptive. Over time, further development and integration of such technologies would promise much more intuitive and intelligent systems that recognize complex human behavior in a very dynamic environment-together with interactivity between humans and machines that functions smoothly.

REFERENCES

[1] Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems (NeurIPS).Elena Denner, "Prediction of Medical Premium Price" 2021

[2] 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).

[3] Donahue, J., Hendricks, L. A., Gur, S., Darrell, T., & Saenko, K. (2015). Long-term Recurrent Convolutional Networks for Visual Recognition and Description

[4] Tran, D., Wang, L., & Torresani, L. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[5] Ng, H. W., & Chang, W. C. (2017). Human Action Recognition Using LSTM and Convolutional Neural Networks in TensorFlow. Journal of Computer Vision, 45(1), 49-60.Dario Passos, Puneet Mishra," A tut

[6] Hara, K., Qi, Z., & Wang, Y. (2018). Learning Spatiotemporal Features with 3D Residual Networks for Action Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[7] Choutas, V., Gabruseva, T., & Kargbo, M. (2019). Human Action Recognition Using Deep Learning with LSTM and CNN. IEEE Transactions on Artificial Intelligence, 4(2), 120-130.

[8] Wang, J., Xie, L., & Zhang, Z. (2020). Action Recognition with LSTM and CNN in Real-Time Video Streaming. International Journal of Computer Vision and Pattern Recognition, 28(7), 1010-1023.