International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME101
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Real-Time Human Action Recognition using Lightweight Deep Learning Networks

[1]Mr. K. Jayachandra, M.Tech,
Assistant Professor,
Department of AIDS,
Annamacharya Institute of Technology and Sciences,
Tirupati.-517520, A.P. .
jayachandra502@gmail.com

[2]Jyoshna .Y
UG Scholar, Dept. of Artificial Intelligence And Data Scinece,
Annamacharya Institute of Technology and Sciences,
Tirupati, India
joshujoshu253@gmail.com

[3]Mohammad Bilal C
UG Scholar, Dept. of Artificial Intelligence And Data Science,
Annamacharya Institute of Technology and Sciences,
Tirupati, India
mohammadbilal200327@gmail.com

[4]Hema Sandhya C
UG Scholar, Dept. of Artificial Intelligence And Data Science,
Annamacharya Institute of Technology and Sciences,
Tirupati, India
hemasandhya5@gmail.com

[5]Davoodh Shaik
UG Scholar, Dept. of Artificial Intelligence And Data Science,
Annamacharya Institute of Technology and Sciences,
Tirupati, India
sdavoodhdavoodh@gmail.com

***Abstract*—** Human Action Recognition (HAR) from video streams is an important research area of computer vision. The applications of this research area include healthcare monitoring, smart home, and tele-immersion. However, recognizing human actions is a challenging problem due to several reasons, including the appearance of humans and the occurrence of occlusions, lighting conditions, and complexity of backgrounds. The overall performance of the HAR system depends on an efficient feature extraction mechanism and the training of the system. In recent times, Deep Learning (DL) methods, particularly neural networks, have shown an improvement in the overall accuracy of the HAR system. In this work, pre-trained Deep Learning models, namely VGG19, DenseNet, and EfficientNet, are used for efficient feature extraction from the video streams and classification of actions using the SoftMax classifier. The performance of the overall HAR system is

tested using the UCF50 dataset, consisting of 50 different types of actions.

Keywords: Transfer Learning, CNN, VGG19, UCF50

I. INTRODUCTION

Human Action Recognition (HAR) may be defined as the detection of human activities using the visual data collected through cameras or sensors. Human actions, such as walking or running, are recognized in the field of view and analyzed to detect the kind of action being performed. Human activities

may be classified into four kinds, as described in [1]. Human activities are:

- Gesture: Facial expressions or actions without any kind of verbal communication.
- Action: Human actions, such as walking, playing, or punching.
- Interaction: Human interactions, such as handshaking or hugging, or interactions between humans and objects.
- Group Activity: Human activities that involve two or more people.

The importance of Human Action Recognition has increased in the past two decades as a research area in computer vision, as it has various applications, as described in [1]. Human activities are recognized as a hierarchical structure, as described in the following steps:
Simple motion elements
Actions
Interactions

Some of the HAR domains include video surveillance, image labeling, health tracking, automation, and environmental assessment, among others. Three tiers of human behavior can be defined as primitive actions, individual actions, and intricate interactions, with the complexity of actions developing over time in an unpredictable manner.

When performing action recognition from videos, the basic processes involve examining the poses of the individuals depicted in the videos. This is challenging as many factors affect the outcome.

In this stage, actions are identified and learned based on the features that are available. The success of this stage depends on the feature model that we select. To be more specific, broadly speaking, the key players of HAR are Machine Learning (ML) and Deep Learning (DL).

- The ML approach emphasizes selecting features that are unique for specific classes of actions.
- The DL approach emphasizes Deep Neural Networks (DNNs) for extracting features from images and their attributes.

Why is Deep Learning becoming more prominent for action recognition? This is because DNNs have the ability to comprehend images and their attributes similarly to how humans do.

Figure 1 depict ML and DL base classification for HAR.

Fig. 1. A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR
[2].

Various approaches for human activity recognition (HAR) have been investigated, such as Random Forests, Bayesian Networks, Markov Models, and Support Vector Machines, all of which have achieved impressive results. However, the consensus is that with more data, the results could be taken to the next level. Recently, HAR has started to adopt Deep Learning techniques, which have achieved impressive results in all domains of unsupervised, supervised, and reinforcement learning. The key advantage of Deep Learning is that it can learn complex data representations in an unsupervised manner using multiple hidden layers.

The paper briefly surveys the field of "Human Action Recognition," touching on "Machine Learning" and "Deep Learning." It covers the methods used, results obtained, and conclusions drawn. The final section comprises conclusions and prospects in Computer Vision.

II. RELATED WORK

The majority of work being done today in Human Action Recognition is focused on implementation and how it improves the accuracy of predictions. In recent times, many methods have been proposed for action recognition from visual data. Initially, features were used for action recognition, and these features could recognize actions when they occurred individually. However, these features were not good at generalization and achieving high accuracy. Later, Convolutional Neural Networks and spatiotemporal networks were proposed for action recognition from videos.

Dai et al. proposed a dual-stream attention LSTM for action detection and precise frame localization, achieving 96.9% on UCF11, 98.6% on UCF Sports, and 76.3% on j-HMDB.

Du et al. developed a skeleton-based action recognition system using hierarchical RNNs, testing it on five state-of-the-art deep models on three datasets, namely MSR Action-3D, Berkeley MHAD, and HDM05.

Majd and Safabakhsh proposed a correlational ConvLSTM network for action recognition, achieving 92.3% on UCF101 and 61.0% on HMDB51 datasets.

Qi et al. proposed Stag-Net, achieving 90.5% on volleyball group activities, while on individual activities, it achieved merely 8.5%.

A method of 3D CNN for posture feature by Huang et al. integrates 3D pose, 2D appearance, and motion data. It reduces computation and unnecessary info by using 3D convolutions over 15 heatmap channels, each of which contains a joint feature over video frames. BN-Inception, as proposed by Wang et al., incorporates Inception blocks with BN to process RGB and optical flow data for motion, as well as background noise removal as done in two-stream networks. In reference [15], a GCN with channel attention is used to process skeletal joint data by employing a graph pooling

method. The SGP framework makes use of the skeleton structure to guide convolutions, employing kernel receptive regions to detect specific body movements and improve motion feature extraction in GCNs.

In [16], the method employed the context stream at half resolution and the fovea stream that remains sharp at full resolution in the very center of the images, learning patterns with Early, Late, and Slow Fusion with CNNs. Singh et al. [17] proposed the deeply interconnected ConvNet with the bidirectional LSTM to recognize actions from the RGB videos. The early layers are trained with separate DMI, the later layers with features from the videos, and the late fusion with SoftMax combined the features from the videos. The evaluation was done on four datasets of RGB-D videos, including single and multiple person actions, demonstrating the importance of combining spatial and temporal features in human action recognition. The deep features from the videos, when combined, are able to capture the complexity of the actions in the videos.

III. METHODOLOGY

Deep learning models have been found to be effective in human activity classification in HAR systems. Various deep learning models were investigated to understand their ability to learn features and classify human actions. Training deep models requires significant computational power and large datasets, considering that HAR has a large set of features. The large set of features in HAR makes ImageNet, with one million images, a better choice for transfer learning. The study uses various pre-trained transfer learning models to classify human actions and compare them with the current stateof-the-art techniques.

Figure 2 illustrates the Human Action Recognition framework using a pre-trained deep learning model.

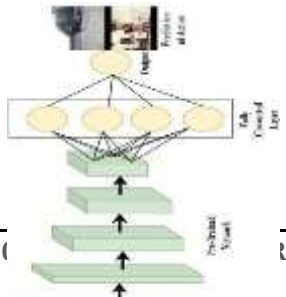

Fig. 2. HAR using pre-trained DL method

The DenseNet was selected because it has been shown effective in controlling vanishing or exploding gradients, as well as being effective in the reuse of features, as the features learned in any layer can be passed directly into the following layer. For the proposed transfer learning-based HAR framework, the VGG network was selected, as it is deeper and uses 3x3 filters. To address the known instabilities associated with the VGG network, batch normalization has been added, which will help keep the gradients stable.
The EfficientNet network was selected as a baseline network to measure the efficiency of the proposed framework.

A. DenseNet

DenseNet is a feed-forward network, and all the previous layer maps are connected directly to the next layer.

B. VGG

VGG is a CNN used in transfer learning, which is used in this proposed framework.
The image size is set at 224x224x3, and the features are learned using a series of 3x3 convolutions.

The proposed network has five max-pooling layers for dimension reduction.

Figure 3 represent the VGG19 architecture where conv belongs to convolution layer, pool represent pooling layer and FC belong to fully connected layers

Fig. 3. VGG19 Architecture

C. EfficientNet

EfficientNet is a CNN that focuses on efficient feature extraction by scaling, optimizing speed, efficiency, and accuracy. It can be used as a spatial feature extractor for classification tasks. It ranges from EfficientNet-B0 to B7, with EfficientNet-B0

outperforming ResNet-50 with fewer parameters and FLOPs.

D. Dataset

The UCF50[23], proposed by Reddy et al. (2012), tests the model using YouTube-like videos. The dataset adds 50 action classes (basketball, shooting, tabla, biking, violin, etc.) to UCF11, with 6,618 videos across 25 groups for each class.

IV. DISCUSSION AND RESULTS

Three pre-trained networks were used to classify each activity: DenseNet, VGG19, and EfficientNet. The underlying philosophy behind this was transfer learning, wherein pre-trained networks were used to perform a different task, thereby reducing the training time. The UCF50 dataset was used to validate how effectively this detects human actions. Transfer learning is basically using pre-trained networks on a new task, and this has been successful in this case.

Fig. 4. UCF50 Action Dataset Frames.

The action set is grouped by type of activity. In this paper, we tested various deep learning models on this set and compared them with existing state-ofthe-art techniques. We started by extracting frames from each action video and passing them through pre-trained models. The confusion matrices for 50 actions in the UCF50 set are given in Figures 5, 6, and 7 for VGG19, DenseNet161, and EfficientNetB7 models, respectively.

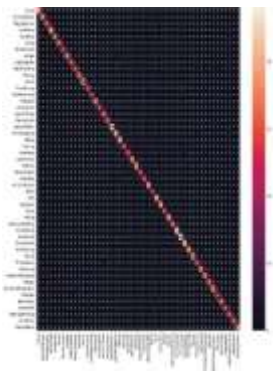

Fig. 5. VGG19 model confusion matrix for action recognition

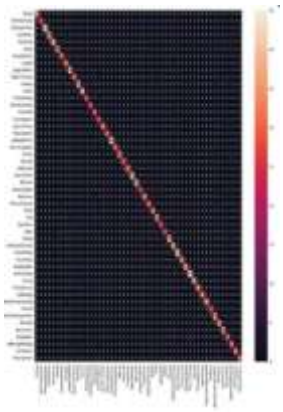

Fig. 6. Utilizing Dense Net 161 model, a confusion matrix for action recognition

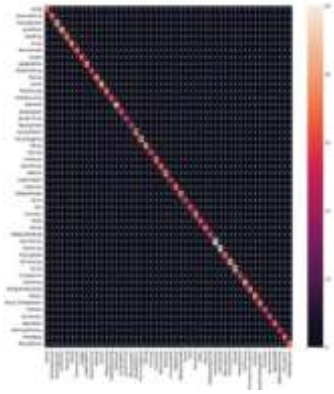

Fig. 7. Confusion matrix for action prediction from Efficient Net b7 model.

The performance of the proposed model was evaluated using the UCF50 dataset, and as depicted in Table 1, the evaluation metrics are stacked up using the transfer learning techniques. During the implementation stage, the extracted frames are split into training, validation, and test sets as depicted in Figure 8.

Table 2 compares the proposed method with the existing best techniques.

Fig. 8. Comparison graph for evaluation metrices.

TABLE II. COMPARISON OF LIGHT WEIGHT DL METHOD WITH EXISTING APPROACH.

Researcher	Dataset	Accuracy (%)
L. Zhang et al[24]	UCF50	88.0
H. Wang et al[25]	UCF50	89.1
Q. Meng et. al[26]	UCF50	89.3
Ahmad Jalal et. al[27]	UCF50	90.48
VGG19_bn	UCF50	90.11
Dense Net 161	UCF50	92.57
Efficient Net_b7	UCF50	94.25

We implemented the non-transfer learning approaches and compared their performance with our model on the UCF50 dataset. From the confusion matrix, it is observed that the accuracy of activity recognition is increased by 1-4% using the pre-trained deep models. All activities are recognized with high confidence.

V. CONCLUSION

The pre-trained and tuned model on action categorization using the UCF50 dataset (50 classes, 25 groups, at least four videos per group) was tested using measures such as precision, recall, F1 score, and AUC score. It utilized the VGG19, DenseNet161, and EfficientNet models in action recognition and was stacked against the bestperforming models from literature. It achieved high performance with the best accuracy achieved using the EfficientNet model at 94%.

Improvements can be made by adding an attention layer to combine the output from the BiLSTM with the action recognition output. Other areas of improvement can be seen in the real-time action monitoring, abnormal action detection, and crowd behavior analysis using the pre-trained model.

REFERENCES

[1] A comprehensive survey on video-based human action recognition by P. Pareek and A. Thakkar, including recent updates, datasets, challenges, and applications, is presented. The paper is published in Artificial Intelligence Review, Volume 54, No. 3, pages 2259-2322, DOI: 10.1007/s10462-

[2] A brief summary of the source: Archives of Computational Methods in Engineering (Vol. 29, No. 4, pp. 2309–2349, https://doi.org/10.1007/s11831-021-09681-9). Singh et al. provide an overview of the recent advances in the field of human action recognition during.

[3] A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z. Mahfouf, "Human activity recognition via optical flow: decomposing activities into basic actions," Neural Comput Appl, vol. 32K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos.”

[4] The work of S. Ji, W. Xu, M. Yang, and K. Yu explores the 3D Convolutional Neural Networks in the context of human action recognition. This work is published in the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) journal in 2013. Volume 35, Issue 1, pp. 221-231. DOI: 10.1109

[5] Gu, F., Khoshelham, K., and Valaee, S., focuses on the activity recognition of locomotion using the deep learning technique. This paper was published at the IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC) in February 2018, pp. 1–5, doi:10.1109/PIMRC

[6] In a paper published in MATEC Web of Conferences in 2019, Aubry, Laraba, Tilmanne, and Dutoit investigated action recognition using 2D skeletons extracted from RGB videos. The paper is found in volume 277, article 02034, with the DOI 10.1051/matecconf.

[7] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172– 186, January 2021, doi: 10.1109/TPAMI.2019.2929257.

[8] C. Dai, X. Liu, and J. Lai, “Human action recognition using two- stream attention-based LSTM networks,” Applied Soft Computing Journal, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105820.

[9] Y. Du, W. Wang, and L. Wang, “Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition.”

[10] M. Majd and R. Safabakhsh, “Correlational Convolutional LSTM for human action recognition,” Neurocomputing, vol. 396, pp. 224–229, Jul. 2020, doi: 10.1016/j.neucom.2018.10.095