ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

A Compact Deep Learning Model for Identifying Human Movements in Videos

[1]Mrs. A Josh Mary, [2]Narava Tarun Eshwar,[3]Rathipalli Veerendra Kumar, [4]Kotari Shyam,

[1]Associate Professor, Department of CSE, Rajamahendri Institute of Engineering & Technology, Bhoopalapatnam, Near Pidimgoyyi, Rajahmundry, E. G. Dist. A.P 533107.

[2,3,4]Student, Department of CSE, Rajamahendri Institute of Engineering & Technology, Bhoopalapatnam, Near Pidimgoyyi, Rajahmundry, E. G. Dist. A.P 533107.

Abstract

Many researchers in the field of computer vision have lately focused on Human Action Recognition (HAR) from a visual stream. Because it has many potential uses, such as health monitoring, home automation, and virtual reality, to name a few. But it still has to deal with occlusion, lighting fluctuations, complex backdrops, and human differences. Proper execution of learning data and features gathering technique are essential to the assessment criteria. Among the many remarkable products of Deep Learning's (DL) success story are neural networks. To be sure, a reliable classifier can't assign a label without a strong features vector. Data sets cannot be complete without features. The computational cost and performance of the method may be impacted by feature extraction. To construct this study framework, we extracted features from the picture sequence using the pre-trained deep learning models VGG19, Dense Net, and Efficient Net, and we categorized each action using the SoftMax layer. The 50-section UCF50 action dataset was used for performance evaluation, which was done using f1-score, AUC, precision, and recall. Model testing accuracy was 90.11 for VGG19, 92.57 for DenseNet, and 94.25 for EfiicientNet.

Keywords—Transfer Learning, CNN, VGG19, UCF50 I.

INTRODUCTION

According to HAR, an action is anything that can be perceived by either the naked eye or a sensor. Actually, paying close attention to someone in one's peripheral vision is essential for activities like walking. Based on the parts of the body that are required to carry out an action, we may classify them into four distinct types. [1]: Yes. Gesture: It relies on the expressions made on the face. Do not need any kind of verbal or physical contact. Human activity included walking, playing, and punching. Humans and inanimate objects engage in connection via gestures such as hugs and handshakes. When more than two things are occurring at once, such as a mix of gestures and interactions, we say that there is group activity. The performance of an action requires the participation of at least two actors. When it comes to computer vision research, HAR has been a mainstay for the last 20 years. The goal of HAR is to identify and detect actions done by one or more people by analyzing a set of observations. This may be done for a variety of individuals. The field of human-computer interaction was born out of this need. This area of study attracts scholars from all around the world because of the vast variety of possible applications. Among its many notable uses are surveillance video, image categorization and retrieval, health monitoring, automation, and environmental modeling [1]. There are three tiers to the inherent hierarchical structure of human activity, which symbolizes their various levels. The most fundamental building block is the atomic element, and the more complex human behaviors are represented by these action primitives. The action primitive level is the first level, while the actions/activities level is the second. Complex interactions reflect the highest degree of human activity classification. A separate field of study is necessary for each of these categories due to their vast nature. This is mainly because real-life human actions are unpredictable and ambiguous. There are a number of challenges that HAR must overcome. Some examples include gender bias, inconsistent results across classes, and interactions involving more than one topic. There are four steps to the process of human activity recognition from videos. As a first stage, we take picture sequences and use them to extract features. There are a variety of handmade approaches that may be utilized in feature extraction, such as SIFT (scale invariant feature transform), SURF (speed up robust feature), shape-based, pose-based, optical flow, and many more [1]. The use of deep learning allows for the extraction of features. This method allows the model to

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

autonomously learn all features from picture sequences. Pose and gesture patterns may be extracted from video sequences and frames showing people going about their daily lives. Size variations, bad lighting, wrong perspectives, and background clutter all contribute to making this an arduous process. The next step involves using the collected characteristics to train and identify actions. Important components of action learning and recognition include learning new models taught by extracted data, identifying which features are related to which action classes, and evaluating those features using classifiers. Some of the most well-known ways to address the HAR problem are the DL method and the Machine Learning (ML) methodology. In the first, more traditional version of AI, the user is still involved in the process of designing, dictating, and fine-tuning the attributes that are retrieved and how actions are described. On the other hand, we expect the DNN to perform better when we use the second approach. The second method relies on the assumption that the DNN can mimic human intelligence and solve all qualities automatically [1][2]. Figure 1 shows the results of HAR base categorization using ML and DL.

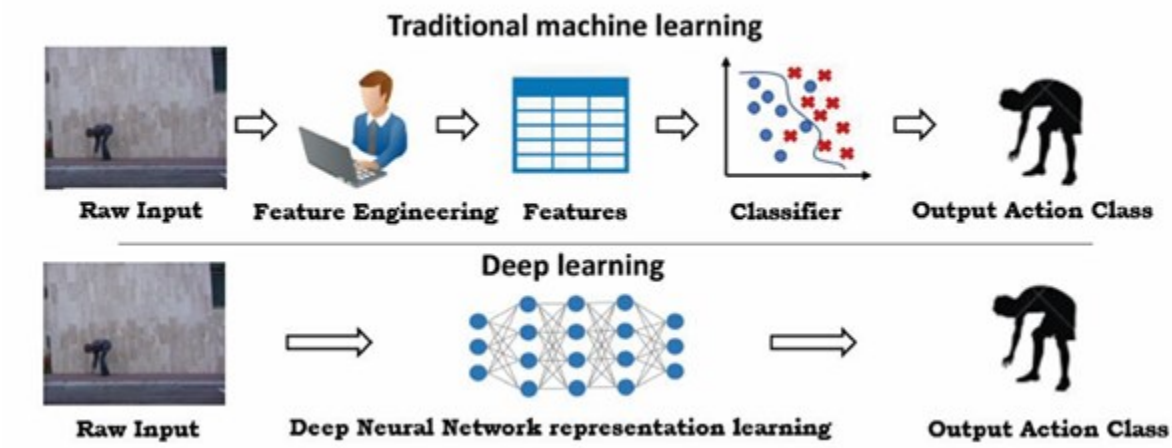

Fig. 1. A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR [2].

There have been decades of attempts to address the HAR issues related to it, such as the clutter backdrop, noise problem, and class similarity issue, using ML base approaches like random forest (RF), Bayesian networks (BN), Markov models (MM), and support vector machine (SVM). Even when faced with very little data inputs and severe constraints, seasoned ML algorithms have shown remarkable performance. The preprocessing stage of machine learning algorithms using handmade features is tedious and requires specific care; these algorithms also need to improve their performance. For really large datasets. DL has made significant progress in recent years. The reason for this is the impressive track record of deep learning research in several domains, including object identification in frames, action recognition, frame categorization, and natural language processing, among others. With its structure proven successful for both unsupervised and reinforced learning, DL drastically cuts down on the work needed to choose the right features when compared to typical ML algorithms. This is all because to the unmanned features pulled out via several hidden layers. This has led to an increase in the number of HAR frameworks that rely on deep learning. This section serves as an overview of the research article: Following a brief introduction to the topic, we go into the specifics of human action recognition using machine and deep learning methods. Then, in Section 2, we shall discuss the degrees of accuracy and modalities of the previously suggested approaches to human action recognition. Section 3 details the procedures and outcomes when applied to the dataset. The current state of computer vision research and its anticipated future directions are reviewed in Section 4.

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

RELATED WORK

There has been a lot of progress in the field of human action recognition. There is a lot of room to enhance the prediction of human activity because of its diverse range of applications. In order to identify human activity in images, many feature-based methods have been developed in the last ten years, some of which rely on human production while others rely on automated learning. Handcrafted features were the backbone of earlier approaches to human activity identification, which tended to focus down on insignificant atomic actions [3]. The main drawback of these approaches is the significant data preparation required and the difficulty in generalizing them to reality, even though they provide an accurate model. Numerous spatiotemporal methods for video activity analysis have been developed since convolutional neural networks (CNNs) achieved success in text and visual classification; these algorithms are capable of autonomously training and classifying from unprocessed RGB video [4]. In order to extract spatial and temporal video data for action recognition, Shuiwang Ji et al.[5] presented a 3D convolution approach. Consequently, the proposed architecture uses the video sequence to generate many data channels, each of which is then subjected to its own set of processing operations, such as convolution and subsampling. For indoor navigation and localization, Gu et al. presented a DL-based approach to detect locomotive movements. They eliminated the need to manually construct important features by using stacked denoising auto-encoders that learned data properties automatically [6]. The suggested study framework boasts a higher level of accuracy compared to another classifier. By analyzing RGB (Color model) video, Aubry et al.[7] developed a novel approach to action detection. First, we need to remove the human skeleton from the movie by removing its motion. This extraction was done using Open Pose [8], a Deep Neural Network (DNN)-employee identification approach that extracts a 2-D skeleton from each body with 18 recognized joints. In the second case, an image classifier is used to transform motion patterns into RGB images. Motion data is stored in the R, G, and B channels. An action sequence RGB image is created in this way. Neural networks now used for picture classification may one day be trained to identify human behaviors as well. According to the dual stream model proposed by Dai et al.[9], which employs an attention-based long short-term memory (LSTM) structure to pinpoint the location of action in visual frames. They claimed to have found a solution to the issue of visually ignoring attention.

Using the UCF Sports dataset, the architecture achieved a 98.6% accuracy rate, the j-HMDB dataset a 76.3% accuracy rate, and the UCF11 dataset a 96.9% accuracy rate. A skeleton-based approach to action recognition using a hierarchical RNN model was developed by Du et al. [10]. They also compared the five deep RNN designs that relied on their proposed approaches. They employed the HDM05 dataset, the Berkeley MHAD dataset, and the MSR Action-3D dataset throughout their examination. The Correlational Convolutional LSTM was developed by Majd and Safabakhsh[11] by adding spatial and motion information to an existing LSTM module and creating temporal links. Their results showed a 92.3% correctness rate and a 61.0% accuracy rate when tested on the popular UCF101 and HMDB51 benchmark datasets, respectively. A novel method for constructing a semantic RNN called stag-Net was proposed by Qi et al.[12] with the aim of identifying both individual and group activities. Using a structural RNN, they expanded their semantic network model to include time as a fourth dimension. Team efforts accounted for 90.5% of the Volleyball dataset, whereas individual efforts accounted for 8.5%. According to Huang et al. [13], a 3D convolutional neural network (ConvNet) is used to extract attributes based on posture by combining information about motion, 2-dimensional appearance, and 3-dimensional stance. We conduct convolution in each of the fifteen channels of the heatmap to decrease the noise, since the computationally demanding characteristics of color joints in frames achieved by 3-D convolutional neural networks (CNNs) are anticipated. In their work on Inception and Batch Normalization, Wang et al. [14] used the (BN-inception) network design. Similar to two stream networks, the previously described method uses RGB variation frames to mimic visual change and optical flow fields in combination with RGB and optical flow frames to prevent background motion. In [15], the author made use of a graph pooling network and a GCN with a channel attention method for joints. Last but not least, the SGP architecture enhanced convolution by including the human skeletal network. In order to get specific information about the human body, kernel receptive areas are used. While reducing calculation costs, the proposed SGP method has the ability to greatly enhance GCNs' ability to collect depending on motion characteristics. The study piece used context stream and fovea stream designs [16]. Frames are sent to the fovea channel at full resolution for the central area, whereas frames to the context channel are sent at half the original resolution. Using a series of short, fixed-length clips from each movie, the research trains a model to recognize three distinct pattern classes: Early Fusion,

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Late Fusion, and Slow Fusion. Through a variety of time-space combinations, CNN is able to generate single-frame animations. According to Singh et al. [17], a bidirectional-LSTM, a highly coupled ConvNet with RGB frames as the top layer, may be used to detect terms related to human activities. Individual DMI are used to learn (train) the bottom layer of the ConvNet model. While the higher layers of the pre-trained ConvNet are fine-tuned to extract temporal information from video streams, the ConvNet-Bi-LSTM model is trained from scratch for RGB frames to enhance the features of the pre-trained CNN. At the decision layer, features are fused using a late fusion approach that follows the SoftMax layer to get a better accuracy value. We use four RGB-D (depth) datasets that include both single-person and multiple-person activities to test the proposed model.

METHODOLOGY

The When it comes to activity categorization, the DL model for HAR reveals the significant outcome. We went over a few deep learning models, how they function, and how precisely they categorize each action. It takes a lot of processing power to train a deep learning model from beginning. Training learning models are superior than transfer learning models. Using ImageNet's massive dataset, they were trained [18]. There are about one million photos in ImageNet that can be used to build transfer learning models. In order to categorize each action, this study contrasted many transfer learning models with state-of-the-art approaches. Several transfer learning models for action recognition were examined in this study. Figure 2 shows the Human Action Recognition model using the deep learning model that has already been trained.

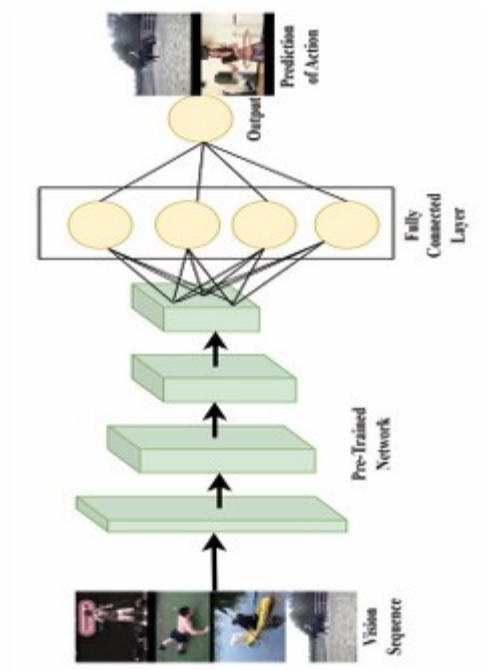

Fig. 2. HAR using pre-trained DL method.

Methods based on transfer learning (TL) are evaluated using Dense Nets [19]. Dense Net neural networks were selected because to their innovative methods for handling decreasing or increasing gradients and its unique architecture that enables a single layer to acquire knowledge from the feature maps of previous layers, enabling the reuse of features. Due to its very deep architecture, achieved by using small (33) filters, VGG[20] is also taught utilizing a transfer learning-based HAR technique. Gradient explosions are common in VGG models because of their intricacy. We used VGG models with batch normalization layers to regulate gradients and solve this issue. When assessing the efficiency of a framework, the Efficient Net[21] approach is also used. A. Narrow Net Dense Net is the name given to a kind of Convolution neural network that is very linked due to the fact that it employs a feed-forward way to connect each successive network layer. After passing through a large-filter-size Conv2D layer, the data is sent via a dense block that forms dense connections with every subsequent layer. New inputs are received by each Dense Net layer from all levels below it, and feature-maps are broadcast to all layers above it. A. VGG We further included VGG [20], a CNN architecture, into the TL-based method for action recognition. The images that are sent into VGG for training are pixel-perfect 512 × 512 images

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

(224, 224, 3). A series of convolutional layers equipped with 3-by-3-pixel filters have been used for the purpose of processing these pictures. Spatial pooling is performed via five max-pooling layers following particular conv2D layers. Dense layers with complete connectivity and a SoftMax prognostication layer follow a stack of convolutional layers. Figure 3 shows the VGG19 design, which includes the convolution layer (conv), the pooling layer (pool), and the fully connected layer (FC).

Fig. 3. VGG19 Architecture

Section C: EfficientNet In order to ensure that the depth, breadth, and resolution parameters of convolutional neural networks are scaled uniformly, Efficient Net[21] use a compound coefficient. Efficient Net scaling uses a set of specified scaling factors to evenly modify the breadth, depth, and resolution of the network, as opposed to the existing method that randomly scales these parameters. A unique convolutional neural network (CNN) with fast and efficient parameter estimation is Efficient Net [21]. By evenly scaling network properties including depth, breadth, and resolution, Efficient Net [21] was able to systematically scale up CNN models using a simple and difficult scaling technique. Efficient Net [21] was also used as a network for extracting spatial features in classification applications. The Efficient Net family included seven convolutional neural network (CNN) models, numbered EfcientNet-B0 through EfcientNet-B7. Results showed that EfcientNet-B0 could efficiently extract features, as it beat Resnet-50[22] with fewer parameters and higher FLOPs (floating-point operations per second) accuracy, all with the same input size. Database (D) The model's performance was evaluated using the UCF50[23] dataset. In 2012, Reddy et al. suggested this dataset. Video sharing websites like YouTube are used for the purpose of collecting videos. Nothing in these films is staged; instead, they all feature natural settings. The UCF11 dataset has been superseded by this one. It has fifty activity lessons including shooting, bicycling, shooting, shooting, playing the tabla, violin, etc. All things considered, there are 6618 films covering everything from basic sports to mundane daily life. There are twenty-five consistent categories for each activity type, with four films allocated to each category. Identical characters, settings, or points of view are common in films that fall under the same genre. The action snippets from the UCF 50 dataset are shown in Figure 4. Basketball Discus Throw Swimming, Tennis, Swing, Drumming, and Punching

Fig. 4. UCF50 Action Dataset Frames. action dataset,

This has many picture sets. Using this strategy, we evaluated the accuracy of multiple deep learning models on the aforementioned dataset in comparison to state-of-the-art approaches. Each set of action films had their frames removed and then put into a deep learning model that had already been trained. Confussion matrices for 50 activities recognition from the UCF 50 dataset employing the VGG19 model, Dense Net 161, and EfficientNet b7 are shown in Figures 5-7.

DISCUSSIONANDRESULTS

We utilized Dense Net, VGG19, and Efficient Net, three pre-trained deep learning models, to categorize each action. Using pre-trained deep learning, we were able to put the data from large datasets like ImageNet to good use. The idea behind the transfer learning method is to train a neural network for a new domain by transferring data from an existing model that has already been trained. The UCF50 was evaluated.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

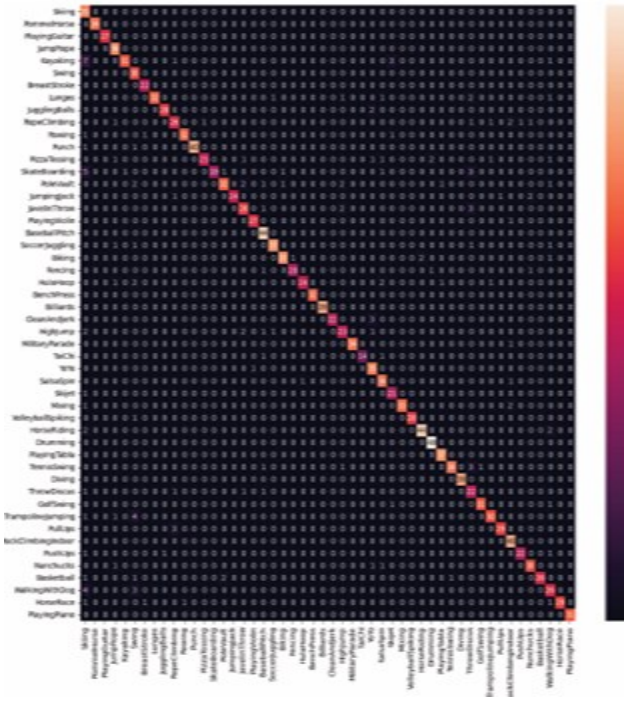

Fig. 5. VGG19 model confusion matrix for action recognition

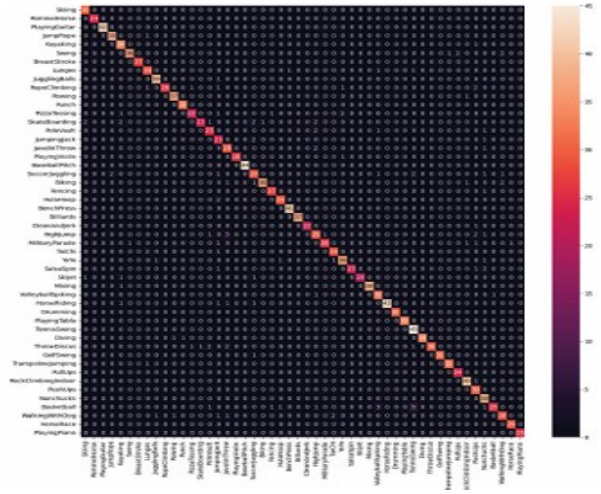

Fig.6. Utilizing Dense Net 161model, a confusion matrix for action recognition.

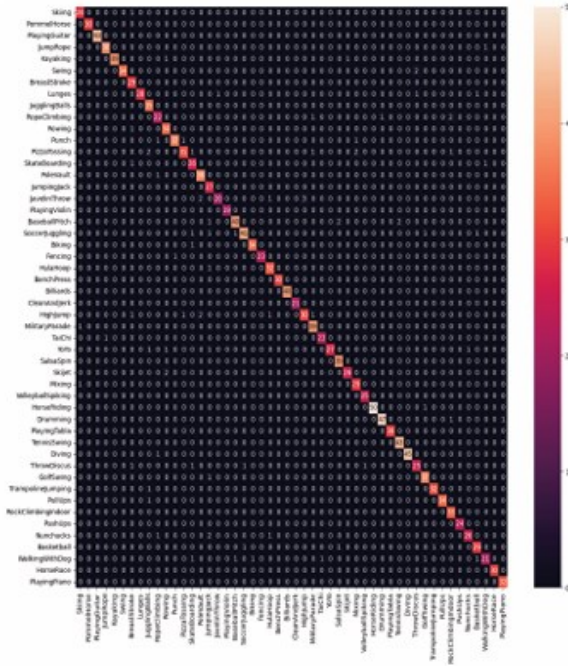

Fig.7. Confusion matrix for action prediction from Efficient Net b7 model.

A confusion matrix displays the classification result on the UCF 50 activity dataset. With a high degree of certainty, the majority of the actions are classified. Table 1 compares model assessment metrics using TL approaches on the UCF50 action dataset. During the implementation phase, the recovered frames were partitioned using the training, validation, and testing stages. An example of this is shown graphically in Figure 8. In Table 2 we may see a comparison with several state-of-the-art methods:

TABLE I. COMPARISON OF VARIOUS LIGHT WEIGHT DLMETHOD.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG19	90.11	91.92	90.34	90.53
Dense Net 161	92.57	93.06	92.45	92.43
Efficient Net b7	94.25	94.92	94.79	94.71

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

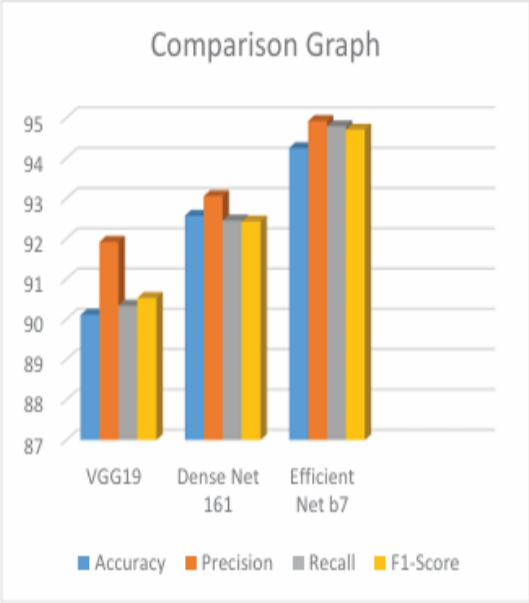

Fig.8. Comparison graph for evaluation metrices.

TABLE II. COMPARISON OF LIGHTWEIGHT 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 compared the efficiency of our technique to that of many non-transfer learning approaches on the UCF 50 dataset. The experiment's results showed if the recognition score increased while using transfer learning on a comparable dataset. While using pre-trained deep learning, their classification performance is improved by 1-4 percent.

CONCLUSION

The UCF 50 action dataset is used to develop deep learning algorithms that can categorize human actions. There are a total of fifty action categories in the UCF50 action dataset, organized into twenty-five groups. Each group has four movies. Precision, recall, f1 score, and AUC score were some of the assessment matrices utilized to measure the efficacy and accuracy of the model. Three models—VGG19, Dense Net 161, and Efficient Net—categorize the dataset's actions. Additionally, this study contrasted cutting-edge approaches that were used with the UCF50 dataset. The performance of these pre-trained deep learning models surpasses that of state-of-the-art approaches. With a 94% accuracy rate, Efficient Net outperforms competing pre-trained deep learning models. Adding more datasets, real-time action tracking, detection of anomalous actions, and crowd behavior classification are all possible future directions for this study. The next step in this study is to modify the pre-trained deep learning model's architecture, for example by adding an attention layer, so that it may be used with Bi-LSTM.

REFERENCES

[1]. P. Pareek and A. Thakkar, “A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications,” Artif Intell Rev, vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi: 10.1007/s10462-020-09904-8.

[2]. P. K. Singh, S. Kundu, T. Adhikary, R. Sarkar, and D. Bhattacharjee, “Progress of Human Action Recognition Research in the Last Ten Years: A Comprehensive Survey,” Archives of Computational Methods in Engineering, vol. 29, no. 4, pp. 2309–2349, Jun. 2022, doi: 10.1007/s11831-021-09681-9.

[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. 32, no. 21, pp. 16387–16400, Nov. 2020, doi: 10.1007/s00521-018-3951-x.

[4]. K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos.”

[5]. S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional neural networks for human action recognition,” IEEE Trans Pattern Anal Mach Intell, vol. 35, no. 1, pp. 221–231, 2013, doi: 10.1109/TPAMI.2012.59.

[6]. F. Gu, K. Khoshelham, and S. Valaee, “Locomotion activity recognition: A deep

learning approach," in IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, Feb. 2018, vol. 2017-October, pp. 1–5. doi: 10.1109/PIMRC.2017.8292444.

[7]. S. Aubry, S. Laraba, J. Tilmanne, and T. Dutoit, "Action recognition based on 2D skeletons extracted from RGB videos," MATEC Web of Conferences, vol. 277, p. 02034, 2019, doi: 10.1051/matecconf/201927702034.

[8]. Rao, S. Govinda, R. RamBabu, BS Anil Kumar, V. Srinivas, and P. Varaprasada Rao. "Detection of traffic congestion from surveillance videos using machine learning techniques." In *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 572-579. IEEE, 2022.

[9]. Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 1, pp. 172 186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[10]. 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.

[11]. 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.

[12]. M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. van Gool, "StagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 549–565, Feb. 2020, doi: 10.1109/TCSVT.2019.2894161.

[13]. Y. Huang, S.-H. Lai, and S.-H. Tai, "Human Action Recognition Based on Temporal Pose CNN and Multi-Dimensional Fusion."

[14]. [Wang Limin et al., Computer Vision – ECCV 2016, vol. 9912. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319 46484-8.

[15]. Y. Chen et al., "Graph convolutional network with structure pooling and joint-wise channel attention for action recognition," Pattern Recognit, vol. 103, Jul. 2020, doi: 10.1016/j.patcog.2020.107321.

[16]. A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02078

[17]. T. Singh and D. K. Vishwakarma, "A deeply coupled ConvNet for human activity recognition using dynamic and RGB images," Neural Comput Appl, vol. 33, no. 1, pp. 469–485, Jan. 2021, doi: 10.1007/s00521-020-05018-y.

[18]. Agrawal, K. K. ., P. . Sharma, G. . Kaur, S. . Keswani, R. . Rambabu, S. K. . Behra, K. . Tolani, and N. S. . Bhati. "Deep Learning-Enabled Image Segmentation for Precise Retinopathy Diagnosis". *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 12s, Jan. 2024, pp. 567-74, https://ijisae.org/index.php/IJISAE/article/view/4541.

[19]. Samota, H. ., Sharma, S. ., Khan, H. ., Malathy, M. ., Singh, G. ., Surjeet, S. and Rambabu, R. . (2024) "A Novel Approach to Predicting Personality Behaviour from Social Media Data Using Deep Learning", *International Journal of Intelligent Systems and Applications in Engineering*, 12(15s), pp. 539–547. Available at: https://ijisae.org/index.php/IJISAE/article/view/4788

[20]. O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int J Comput Vis, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[21]. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Nov. 2017, vol. 2017-January, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[22]. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556