# A Compact Deep Learning Model for Identifying Human Movements in Videos

[1] P. Premchand, [2] B. Vinisha,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

## Abstract

In the field of computer vision, Human Action Recognition (HAR) from a visual stream has lately garnered a lot of attention from researchers. Because it has several potential uses, such as health monitoring, home automation, and teleimmersion. Nonetheless, it continues to encounter challenges like as human variation, occlusion, lighting variations, and intricate backdrops. 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. The backbone of every data collection is its features. The computational cost and performance of the method can be impacted by feature extraction. In this study framework, we employed the SoftMax layer to categorize each action and pre-trained deep learning models VGG19, Dense Net, and Efficient Net to extract features from the picture sequence. The UCF50 action dataset was utilized; it is 50 sections long and uses f1-score, AUC, precision, and recall to measure performance. Models acquired VGG1990.11, DenseNet-92.57, and EfiicientNet-94.25 for accuracy testing.

## Keywords—

Subjects: UCF50, VGG19, CNN, Transfer Learning

## I. INTRODUCTION

What constitutes an action in HAR is its ability to be seen by either the naked eye or a sensor. Keeping one's focus on an object in one's line of sight is essential for activities like walking. Based on what parts of the body are required to carry out an action, we may classify them into four distinct types. [1]. x Gesture: expressions on the face serve as its foundation. No verbal or physical means of communication are required. Human activity included walking, playing, and punching. x Interaction: It includes not just human-object interactions but also human-to-human interactions such as embracing and handshakes. x Group activity: Combinations of two or more actions, such as gestures and interactions, are called group activities. The performance of an action requires the participation of at least two actors. Research into computer vision has been more dependent on HAR within the past 20 years. A person or people's actions can be detected and identified using HAR, which is based on a set of observations. This may be done for a variety of individuals. As a result, there was a growing demand for advancements in human-computer interface. This area of study attracts scholars from all around the world because of the vast variety of possible applications. Some of its most common uses are environmental modeling, automation, health monitoring, visual labeling and retrieval, and surveillance video[1]. There is an inherent hierarchical structure to human actions, which indicates their numerous levels. These levels may be broadly classified into three types. To begin, at the most fundamental level, there is an atomic element; these action primitives stand in for the more complex human acts. As a second level, the actions/activities level follows the performance primitive level. In terms of human activity classification, complex interactions constitute the highest degree. A separate field of study is necessary for each of these categories due to their vast nature. The main cause of this is the inherent vagueness and unpredictability of human behavior in the actual world. 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 subject. There is a four-step procedure for human activity recognition from videos. Feature extraction from provided picture sequences is the initial stage. Some examples of handmade approaches that may be utilized in feature extraction include SIFT (scale invariant feature transform), SURF (speed up robust feature), shape-based, pose-based, and optical flow [1]. The deep

learning approach is capable of feature extraction. The model in this technique learns all features automatically from picture sequences. Pose and gesture patterns may be extracted from video sequences and frames showing people going about their daily lives. Consequently, it is a difficult endeavor because of things like size variations, bad lighting, wrong views, and background clutter. Learning and recognition of actions based on extracted characteristics is the next level. A key component of action learning and recognition is learning new models that are instructed by extracted features. Other crucial steps include identifying which features are relevant to which action classes and evaluating them using classifiers. Notable methods for addressing the HAR problem include the Machine Learning (ML) approach and the DL method. 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 employ the second approach. Conversely, in the second method, we anticipate that the DNN will autonomously resolve all characteristics by mimicking the human brain's function [1][2]. Base classification for HAR using ML and DL is shown in Figure 1.

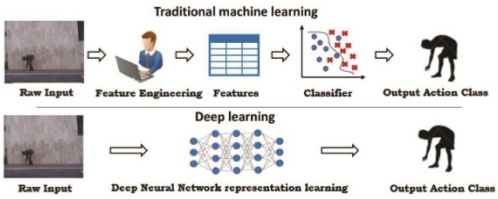


**Fig.1. A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR [2].**

There have been decades of attempts to address the HAR issues related to it, such as the clutter backdrop, noise issue, and class similarity issue, using ML base approaches like random forest (RF), Bayesian networks (BN), Markov models (MM), and support vector machine (SVM). Experienced ML algorithms have proven themselves capable of achieving remarkable results even when faced with very little data inputs and severe constraints. The preprocessing stage of machine learning algorithms with handmade features is tedious and requires specific care; these algorithms also need to improve their performance. If the amount of data is enormous. DL has made significant progress in recent years. The reason behind this is the impressive track record of deep learning research in several domains, including object detection 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 through several hidden layers. As a result, more and more HAR frameworks based on deep learning have been presented. A brief overview of the research article is as follows: First, we give a brief introduction to human action recognition, and then we look at how machine learning and deep learning approaches this problem. Afterwards, in Section 2, we shall discuss the methods and degrees of accuracy of earlier approaches to human action recognition. Section 3 delves into the methods and outcomes of the dataset. The current state of computer vision research and its anticipated future directions are reviewed in Section 4.

# II. RELATED WORK

It's done. Given its adaptability, there is a great deal of room to enhance the forecasting of human behavior. In order to identify people's movements in images, several feature-based methods have been created in the past ten years, some of which rely on human input and others on machine learning. Traditional methods for identifying human actions depended on subjective features, with an emphasis on insignificant atomic processes that don't appear to have any practical relevance whatsoever [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 a very accurate model. Algorithms that can automatically train and classify from raw RGB video have been developed for video activity analysis, building on the success of convolutional neural networks (CNNs) in text and visual classification [4]. To identify actions in videos, Shuiwang Ji et al.[5] presented a 3D convolution technique for extracting spatial and temporal data. 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 an aDL-based

approach to detecting 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. A novel method for identifying an action from RGB (Color model) video was developed by Aubry et al. [7]. In order to accomplish this, the motion in the film must be removed in order to extract the human skeleton. The process of extracting a 2-dimensional skeleton with 18 known joints from each body was carried out using Open Pose [8], a Deep Neural Network (DNN)-employee identification approach. In the second case, an image classifier is used to transform motion patterns into RGB images. The R, G, and B channels are utilized for the storage of motion data. An action sequence RGB image is created in this way. Neural networks now employed for picture classification may one day be trained to identify human behaviors as well. Dai et al.[9] proposed a dual-stream model that uses an attention-based LSTM structure to pinpoint where in a visual frame the activity is taking place. They claimed to have found a solution to the issue of visually ignoring attention. The accuracy of the architecture varied among datasets; it was 96.9% on the UCF11 dataset, 98.6% on the UCF Sports dataset, and 76.3% on the j-HMDB dataset. Using a hierarchical RNN model, Du et al. [10] developed a skeleton-based approach to action recognition. 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. By creating temporal links and adding spatial and motion information to an existing LSTM module, Majd and Safabakhsh[11] created the Correlational Convolutional LSTM. 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] for 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. With this strategy, teams were able to finish 90.5% of the volleyball dataset, whereas individuals only managed 8.5%. In the study by Huang et al. [13], features based on posture are derived from a 3D convolutional neural network (ConvNet) by combining data on motion, 2-dimensional appearance, and 3-dimensional stance. We apply convolution to each of the fifteen channels of the heatmap to lessen the noise, since we anticipate that the 3-D convolutional neural networks (CNNs) acquired color joint features in frames would be computationally demanding. In their works Inception and Batch Normalization, Wang et al. [14] employed the (BN-inception) network architecture. Similar to twostream networks, the previously described method mixes RGB and optical flow frames to prevent background motion, and RGB variation frames to mimic changes in appearance. In [15], the author made use of a graph pooling network and a GCN with a channel attention method for joints. The SGP architecture, which enhanced convolution and included a human skeleton network, was the last step. Specific information about the human body is retrieved by use of kernel receptive areas. While reducing calculation costs, the proposed SGP method has the ability to greatly enhance GCNs' ability to gather depending on motion characteristics. The study piece used context stream and fovea stream designs [16]. The resolution that the fovea channel receives is complete, whereas the context channel receives half of the original resolution. Training a model to recognize Early Fusion, Late Fusion, and Slow Fusion patterns from collections of tiny, fixed-length segments in each movie is the focus of the work. Through a variety of time-space combinations, CNN is able to generate single-frame animations. For the purpose of human activity recognition, Singh et al.[17] presented bidirectional-LSTM, a highly connected ConvNet with RGB frames as its top layer. Each DMI contributes to the learning process that forms the base of the ConvNet model. In order to enhance the pre-trained CNN's features, the ConvNet-Bi-LSTM model is trained from the ground up for RGB frames. In order to extract temporal information from video streams, the topmost layers of the pre-trained ConvNet are tweaked or fine-tuned. At the decision layer, features are fused using a late fusion approach that follows the SoftMax layer to get a better accuracy value. By utilizing four RGB-D (depth) datasets that encompass both single-person and multiple-person behaviors, we assess the efficacy of the proposed model.

# III. METHODOLOGY

An effect on activity categorization is demonstrated by the DL model for HAR. We went over a few deep learning models, how they function, and how precisely they categorize each action. Training a deep learning model from start necessitates a lot of processing power. Learning models, as opposed to transfer learning models, undergo training. Using ImageNet's massive dataset, they were trained [18]. For the purpose of training transfer learning models, ImageNet contains over 1 million photos. This study evaluated many transfer learning models with state-of-the-art approaches for action classification. 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.

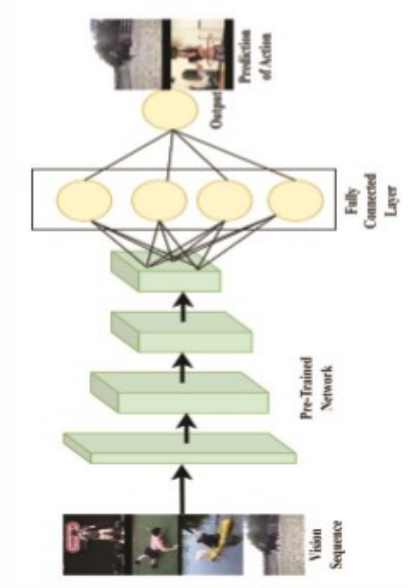


**Fig.2. HAR using pre-trained DL method**

In order to assess methods that rely on transfer learning (TL), Dense Net[19] is utilized. 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. The extremely deep architecture of VGG[20] is achieved by employing small (33) filters, and this is achieved via a transfer learning-based HAR method. 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 problem. Another tool for gauging the framework's efficiency is the Efficient Net[21] technique. The Dense Net The term "Dense Net" refers to the highly linked nature of a dense convolution neural network, which is defined as an architecture that employs a feed-forward way to connect each succeeding 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. Every Dense Net layer takes data from the ones below it and sends out feature maps to the ones above it. Class B. VGG Our TL-based method for action recognition additionally makes use of VGG [20], a CNN architecture. When training with VGG, the images must adhere to a certain ratio, meaning they must be 512 × 512 pixels (224, 224, 3). A series of convolutional layers equipped with 3-by-3-pixel filters have been employed for the purpose of processing these pictures. In the three conv2D levels that follow,
Spatial pooling is performed using five max-pooling layers. Next come dense fully connected layers, followed by a stack of convolutional layers, and finally a SoftMax prediction layer. The VGG19 architecture is shown in Figure 3, with the components conv, pool, and FC corresponding to the various layers.


**Fig.3. VGG19 Architecture**

EfficientNet Efficient Net[21] is a method for designing and scaling convolutional neural networks that uses a compound coefficient to uniformly scale the depth, width, and resolution parameters. The Efficient Net scaling method consistently modifies the depth, breadth, and resolution of the network based on a set of specified scaling parameters, as opposed to the existing approach that randomly scales these aspects. A unique convolutional neural network (CNN) with fast and efficient parameter estimation is Efficient Net [21]. To more methodically scale up CNN models, Efficient Net [21] uses a simple and difficult scaling methodology to consistently scale network features including depth, breadth, and resolution. As a spatial feature extraction network, Efficient Net [21] was also used in classification tasks. The Efficient Net family included seven convolutional neural network (CNN) models, numbered EfcientNet-B0 through fcientNet-

B7. Using the same input size, EfcientNet-B0 achieved better results than Resnet-50[22] with fewer parameters and higher FLOPs (floating-point operations per second) accuracy, suggesting that EfcientNet-B0 can efficiently extract features. Table D. The model's performance was evaluated using the UCF50[23] dataset. Reddy et al. (2012) first suggested this dataset. Videos are compiled using online sites such as YouTube. None of the videos were shot in a studio; 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 a minimum of four films assigned to each activity, and each class is further divided into 25 similar groups. Movies that fall under the same genre often share elements like characters, settings, or points of view. The action snippets from the UCF 50 dataset are shown in Figure 4.

## IV.DISCUSSION AND RESULTS

We employed three pre-trained deep learning models—Dense Net, VGG19, and Efficient Net—to categorize each action. To make the most of the data collected from large datasets like ImageNet, we used pre-trained deep learning. A neural network may be taught to handle new domains using the transfer learning method, which involves transferring data from a model that has already been trained. Test Results for the UCF50


**Fig.4. UCF50 Action Dataset Frames**

database of actions, which includes several categories of photos. Using this strategy, we evaluated the accuracy of multiple deep learning models on the aforementioned dataset in comparison to state-of-the-art approaches. To begin, a pre-trained deeplearning model was fed frames collected from each set of action footage. Using the VGG19 model, Dense Net 161, and EfficientNet b7, the confusion matrix for detecting 50 actions from the UCF 50 dataset is shown in Figures 5-7.

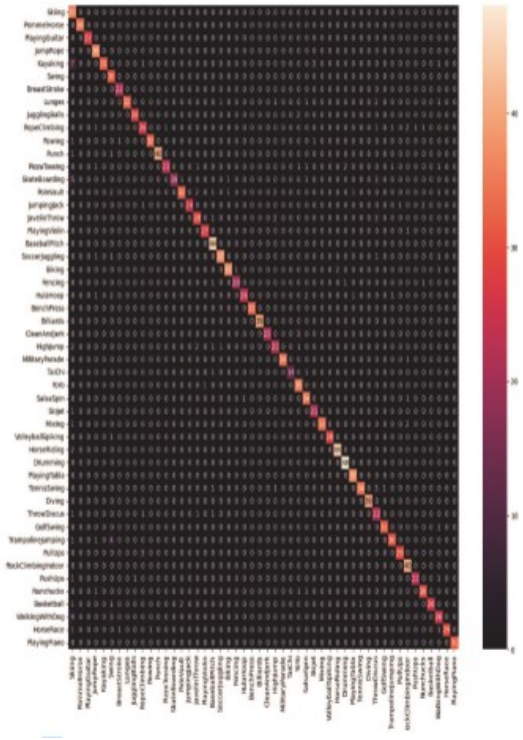

**Fig.5. VGG19 model confusion matrix for action recognition.**

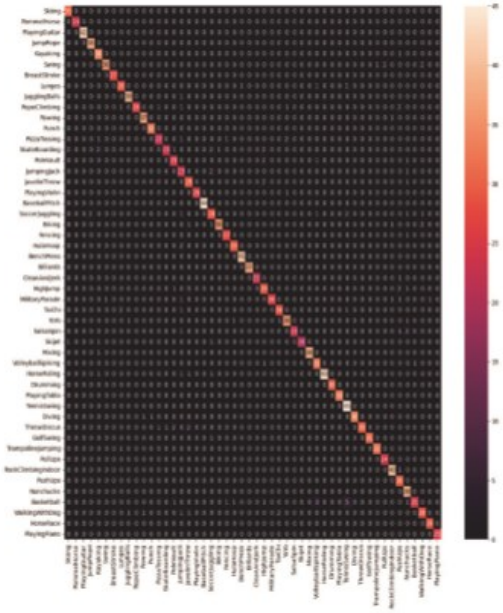

**Fig.6. Utilizing Dense Net 161model, a confusion matrix for action recognition**.

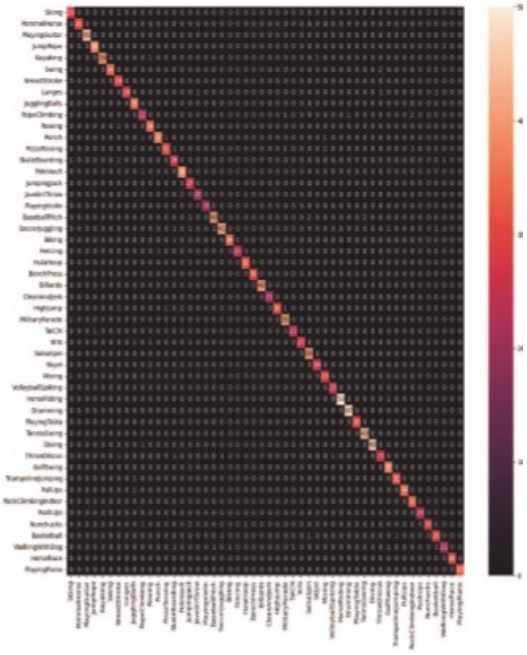

**Fig.7. Confusion matrix for action prediction from Efficient Net b7 model.**

Using the UCF 50 activity dataset, the classification result is displayed as a confusion matrix. We can confidently and properly categorize the majority of the actions. Model assessment metrics using TL approaches are compared in Table 1 on the UCF50 action dataset. The recovered frames were partitioned during the implementation phase using the training, validation, and testing stages. A visual representation of this may be seen in Figure 8. Table 2 displays comparisons with various 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 |

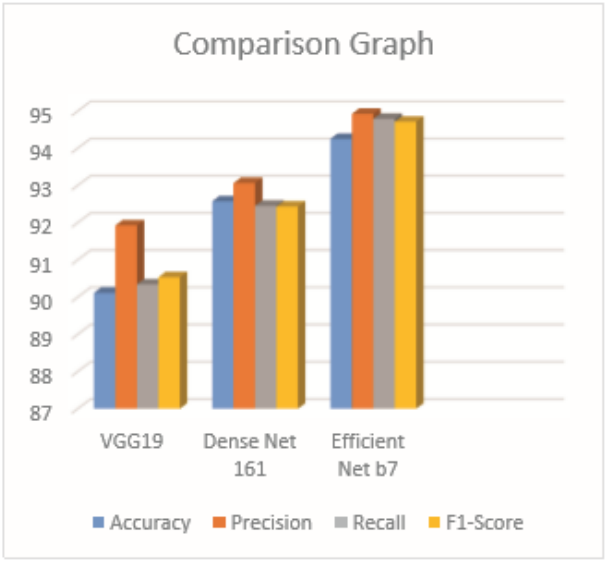


**Fig.8. Comparison graph for evaluation metrices.**

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 tested our method against several others that did not use transfer learning on the UCF 50 dataset to see how well it performed. The results showed if the recognition score was increased by applying transfer learning on a comparable dataset. After applying pretrained deep learning, their classification performance is improved by 1-4 percent. No. 87

# V. CONCLUSION

The UCF 50 action dataset is used to develop deep learning algorithms that can categorize human actions. The UCF50 action dataset is comprised of fifty distinct action categories, organized into twenty-five groups, with a minimum of four films per group. The accuracy and efficacy of the model were tested

using several evaluation matrices, such as precision, recall, f1 score, and AUC score. For every dataset activity, three models—VGG19, Dense Net 161, and Efficient Net—classify it. The UCF50 dataset was also used to compare state-of-the-art algorithms in this paper. When compared to cutting-edge techniques, our pretrained DL models outperform them. With a 94% accuracy rate, Efficient Net outperforms competing pre-trained deep learning models. We may build on this work to categorize actions in different datasets, monitor actions in real-time, identify aberrant actions, and study crowd behavior in the future. In order to make the pre-trained deep learning model operate with Bi-LSTM, this study modifies its structure, for example by adding an attention layer.

# 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] 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.

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

[10] Y. Du, W. Wang, and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition."

[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: 10.1007/978-3-31946484-8.

[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] 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.

[19] 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.

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

[21] M. Tan and Q. v. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016-December, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[23] K. K. Reddy and M. Shah, "Recognizing 50 Human Action Categories of Web Videos." [24] L. Zhang and X. Xiang, "Video event classification based on twostage neural network," Multimed Tools Appl, vol. 79, no. 29–30, pp. 21471–21486, Aug. 2020, doi: 10.1007/s11042-019-08457-5.

[25] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A Robust and Efficient Video Representation for Action Recognition," Int J Comput Vis, vol. 119, no. 3, pp. 219–238, Sep. 2016, doi: 10.1007/s11263015-0846-5.

[26] Q. Meng, H. Zhu, W. Zhang, X. Piao, and A. Zhang, "Action recognition using form and motion modalities," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 16, no. 1s, Apr. 2020, doi: 10.1145/3350840.

[27] A. Jalal, I. Akhtar, and K. Kim, "Human posture estimation and sustainable events classification via Pseudo-2D stick model and K-ary tree hashing," Sustainability (Switzerland), vol. 12, no. 23, pp. 1–24, Dec. 2020, doi: 10.3390/su12239814.