Hybrid HAR-CNN Model: A Hybrid Convolutional Neural Network Model for Predicting and Recognizing the Human Activity Recognition

[1]Venugopal Rao A, [2]Santosh Kumar Vishwakarma, [3]Shakti Kundu and [4]Varun Tiwari
[1,2]School of Computer Science and Engineering, Manipal University Jaipur, India.
[3]School of Engineering and Technology, BML Munjal University, Kapriwas, Haryana, India.
[4]Department of AI & ML, Manipal University, Jaipur, India.
[1]vengopal.229351007@muj.manipal.edu, [2]santosh.kumar@jaipur.manipal.edu
[3]shakti.kundu@bmu.edu.in, [4]drvaruntiwari2020@gmail.com

Correspondence should be addressed to Shakti Kundu : shakti.kundu@bmu.edu.in.

Article Info
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202404040
Received 18 September 2023; Revised from 26 January 2024; Accepted 16 February 2024.
Available online 05 April 2024.

Abstract – Human activity recognition (HAR) is an active research area in computer vision for the past several years and research is continuing in this field due to the unavailability of perfect recognition system. The human activity recognition system covers e-health, patient monitoring,assistive daily living activities, video surveillance, security and behavior analysis, and sports analysis. Many researchers have suggested techniques that use visual perception to detect human activities. Researchers will need to address problems including light variations in human activity detection, interclass similarity between scenes, the surroundings and recording setting, andtemporal variation to construct an efficient vision-based human activity recognition system. However, a significant drawback of many deep learning models is their inability to achieve satisfactory results in real-world scenarios due to the conflicts mentioned above. To address this challenge, we developed a hybrid HAR-CNN classifier aimed at enhancing the learning outcomes of Deep CNNs by combining two models: Improved CNN and VGG-19. Using the KTH dataset, we collected 6,000 images for training, validation, and testing of our proposed technique. Our research findings indicate that the Hybrid HAR-CNN model, which combines Improved CNN with VGG-19 Net, outperforms individual deep learning models such as Improved CNN and VGG-19 Net.

Keywords – Human Activity, Improved CNN, Deep Learning, Activity Recognition, Artificial Intelligence.

I. INTRODUCTION

The Human Activity Recognition system has many applications due to the ability to understand the complex scenes from the surveillance video, for example surveillance system deployed at the different public places like railway stations, airport, Banks, they all require recognition ofsuspicious or abnormal activities rather than normal activity [1-3]. For instance, railway surveillance system must be capable of recognizing some abnormal activities like “intentionally dropping the bag at railway station”, “push the security personals”, or “putting the bag in the trash bin” [4]. Similarly, banking surveillance system must also be able to recognize some unusual events in the surveillance video such as “person loitering outside the bank premises”, “face covering using mask” etc. Now, HAR system has become an active component for many different applications such as understanding the human behavior, health-care sector, patient monitoring system, human security application, virtual reality games, automatic videosurveillance systems, object detection and tracking [5, 6].

Some simple atomic activities are like eating or drinking in which there is an elementarymovement of the user’s body part. Apart from the simple activity, there exist some complexactivities in which multiple body part movement occurs for example shaking the hands, activitiesduring playing sports and fighting [7]. Since the degree of the complexity of an activity significantly depends on type of activity therefore, the purpose of the recognition system is to be able to completely recognize the activity, in any possible scenario [8].

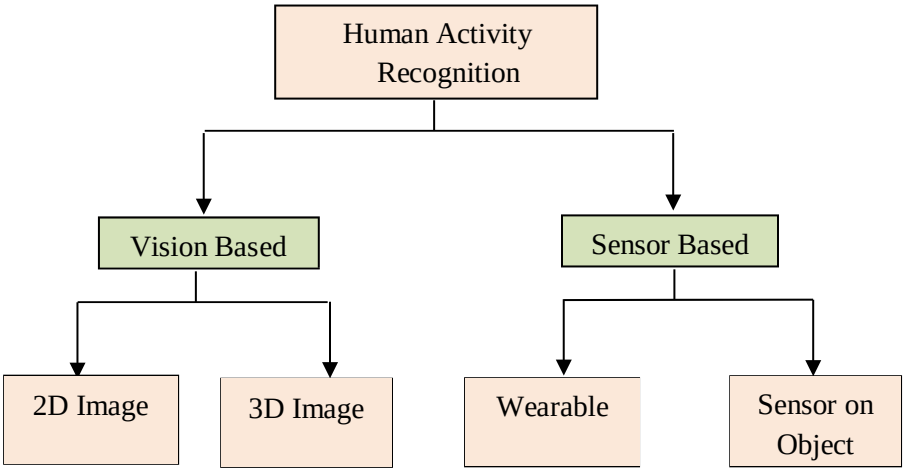

Fig 1. Types of Human Activity Recognition

However, it is very complicated task to do and researchers are trying to address this issue. The activities such as sitting, standing, walking, eating or drinking appears to be simple in nature however many complex challenges are still present, because the different people perform the same activity in the different way (intra- class dissimilarity) [9]. For example, "walking" activity of one person may be similar to the "running" activity of another person; likewise "sitting" activity of one person may be similar to "sleeping on chair" activity of another person. These challenges are come under "intra-class similarity" problem exist in the activity [10].

As shown in **Fig 1**, the human activity recognition may classified into two categories namely vision based HAR and sensor based HAR. The vision based HAR is possible with still images (2D images) or video sequences (3D images). Similarly, the sensor based HAR system isclassified into wearable sensor and sensor on object.

*Background of research Problem***:** Nevertheless, human activity recognition is a difficult problem in the field of machine learning, and many key difficulties, such as intra-class variation, changes in illumination, occlusion, actions that are similar, viewpoint variations, changes in scale,appearance, age, frame resolutions, and lighting conditions, remain unresolved.

This research paper comprises five main sections. **Section 2** conducts an in-depth review of existing literature on vision-based human activity recognition, highlighting research gaps and outlining the research objectives. The proposed Hybrid HAR-CNN model is elaborated upon in **Section 3. Section 4** delves into the experimental results pertaining to the Hybrid HAR-CNN model, comparing them with those of the enhanced CNN model and VGG-19 Net model. Finally, **Section 5** presents the findings, summarizes the study, and suggests future research directions.

II. LITERATURE SURVEY

Human activity recognition within video sequences stands as a prominent and steadily growing field within computer vision research, owing to its myriad applications across various domains such as safety, surveillance, healthcare, robotics, animations, sports analysis, content-based video summarization, behavioral analysis, and smart homes, among others. Over the past few decades, numerous feature-based approaches, both handcrafted and automatically learned, have been developed for this purpose, as depicted in **Figure 2**. Initially, human activity recognition methods relied heavily on handcrafted features, primarily targeting basic atomic actions, which often proved limited in practical applications. One notable drawback of these approaches lies in the intricate data preprocessing requirements and their struggle to generalize effectively in real-world scenarios, despite achieving high accuracy models. For instance, Bobick and Davis [11] pioneered the extraction of motion features from video sequences in the form of Motion History Images (MHI) and Motion Energy Images (MEI) temporal templates to identify human actions within static background conditions. Their focus was on specific human motion activities observed over time [37-38].

Shechtman and Irani [12] introduced a novel approach based on behavior-based similarity matrix templates for quantifying the similarity between human actions. Their method extended the concept of 2D image correlation into 3D space-time volumes, enabling the correlation of dynamic behaviors and actions. Similarly, Rodriguez et al. [13] proposed a template-based approach utilizing maximum average correlation height (MACH) filters to identify actions within videos. Notably, their model effectively tackles the challenge of intra-class variations while maintaining computational efficiency.

Chakraborty et al. [14] introduced a method for action recognition utilizing local interest points, termed Space-Time Interest

Points (STIPs), which extends the 2D Harris detector into a 3D corner detector. Their approach demonstrated promising results in representing STIP features, particularly in scenarios involving occluded backgrounds and variations in viewpoints. However, it's worth noting that their method is susceptible to camera motion, such as camera jitters. Alternatively, Willems et al. [15] proposed an extension of the 2D Hessian detector into 3D, presenting a technique for localized action detection based on second derivatives of the corner detector.

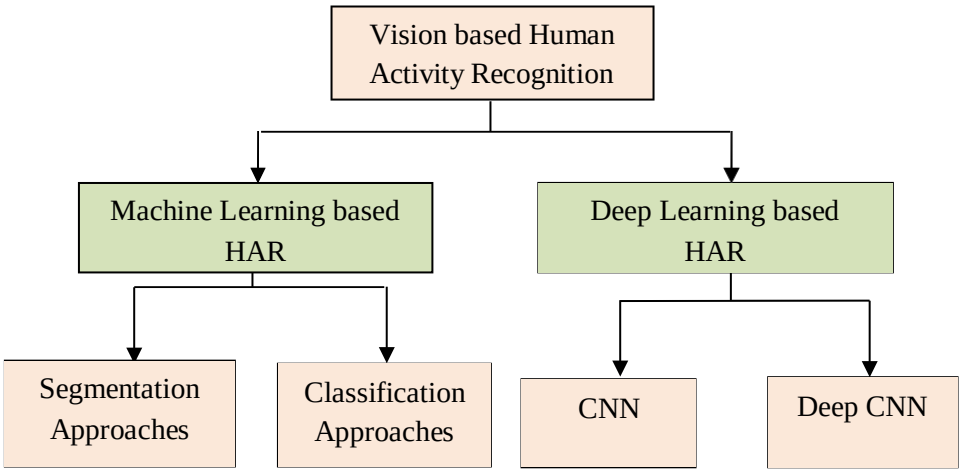

Fig 2. Different Kind of Vision Based Human Activity Recognition Approaches

A novel approach to automatically annotate the movie clips for training the action classifier, Laptev et al. [16] introduced the Histogram of Optical Flow (HOF) based spatio temporal An extension of the 2D Harris interest point detector has been proposed for action recognition in realistic videos, employing a descriptor. Additionally, a bag-of-features-based approach has demonstrated resilience against view variations, changes in illumination, and cluttered background conditions. Dalal et al. [17] devised a human pose descriptor utilizing the Histogram of Oriented Gradients (HoG) to identify actions in dynamic environmental settings. This method combines gradient features with a differential optical flow motion descriptor to represent human activities in realistic cinematic scenarios, yielding promising results across various challenging conditions [39-40].

Gaidon et al. [18] introduced the Actom Sequence Model (ASM), which extends the bag-of-features approach temporally to recognize action videos of varying lengths. Actoms are constructed based on sequences of atom units, with visual features represented as histograms of Actoms. Thurau and Hlavac [19] proposed a feature descriptor for action recognition based on human pose modeling. Their method utilizes Histogram of Oriented Gradient (HoG) on a designated region of interest (RoI) and represents feature vectors using non-negative matrix factorization.

Numerous deep learning models leveraging Convolutional Neural Networks (CNNs) have been devised by researchers, particularly for scene classification tasks. LeCun et al. [20,21] pioneered the initial CNN model, akin to a conventional Artificial Neural Network (ANN), laying the groundwork for contemporary CNN architectures. The structure of CNNs draws inspiration from the neurons found in animal and human brains. In recent times, researchers have expanded upon this foundation, creating a plethora of models tailored to address various image classification challenges.

For example, Karpathy et al. [22] presented four different fusion techniques along temporal dimension. They also presented slow fusion technique in which higher layers acquire more global information along both temporal and spatial dimension. The connectivity of all convolutional layers increased in time dimension by implementation and performing the time convolution. Yue-Hei Ng et al. [23] presented a method to handle a large sized video. The proposed approach was divided into two types, in the first type various convolutional temporal feature pooling techniques have been applied and in the second type a recurrent neural network such as long short-term memory network (LSTM) has been employed to process the ordered sequence of the input video. The presented approach was implemented on two popular datasets such as Sports 1M and UCF 101 datasets and achieved 12.2% (73.1% from 60.9%) and 0.6% (88.6% from 88.0%) improvement in the recognition accuracy with optical flow.

In another type of approach is to expand the convolutional operation along the time direction. The work is implemented by Ji et al. [24] presented a 3D convolutional network with the help of3D kernels or 3D filters that were extended along the temporal axis. These 3D kernels extract thefeatures from both space and time directions. To improve the performance of two-stream networks, Tu et al. [25] suggested a novel architecture called human-related multi-stream CNN (HR-MSCNN) architecture which combines human motion, appearance and human-related region altogether. They considered human-related region using improved foreground detection and region of interest corresponding to human motion.

Piergiovanni et al. [26] introduced a novel motion representation method, the convolutional layer inspired by optical flow, aimed at capturing motion within video sequences for activity recognition. This representation, termed the flow layer, is fully differentiable and effectively captures motion flow within the video sequence. The algorithm optimizes its parameters iteratively along with other model parameters to enhance activity recognition performance. Similarly, Simonyan et al. [27] proposed a two-stream ConvNet architecture, where one stream extracts appearance features from individual video frames, while the other stream learns motion information between frames using multi-frame dense optical flow. These streams are then fused at the softmax score level. Despite limited training data, the proposed two-stream network demonstrates robust recognition performance.

Identification of the gaps: Through a comprehensive literature review, various machine learning and deep learning algorithms have been explored for predicting human activity. This review has revealed several gaps in existing research.

- Traditional machine learning techniques relying on descriptors are observed to be insufficiently flexible in capturing the diverse variations present in video frames, such as changes in scale, viewing angles, and occlusions.
    - Moreover, conventional feature extraction methods often struggle to discern finer details within silhouettes, limiting their effectiveness.
    - Abrupt scene changes are frequently encountered in videos, presenting a challenge for activity recognition algorithms.
    - The variability in the speed of motion across different videos poses a challenge, as standard algorithms for extracting key poses frames may result in inaccuracies and lower prediction accuracy.

*Research Questions***:** Here are the challenging research questions addressed in this paper are as follows:

- How to effectively apply hybrid HAR-CNN model for human activity recognition.
- How to reduce the training time of hybrid human activity recognition (HAR-CNN) model.
- How to increase / improve the prediction accuracy of hybrid human activity recognitionmodel.

Research Methodologies: To conclude the ongoing study with the existing Human Activity Recognition (HAR) system, the research primarily focused on conducting secondary research utilizing established datasets referenced in contemporary methodologies. No unique dataset was developed specifically for this study. Additionally, previous literature on HAR was reviewed extensively to gain insights into the background work related to the topic at hand. This study proposes a solution utilizing deep learning techniques applied to the KTH dataset and assesses its performance using various statistical measures such as accuracy, precision, recall, and confusion matrix analysis. Consequently, the current study falls under the category of empirical research.

Objective of the Proposed Work

- The main objective of this research is to develop a hybrid HAR-CNN model for human activity recognition from static actions (2D images).
- The hybrid model is based on combining the features of the improved CNN model with VGG- 19 Net model.
- The hybrid model reduces the computational complexity by using the Global Average Pooling layer instead of Flatten layer.
- The hybrid model (I-CNN + VGG-19 Net) improves the prediction accuracy of various human activity recognition.
- All of the models that were discussed above are inefficient because it takes more computational time to train and evaluate the data using those models. We suggested a hybrid HAR-CNN model for the human activity recognition after taking into consideration the drawbacks that were discussed earlier.

III. PROPOSED WORK

The purpose and intention of this research is to concatenate the features of two deep learning models namely improved CNN and VGG-19 Net model for improving the prediction and recognition of human activities in effective manner. There are five primary parts to the human activity recognition system, which are: collecting dataset, extraction of features from improved CNN and VGG-16 Net respectively, then training, validate the model, finally predict the recognition results with combined model.

Primary Contribution of This Research Work

The most important findings from these studies are summarized in the following manner:

- A hybrid HAR-CNN model for predicting and recognizing the various human activities like walking, running, jogging, down stair and up-stairs.
- Combining the features of two different types of deep learning models—namely, aimproved CNN and a VGG-19 Net model—is known as feature concatenation or feature fusion.
- In order to reduce the parameters, we have used Global Average Pooling method (GAP layer) in Flatten.

Dataset Collection

The KTH dataset was established by the Royal Institute of Technology in Sweden in 2004, comprising six distinct human actions: walking, jogging, running, boxing, hand clapping, and hand waving. These actions were performed by 25 different actors across four unique scenarios, resulting in a total of 600 video sequences (25 actors × 6 actions × 4 situations). Recorded using a stationary camera and background, this dataset is often regarded as a straightforward benchmark for evaluating human activity recognition algorithms**. Fig 3** illustrates a single image exemplifying each action within one of six possible scenarios. Utilizing a split of 70% for training, 20% for validation, and 10% for testing, as detailed in **Table 1**, facilitates comprehensive model assessment.

Table 1. KTH Images Dataset Information

S. No.	Expression Type	Total no. of Images	Training Images	Validation Images	TestingImages
1.	Walking	1000	700	200	100
2.	Jogging	1000	700	200	100
3.	Running	1000	700	200	100
4.	Boxing	1000	700	200	100
5.	Hand waving	1000	700	200	100
6.	Hand clapping	1000	700	200	100

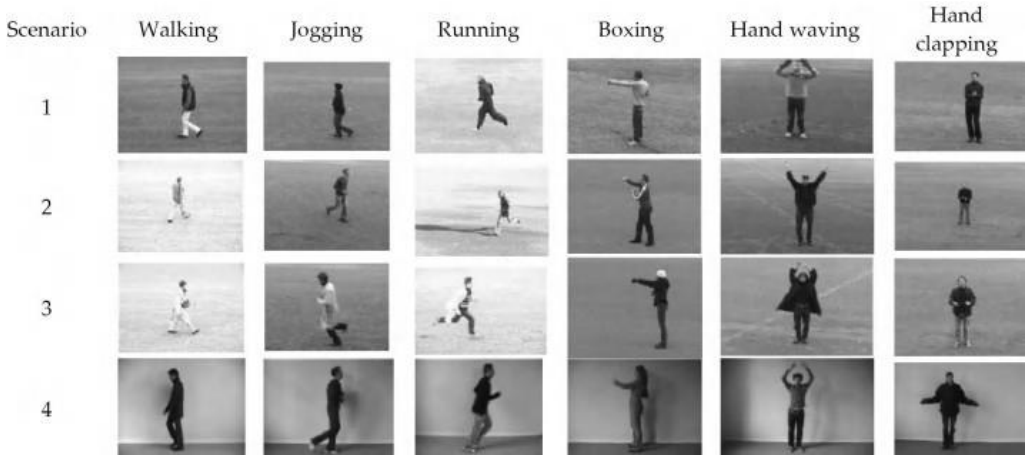

Fig 3. One Frame Example of Each Action in KTH Dataset

Improved CNN Model

Deep learning techniques, such as convolutional neural networks (CNNs), are very popular in thefield of computer vision due to their strong similarity to human brains. In order to identify images, the system employs convolutional neural networks, often known as CNNs. **Fig 4** demonstrates that a typical CNN model has three distinct levels: an input layer, several hidden layers, and an output layer. These five kinds of layers make up the model.

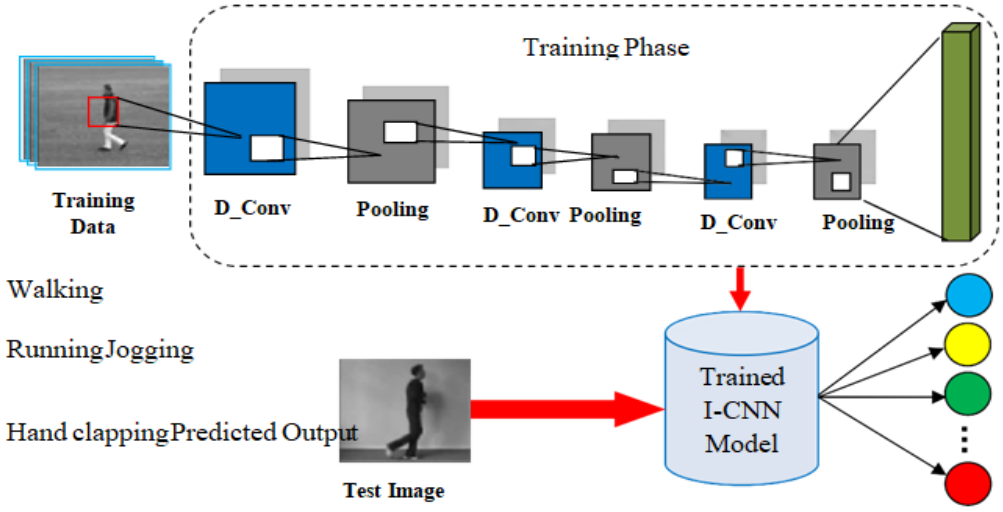

Fig 4. Architecture of Improved CNN model

- Input Training Dataset
- Convolutional layer
- Pooling layer
- Global Average Pooling (GAP) layer, in addition to Fully connected layer
- Recognition results

When processing an image, several types of layers, including convolutional, non-linear, pooling, and fully connected, are used. The height, width, and depth of a layer are the three dimensions that describe it. The first layer of a CNN is referred as the convolution layer. Every time the filteris dragged over the image, the original values of the pixels in the input image are multiplied by the values of the filter. These values are then joined together to produce a single value. A single number denotes the spatial position of the filter at the time it was used. This is done across the whole of the input image. The first input layer is responsible for producing an activation map for the image, which highlights significant parts of the image.

The second convolution layer takes in the activation map and uses it both as input and as output. Each of the image's input layers specifies a set of coordinates that point to regions of the original image where certain features, including curves and edges, may be seen to exist. High-level features such as hands, feet, and squares are produced when a first convolution layer is iterated through a second convolution layer. The complexity of the activation map's output increases proportionately with the number of convolution layers that it undergoes. As a consequence of this, the filters were able to highlight more subtly differentiated aspects of the picture, such as certain colours and text. Reduce the amount of parameters that must be monitored duringtraining with the help of the pooling layer. Additionally, it guarantees that vital information on the image is preserved while simultaneously lowering the dimensionality.

We found that the number of trainable parameters in CNN is high when we are using the flatten layer. If there are too many parameters, it will reduce the training speed and leads in overfitting. In order to solve this issue, we have used Global Average Pooling (GAP) layer to replace the flatten layers in CNNs at the end of convolution and pooling process. Each feature point is generated from a feature map recovered from the final convolutional layer, and the points are pooled together to form a vector, which is then averaged before being input into a fullyconnected layer. When a dropout occurs in the fully connected layer, the output is sent to a ten- neuron soft-max classifier for further processing. The illustration of Flatten and GAP layer is shown in **Fig 5**. The fully connected layer is the final feature map that will be used for categorization.

VGG-19 Net Model

The Deep CNN is an improved version of the CNN with additional convolutional layers. There are several different Deep CNN models besides AlexNet, which is among the most popular. The VGG-19 model (Visual Geometry Group) presented by Krizhevsky et al. [29] is a powerful DeepCNN that was entered in the ILSVRC-2014 competition. As shown in **Fig 6**, the VGG-19 Net model has 16 convolution layers, five pooling layers,two fully connected layers, and one soft-max classifier. In all, there were five sections that made up the layer. There were five distinct blocks within the overall architecture. Each block has two convolutions and one pooling layer. The remaining three blocks each contains three convolutionsand a pooling layer.

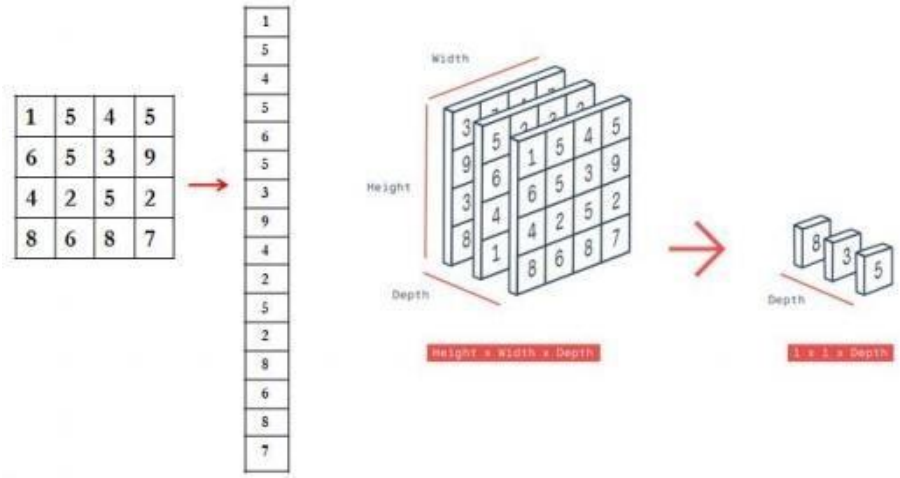

Fig 5. Variation Between Flatten Vs. Global Average Pooling Layer

Hybrid HAR-CNN Model using Improved CNN and VGG-19 Net Model

In general, ensemble learning entails training more than one network on the same dataset, then utilizing each of the trained models to make a prediction, followed by some kind of combination of all of the predicted outcomes or predictions in order to obtain better result at a final outcomeor prediction. To increase the recognition and prediction accuracy of human activity recognition, we combined the improved CNN with the VGG-19 Net model in this study. Finally, the performance of each individual model was compared against the performance of hybrid HAR- CNN prediction models to determine which model performed better overall. The overall framework for the proposed effort is depicted in **Fig 7**.

Input Image

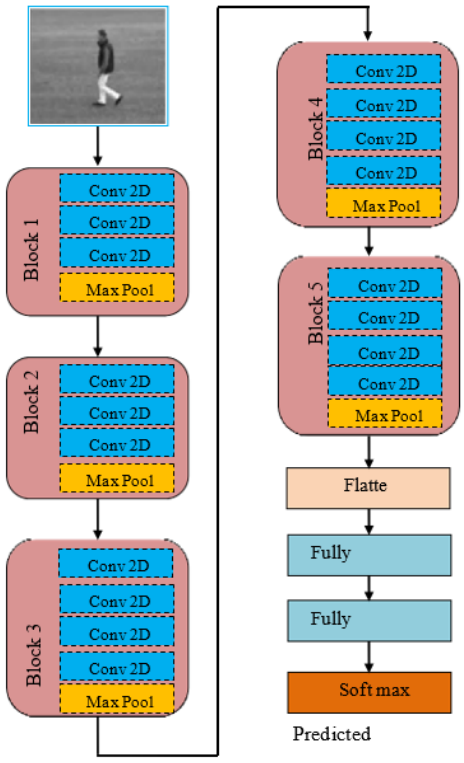

Fig 6. Workflow of VGG-19 Net Deep Learning Architecture Model

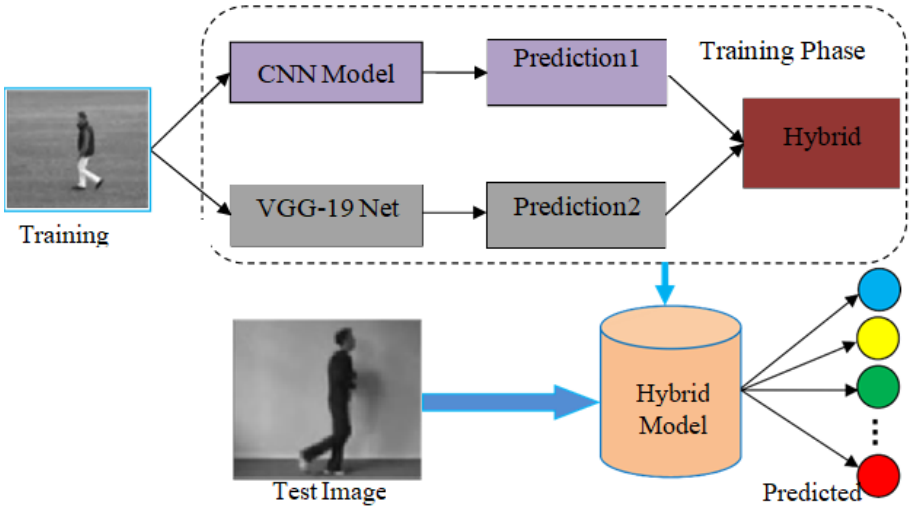

Fig 7. Architecture of Hybrid HAR-CNN Recognition System

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conducted experiments to evaluate the effectiveness of the traditional CNN model. The experiments were implemented using popular [30], NumPy [31], Matplotlib [32], and scikit-learn [33], within environments such as Jupyter Notebook and Anaconda Prompt IDE. For training and evaluation, Keras [34] and TensorFlow [35] were utilized on a system equipped with a Core i7 CPU running at 2.6GHz, a 1TB hard disk drive, and 8GB of RAM. The experiments were conducted on the KTH human activity dataset [41].

Evaluation Metrics

The proposed model's performance is assessed using various metrics including Precision, Recall, Accuracy, and F1-measure. These metrics are calculated using a confusion matrix, illustrated in **Figure 8**, which is a two-dimensional table where actual values are represented in columns and predicted values in rows. Within this matrix, TP (True Positive) indicates the instances where the model correctly predicts the positive class, TN (True Negative) signifies correct predictions of the negative class, FP (False Positive) denotes incorrect predictions of the positive class, and FN (False Negative) represents incorrect predictions of the negative class.

	P	**N**
Y	**True Positive**	**False Positive**
N	**False Negative**	**True Negative**
	P	**N**

Fig 8. Confusion Matrix

Precision

Precision is one of the best measures to show how the model is precise. It can be measured by the ratio of correctly predicted positive observations to the total predicted positive observations. Precision value can be calculated using the equation (1).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

Recall

Recall is the ratio of correctly predicted positive observations to all observations in the actual class. It is used to calculate how many of the actual positives the model catches by labeling it as positive. Recall value can be calculated using the equation (2).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

Accuracy

The Accuracy can be calculated by the number of properly classified data in a dataset divided by the total number of samples, as shown in the equation (3).

$$\text{Accuracy} = \frac{\text{TP} + \text{FP}}{\text{TP+FP+TN+FN}} \tag{3}$$

FI-Measure

The F1-measure (harmonic mean) is used to show the balance between the precision and recall measures. The F1- score measure can be calculated using the equation (4).

$$\text{FI Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{4}$$

V. RESULTS AND DISCUSSIONS

First, we have trained and evaluated two different CNN Models such as improved convolutional neural networks and VGG-19 Net model individually for recognizing the various human activities of KTH images. From **Table 2** and **Fig 9,** we found that both Improved CNN model and VGG-19 Net model achieved nearly 97% of accuracy. As shown in **Fig 9a** and **Fig 9b**, the improved CNN model and VGG-19 Net model some of the classes are misclassified. For example, hand clapping and running misclassified as walking. In order to improve the human activity recognition accuracy further we have hybrid the two models namely improved CNN and VGG-19 Net at decision level by averaging method. The hybrid model used the same input as the individual base models and calculated the average prediction.

Table 2. Performance Comparison of Proposed Model with Traditional CNN Model

S. No.	Model	Accuracy	Precision	Recall	F1- Score
1.	Improved CNN Model	96.85	97.17	96.86	96.89
2.	VGG-19 Net Model	95.76	95.82	95.79	95.77

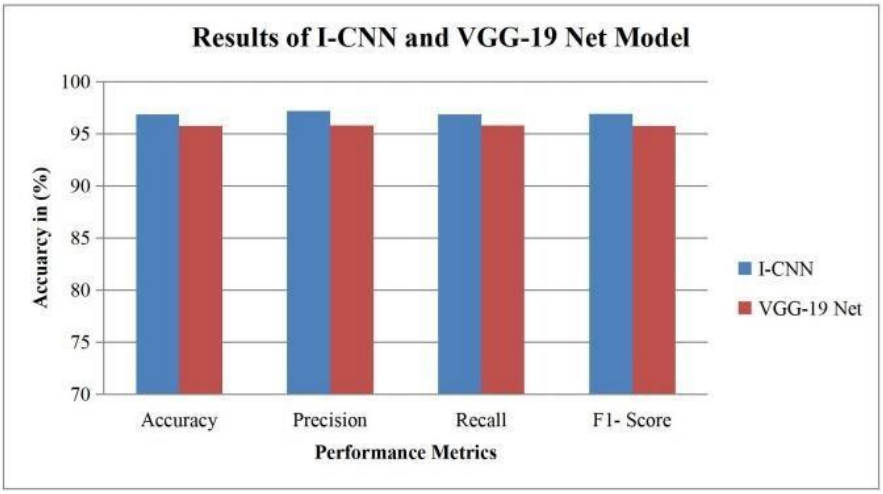

Fig 8. Classification accuracy of Improved CNN and VGG-19 Net for KTH dataset

The results shown in **Table 3** and **Fig 10** contributed us to establish that the hybrid model (Improved CNN+VGG-19 Net) was the most effective method, yielding the maximum accuracy as 98.8%. The confusion matrix was used to calculate several performance assessment measures,which are shown in **Fig 11.**

Table 2 and **Table 3** show the results of a comparison between the recommended hybrid model and the base models (the improved CNN and the VGG-19 Net model) in terms of standard metrics including trained accuracy, validated accuracy, trained loss, and validated loss after 15 epochs with dropout. These parameters are computed in order to provide an estimation of the trained models using a learning rate of 0.00001 and SGD optimization. These parameters are computed in order to provide an estimation of the degree to which the training models have been overfit.

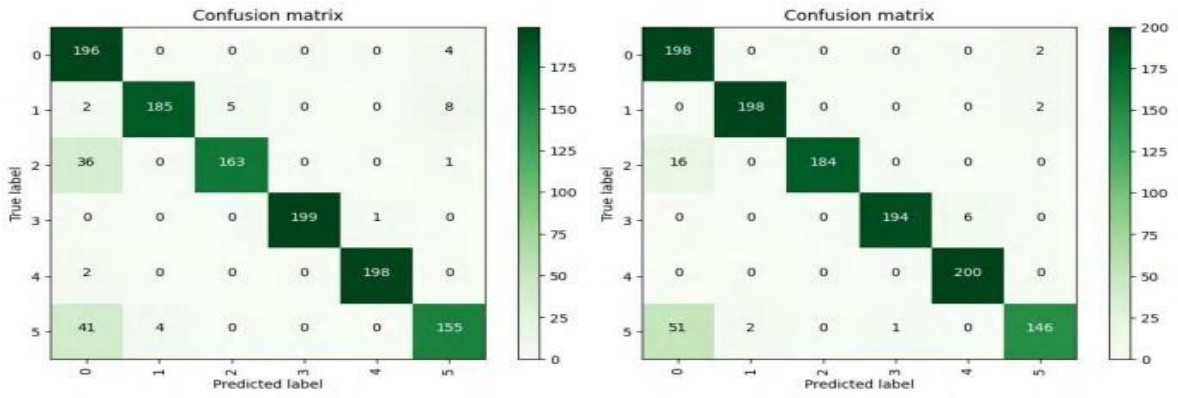

Fig 9a. Confusion matrix of Improved CNN, **Fig 9b.** Confusion matrix of VGG-19 Net

Table 3. Performance of Proposed Model

S. No.	Model	Accuracy	Precision	Recall	F1- Score
1.	Hybrid CNN Model (ImprovedCNN + VGG-19 Net Model)	98.8	98.15	98.45	98.3

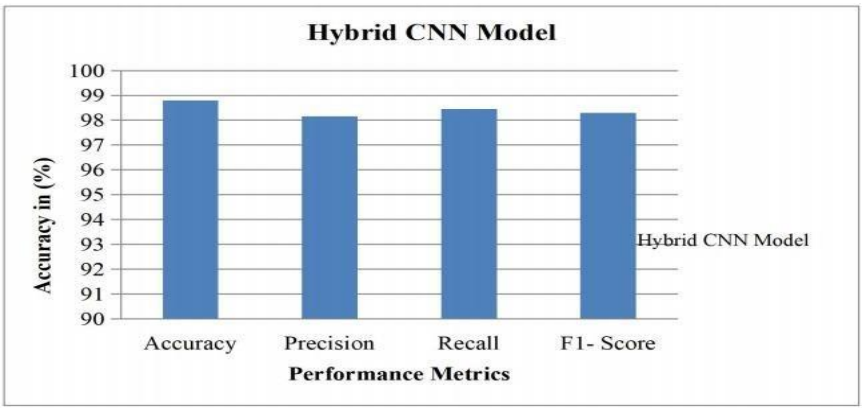

Fig 10. Performance comparison of Hybrid CNN model

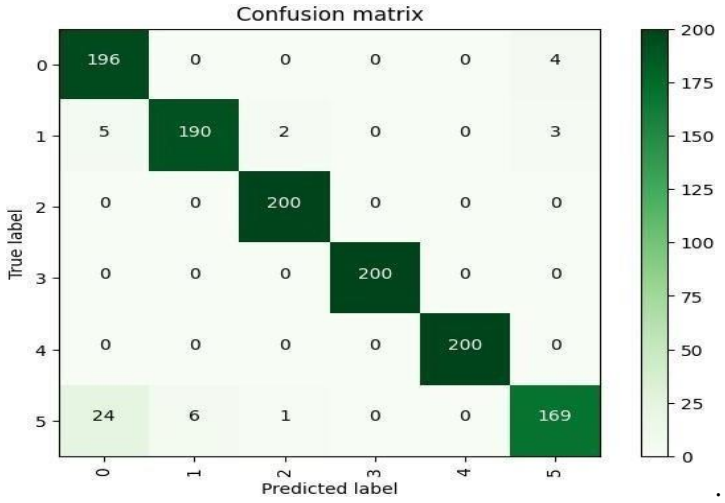

Fig 11. Confusion matrix of Hybrid CNN Model (Improved CNN + VGG-19 Net)

VI. CONCLUSION

Using an Ensemble HAR-CNN model, we have presented a method for improving the performance and prediction accuracy of various human activities like walking, running, jogging, down stair and up-stairs in KTH dataset images. To begin, we trained the improved CNN model and VGG-19 Net model for human activity recognition separately, achieving accuracy rates of 96.65% and 97.5% respectively. Then the two models are combined in order to increase the accuracy of the final result. The ensemble model achieved accuracy of 98.8%. According to the results of a comparative examination of two base models, the proposed ensemble HAR model outperforms the others in terms of performance and achieves a significant level of prediction accuracy.

In Our future objectives include integrating our proposed Ensemble HAR model into a GPU environment to expedite computational processes. Furthermore, we plan to develop an automatic mobile application for real-time human activity recognition.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

References

[1] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis,” ACM Computing Surveys, vol. 43, no. 3, pp. 1–43, Apr. 2011, doi: 10.1145/1922649.1922653.

[2] K. K. Verma, B. M. Singh, and A. Dixit, “A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system,” International Journal of Information Technology, vol. 14, no. 1, pp. 397–410, Sep. 2019, doi: 10.1007/s41870-019-00364-0.

[3] S. Bosch, R. Marin-Perianu, P. Havinga, A. Horst, M. Marin-Perianu, and A. Vasilescu, “Automatic recognition of object use based on wireless motion sensors,” International Symposium on Wearable Computers (ISWC) 2010, Oct. 2010, doi: 10.1109/iswc.2010.5665858.

[4] D. Metaxas and S. Zhang, “A review of motion analysis methods for human Nonverbal Communication Computing,” Image and Vision Computing, vol. 31, no. 6–7, pp. 421–433, Jun. 2013, doi: 10.1016/j.imavis.2013.03.005.

[5] O. D. Lara and M. A. Labrador, “A Survey on Human Activity Recognition using Wearable Sensors,” IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1192–1209, 2013, doi: 10.1109/surv.2012.110112.00192.

[6] T. B. Moeslund and E. Granum, “A Survey of Computer Vision-Based Human Motion Capture,” Computer Vision and Image Understanding, vol. 81, no. 3, pp. 231–268, Mar. 2001, doi: 10.1006/cviu.2000.0897.

[7] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, “Human activity recognition using multi-features and multiple kernel learning,” Pattern Recognition, vol. 47, no. 5, pp. 1800–1812, May 2014, doi: 10.1016/j.patcog.2013.11.032.

[8] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3D points,” 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, Jun. 2010, doi: 10.1109/cvprw.2010.5543273.

[9] M. J. Cheok, Z. Omar, and M. H. Jaward, “A review of hand gesture and sign language recognition techniques,” International Journal of Machine Learning and Cybernetics, vol. 10, no. 1, pp. 131–153, Aug. 2017, doi: 10.1007/s13042-017-0705-5.

[10] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Jun. 2012, doi: 10.1109/cvprw.2012.6239233.

[11] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257–267, Mar. 2001, doi: 10.1109/34.910878.

[12] E. Shechtman and M. Irani, “Matching Local Self-Similarities across Images and Videos,” 2007 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2007, doi: 10.1109/cvpr.2007.383198.

[13] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition,” 2008 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2008, doi: 10.1109/cvpr.2008.4587727.

[14] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, “Selective spatio-temporal interest points,” Computer Vision and Image Understanding, vol. 116, no. 3, pp. 396–410, Mar. 2012, doi: 10.1016/j.cviu.2011.09.010.

[15] G. Willems, T. Tuytelaars, and L. Van Gool, “An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector,” Computer Vision – ECCV 2008, pp. 650–663, 2008, doi: 10.1007/978-3-540-88688-4_48.

[16] A. Gilbert, J. Illingworth, and R. Bowden, “Action Recognition Using Mined Hierarchical Compound Features,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 883–897, May 2011, doi: 10.1109/tpami.2010.144.

[17] N. Dalal, B. Triggs, and C. Schmid, “Human Detection Using Oriented Histograms of Flow and Appearance,” Lecture Notes in Computer Science, pp. 428–441, 2006, doi: 10.1007/11744047_33.

[18] A. Gaidon, Z. Harchaoui, and C. Schmid, “Temporal Localization of Actions with Actoms,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 11, pp. 2782–2795, Nov. 2013, doi: 10.1109/tpami.2013.65.

[19] C. Thurau and V. Hlaváč, “Recognizing Human Actions by Their Pose,” Statistical and Geometrical Approaches to Visual Motion Analysis, pp. 169–192, 2009, doi: 10.1007/978-3-642-03061-1_9.

[20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.
[21] N. Mohan et al., "Statistical Evaluation of Machining Parameters in Drilling of Glass Laminate Aluminum Reinforced Epoxy Composites using Machine Learning Model," Engineered Science, 2022, doi: 10.30919/es8e716.
[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2014, doi: 10.1109/cvpr.2014.223.
[23] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, doi: 10.1109/cvpr.2015.7299101.
[24] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/tpami.2012.59.
[25] Z. Tu et al., "Multi-stream CNN: Learning representations based on human-related regions for action recognition," Pattern Recognition, vol. 79, pp. 32–43, Jul. 2018, doi: 10.1016/j.patcog.2018.01.020.
[26] S. Alam et al., "Effective sound detection system in commercial car vehicles using Msp430 launchpad development," Multimedia Tools and Applications, May 2023, doi: 10.1007/s11042-023-15373-2.
[27] "Preprint repository arXiv achieves milestone million uploads," Physics Today, 2014, doi: 10.1063/pt.5.028530.
[28] https://www.csc.kth.se/cvap/actions/ in Proc. ICPR'04, Cambridge, UK. (2004).
[29] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition, Proceedings of the International Conference on Learning Representations (ICLR), pp. 1-8, (2015), doi: https://doi.org/10.48550/arXiv.1409.1556
[30] Bradski, G, :The OpenCV Library". Dr. Dobb's Journal of Software Tools, (2000).
[31] C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[32] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/mcse.2007.55.
[33] https://scikit-learn.org/stable/ (2011) Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[34] https://keras.io/. (2017) Keras: The python deep learning library
[35] https://www.tensorflow.org/. (2017) An open-source software library for machine intelligence
[36] "SINR Pricing in Non Cooperative Power Control Game for Wireless Ad Hoc Networks," KSII Transactions on Internet and Information Systems, vol. 8, no. 7, Jul. 2014, doi: 10.3837/tiis.2014.07.005.
[37] L. Bhagyalakshmi, S. K. Suman, and T. Sujeethadevi, "Joint Routing and Resource Allocation for Cluster Based Isolated Nodes in Cognitive Radio Wireless Sensor Networks," Wireless Personal Communications, vol. 114, no. 4, pp. 3477–3488, Jun. 2020, doi: 10.1007/s11277-020-07543-4.
[38] K. Mahalakshmi et al., "Public Auditing Scheme for Integrity Verification in Distributed Cloud Storage System," Scientific Programming, vol. 2021, pp. 1–5, Dec. 2021, doi: 10.1155/2021/8533995.
[39] S. K. Suman et al., "Detection and Prediction of HMS from Drinking Water by Analysing the Adsorbents from Residuals Using Deep Learning," Adsorption Science & Technology, vol. 2022, Jan. 2022, doi: 10.1155/2022/3265366.
[40] "Avoiding Energy Holes Problem using Load Balancing Approach in Wireless Sensor Network," KSII Transactions on Internet and Information Systems, vol. 8, no. 5, pp. 1618–1637, May 2014, doi: 10.3837/tiis.2014.05.007.
[41] S. Singh, S. V. Singh, D. Yadav, S. K. Suman, B. Lakshminarayanan, and G. Singh, "Discrete interferences optimum beamformer in correlated signal and interfering noise," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 2, p. 1732, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1732-1743.