A Unified Framework for Human Activity Recognition using Various Deep Learning Models

[1]E.Ramya, [2]A.Manthra, [3]Dr. M.Tamilselvi, [4]Dr. H. Niroshini Infantia
[1]Assistant Professor, Department of AI&DS, Karpaga Vinayaga College of Engineering and Technology, Chengalpattu, Tamilandu.
[2]Assistant Professor, Department of CSE, Arasu Engineering College, Kumbakonam, Tamilnadu.
[3]Associate Professor& Head, Department of CSE, Roever Engineering College, Perambalur, Tamilnadu
[4]Associate Professor, St.Joseph's Institute of Technology, OMR, Chennai, Tamilnadu
[1]ramyavinoth734@gmail.com, [2]manthraarumugam1198@gmail.com, [3]tamilnaveena@gmail.com, [4]niroshiniinfantiah@gmail.com

Abstract – The recognition of human movement in videos has emerged as a top priority for researchers working in the field of computer vision due to the extensive assortment of real-world applications it offers. E-health, patient monitoring, activities that require assistance with daily life, video surveillance, security and behavior analysis, sports analysis, and a great deal more are all included in this category. To recognize human activities, a significant number of researchers have proposed approaches that rely on eyesight as the primary identifying factor. Researchers will need to address challenges such as illumination fluctuations in human activity detection, interclass similarity between images, the surroundings and recording setting, and temporal variation in order to establish a vision-based human activity recognition system that is capable of producing accurate results.

In order to address this issue, we have developed and implemented a system that is based on deep learning and is capable of producing predictions and classifications regarding human activity identification. Specifically, the ordinary CNN model, the Alex Net model, and the ResNet-50 model are the ones responsible for accomplishing this. On the basis of the findings of our research, it has been noted that the performance of a ResNet-50 model is superior to that of other two models, such as the Traditional CNN model and the Alex Net model. This is the case after comparing the ResNet-50 model to the other two models. In order to evaluate the effectiveness of our proposed approach, we have produced a benchmark dataset that is available to the public from KTH. Through the utilization of the ResNet-50 model as a feature extractor and Soft-max as the classifier, the model is able to achieve the best level of performance that is attainable. The accuracy is 98.44%, the precision is 98.5%, the recall is 98.5%, and the F1-score is 98.5% with this configuration.

Keywords: ***Deep learning, activity recognition, human activity, vision recognition, sensor recognition and computer vision.***

1. INTRODUCTION

The recognition of human movement in videos has emerged as a top priority for researchers working in the field of computer vision due to the extensive assortment of real-world applications it offers. E-health, patient monitoring, activities that require assistance with daily life, video surveillance, security and behavior analysis, sports analysis, and a great deal more are all included in this category [1]. In other words, it is the problem of identifying or categorizing the various activities that are carried out by a person in the movies (that is, the sequence of image frames). There are numerous kinds of activities that a person can engage in a variety of settings, including activities that take place indoors and outside, activities that are conducted on a daily basis, activities that take place in public areas or retail malls, and even more [2]. According to the data presented in Figure 1, the recognition of human activities can be broken down into two distinct categories: vision-based activity and sensor-based activity.

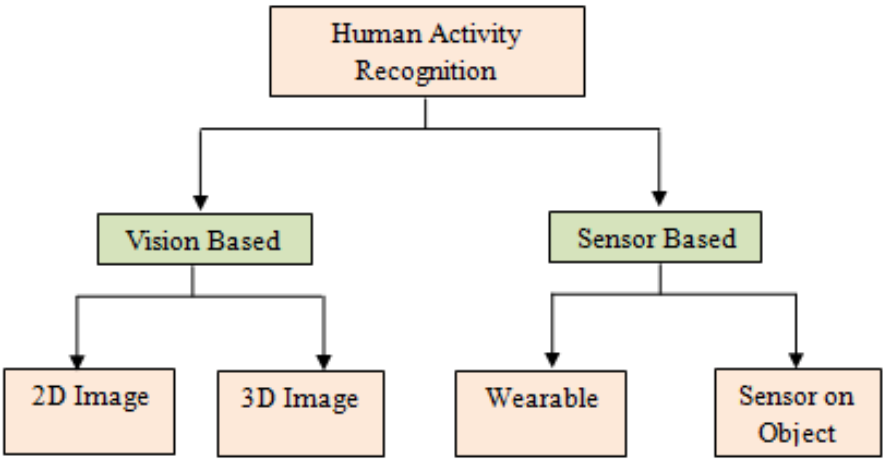

Fig. 1 Types of human activity recognition

The development of an autonomous system that is capable of accurately recognizing and comprehending human behavior and actions is one of the primary goals of the civilization founded on artificial intelligence. For instance, a robot assistant could be able to provide assistance to a patient who is being monitored at home, as well as analyze the appropriate approach to exercise and avoid the patient from experiencing any further injuries in the future [3]. This would allow the robot to serve society in a more effective manner. As a result, such an intelligent system will be of great assistance to us since it will free up time that would otherwise

be spent going to the doctor, which will in turn reduce the amount of money spent on medical care, and it will also provide constant remote monitoring of the patient [4].

The past twenty years have seen the development of a great number of feature-based techniques that are both manually designed and automatically taught for the purpose of human action recognition in videos. Handcrafted features that primarily concentrate on basic atomic actions were the foundation of earlier methods to human activity recognition. It is based on spatial background subtraction, optical flow, dense trajectories, and human position variations [5] that the handcrafted feature extraction methodologies for activity recognition will be utilized.

It has been noted that when it comes to action categorization, handcrafted features solutions provided promising results, although they relied more on feature descriptors. The implementation of these solutions needed more labor and skill in the relevant field [6]. There are still many key challenges that have not been resolved in the field of machine learning, including intra-class variation, illumination changes, occlusion, actions similarities, viewpoint variations, change in scale, appearance, age, frame resolutions, and lighting conditions [7].

Challenges in HAR: Human activity recognition is a challenging problem in machine learning.\ The complexity of the recorded films and the varied changes in the actions of humans, as are depicted in Figure 1.3, make Human Activity Recognition (HAR) a difficult study subject in the field of computer vision. In the process of human activity recognition, the researchers encountered a number of problems, some of which are listed below.

Fig. 2 The various obstacles encountered in the recognition of human activities

- Occlusion
- Background and Environment Conditions
- Viewpoint Variation
- Human activities exhibit significant intra-class variety.
- There is a significant amount of similarity between activities across different classes.

Summary of the paper: There are five main parts to this study paper, which are: In Section 2, a lot of research on vision-based human activity recognition is reviewed. Research gaps are then pointed out, and the goal of the research is talked about. The suggested better convolutional neural network model will be talked about in Section 3. In Section 4, the suggested improved CNN model's test results are talked about and compared to the results of the traditional CNN model. The results will be talked about and a summary will be given in Section 5. We will also suggest ways to do future study.

2. LITERATURE SURVEY

Human activity identification in video sequences is the most popular and quickly growing area of study in the field of computer vision. This is because it has so many uses in everyday life. In this group are things like safety, surveillance, healthcare, robots, animations, sports analysis, content-based video summary, behavioral analysis, smart homes, and a lot more. Figure 2 shows that over the last few decades, many feature-based methods have been created over the goal of recognizing human actions in movies and scenes [8]. These techniques can be taught automatically or by hand. The very first ways to figure out what a person was doing were based on custom features that were mostly focused on basic atomic actions. It looked like these functions weren't as useful for real-world situations. Because they produce a very accurate model, these methods have some problems. The main one is that they need to be used with pre-processed data and are hard to use in real life. As an example, Bobick and Davis [9] got the motion feature from video frames as Motion History Images (MHI) and Motion Energy Images (MEI) temporal template to tell the difference between people acting in situations where the background was still. They have focused on certain types of human motion, and they have thought about how motion changes over time.

Shechtman and Irani [10] proposed a template for a behavior-based similarity matrix in order to quantify the degree of similarity between human acts. For the purpose of correlating the dynamic

behavior and activities, they extended the 2D picture correlation to the 3D space-time volume. For the purpose of identifying the activity in movies, Rodriguez et al. [11] presented a maximum average correlation height (MACH) filter template-based approach. Their model is able to solve the problem of intra-class variances while incurring the least amount of computing expense possible.

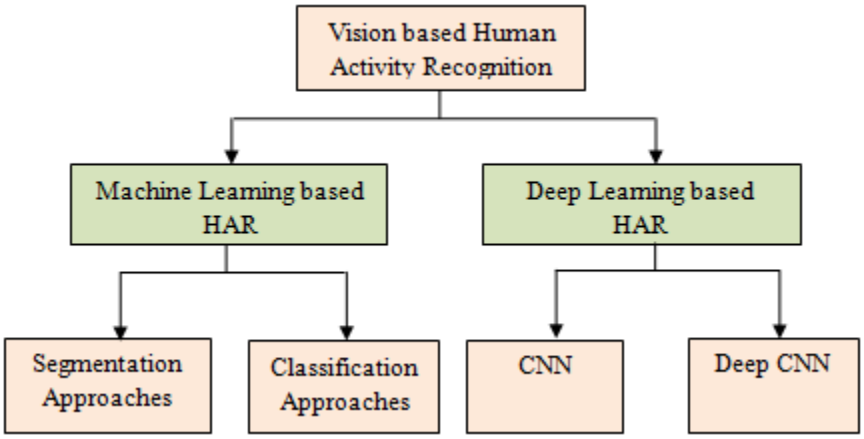

Fig. 3 Different kind of vision based human activity recognition approaches

A Space-Time Interest Points (STIPs) approach was proposed by Chakraborty et al. [12] in order to recognize the activity based on local interest points. This was accomplished by extending the 2D Harris detector to a 3D corner detector from the previous approach. Under conditions of obstructed background and view changes, the STIP features-based representation has demonstrated great results in terms of posture estimation. Their approach, on the other hand, is dependent on the motion of the camera, often known as camera jitters. Willems et al. [13] provided an approach to localized action that makes use of second derivatives of the corner detector. This was in addition to the expansion of the 2D Hessian detector to the 3D space dimensions.

The Histogram of Optical Flow (HOF) based spatial temporal descriptor was introduced by Laptev et al. [14] as a novel way to automatically annotate movie clips for the purpose of training the action classifier. This descriptor is an extension of the 2D Harris interest point detector, and it is designed to classify actions in action-packed videos. Furthermore, the bag of features-based technique demonstrated a high level of robustness in the face of perspective fluctuations, variation in illumination, and background conditions characterized by clutter. In order to recognize the action in moving ambient conditions, Dalal et al. [15] developed a human

pose descriptor by utilizing the Histogram of Oriented (HoG). The gradient features and the differential optical flow motion descriptor are being used in this method for the purpose of accurately portraying human actions in realistic movie scenarios. In a variety of difficult circumstances, the cumulative characteristics of the descriptor demonstrated promising results.

In their study [16], Gaidon and colleagues introduced an Actom Sequence Model (ASM) that was designed to recognize action movies of varying lengths. This model was based on the temporal extension of the bag-of-features technique. The formulation of actoms is based on the sequence of atom units, and the visual characteristics are expressed as a sequence of the histogram of actoms. A human pose model feature descriptor for action recognition was published by Thurau and Hlavac [17]. This feature descriptor was based on a histogram of the gradient (HoG) on a particular region of interest (RoI), and it represented a feature vector by utilizing non-negative matrix factorization.

The same may be said for the numerous deep learning models that researchers have constructed through the utilization of CNN. Convolutional Neural Networks, also known as CNN, are a flexible concept that may be utilized to implement various scene classification techniques. The very first CNN model was developed by LeCun and colleagues [18]. This model is comparable to a conventional Artificial Neural Network (ANN) and serves as the foundation for contemporary CNN. A significant source of inspiration for the construction of the CNN model is the neurons seen in both animal and human brains. In recent times, researchers have generated a great deal of models that are associated with pictures that have classification issues.

An illustration of this would be the presentation of four distinct fusion approaches along the temporal dimension by Karpathy et al. [19]. In addition to this, they presented a technique known as delayed fusion, which allows higher layers to gather more global knowledge along both the temporal and territorial dimensions. The implementation and execution of the time convolution resulted in an increase in the connectivity of all of the convolutional layers in the chronological dimension.

3. PROPOSED WORK

In this section, we have used different deep learning models such as traditional CNN, AlexNet and ResNet-50 models for classification and prediction of human activity recognition.

3.1 CNN Model

Convolutional neural networks are a subcategory of neural networks that are used to perform specific tasks. The receptive field is a biological neuron that is meant to reproduce the connectivity pattern of neurons present in the human brain. Its design is based on the idea of a biological neuron that is capable of receiving information. A feed forward neural network is the CNN model. This type of neural network is made up of a stack of filters (the convolutional layer) and sub-sampling layers (the pooling layer) that repeat themselves in a different order. At the very end of the network, it is composed of one or more neurons that are completely coupled to one another (a layer that is densely connected and fully connected).

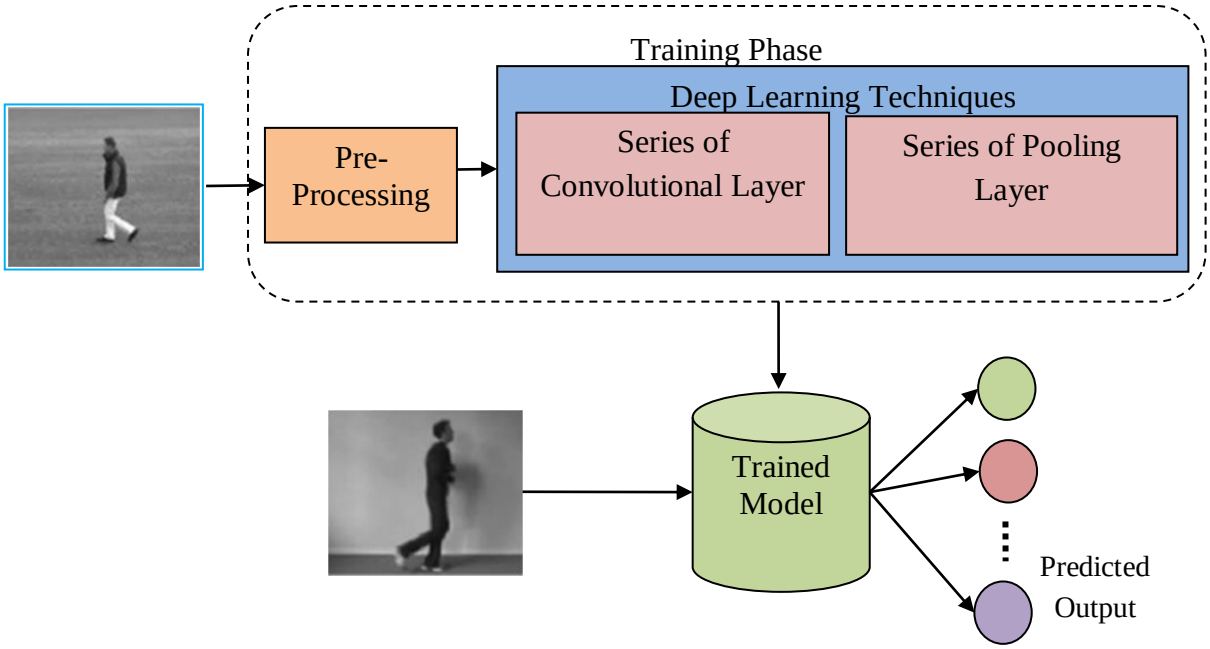

Fig. 4 General Architecture of Convolutional Neural Network

In spite of the fact that this model is employed in a range of disciplines, it still results in the best possible outcomes when it is applied to applications that include image processing. The CNN is created by concatenating discrete blocks or layers. This is the method that is used. A variety of responsibilities are being carried out by the components of various tiers as they come together. Figure 4 is a representation of the general architecture of the typical convolutional neural network. This graphic may be found here. Following is a list of the layers that make up this network.

- ❖ Feature Extraction Stage
- ❖ Feature Reduction Stage
- ❖ Flatten Stage
- ❖ Fully Connected Layer
- ❖ Soft-max classifiers

3.2 Alex Net Model

AlexNet [20] is a CNN with eight layers: three maximum pooling levels, three fully linked layers, five convolution layers, and so forth. In order to train AlexNet, the ImageNet database was mined for over a million photos and over a thousand categories. It has a maximum input size of 227x227x3 pixels:

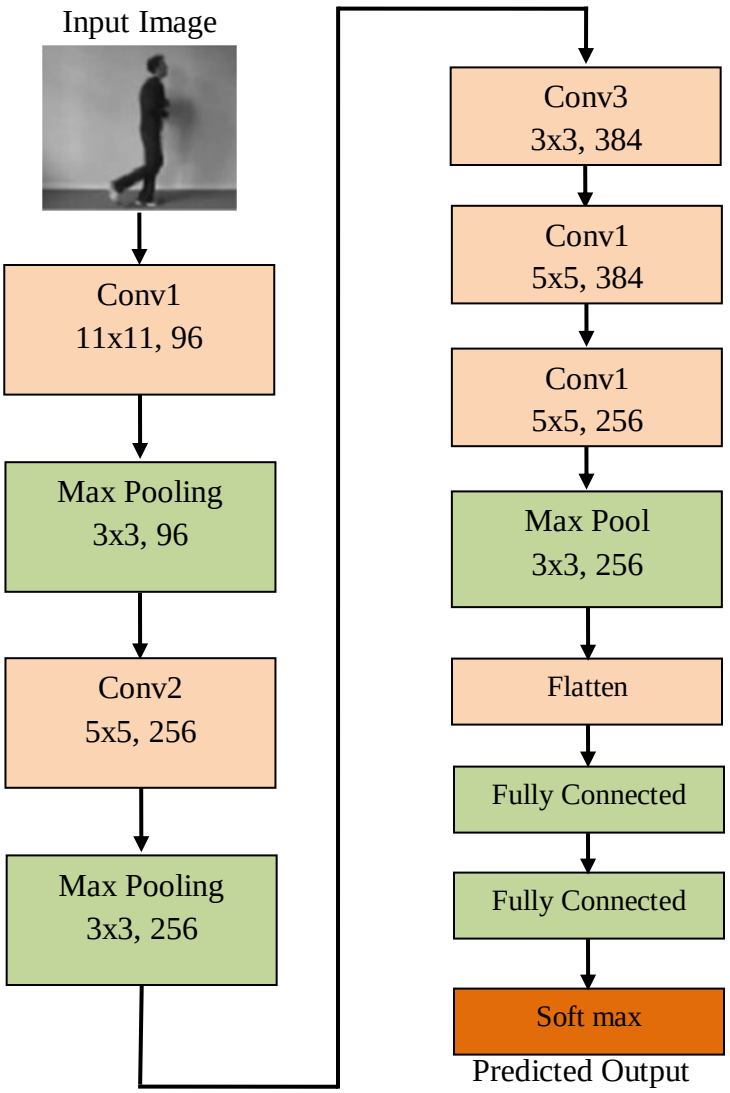

Fig. 5 Architecture of AlexNet Model

The resolution of 227×227 represents the width and height of the input image, while the number 3 signifies that the images are RGB color images. Four strides and ninety-six filters with a dimension of eleven by eleven make up the first convolution layer. For the second convolution layer, we use 256 filters, each with a 5x5 pixel size and a 1-pixel stride. Filters in the third convolution layer are 384 in number and have a 3x3 grid with one stride.

Table 1 Information about Alex Net model

Layer	Filters	Filter Size	Stride
1	96	11×11	4
2	256	5×5	1
3	384	3×3	1
4	384	3×3	1
5	256	3×3	1

There are 384 filters in the fourth convolution layer, each with a 3x3 filter size and a stride of 1. There are 256 filters in the fifth convolution layer, and each one is 3x3 and one stride long. Information about the number of convolutions, filter size, and stride can be found in Table 1. Max pooling and ReLU both employ a 3x3 pool size to normalize each convolutional layer that follows. Figure 5 illustrates the fundamental basic layout of the AlexNet system.

3.3 ResNet-50 Net Model

A CNN that contains fifty levels of nested sub networks is referred to as ResNet-50. According to Kustina et al. [21] and Bawaningtyas et al., the design consists of a total of 48 convolution layers, 16 bottleneck blocks, and one completely linked layer. As shown in Figure 6, there are a number of different bottleneck components that are both the same and distinct from one another. Each of the first three bottleneck blocks contains a convolution layer with 64 filters; these filters range in size from 1x1 to 3x3, and there are also 64 1x1 filters. In blocks 4–7, there are convolution layers with 512 and 128 filters, respectively; the last two have 1x1 and 3x3 filter sizes. A third layer with 1024 1x1 filters rounds up the architecture, which also includes two convolution layers with 256 1x1 filters and one 3x3 filter. Layer 2048 contains 2048 filters, all with a 1x1 size, in contrast to the convolution layers of building blocks 14–16, which have filter sizes ranging from 1x1 to 3x3. A few examples of ResNet models are ResNet-50, ResNet-101, and ResNet-18.

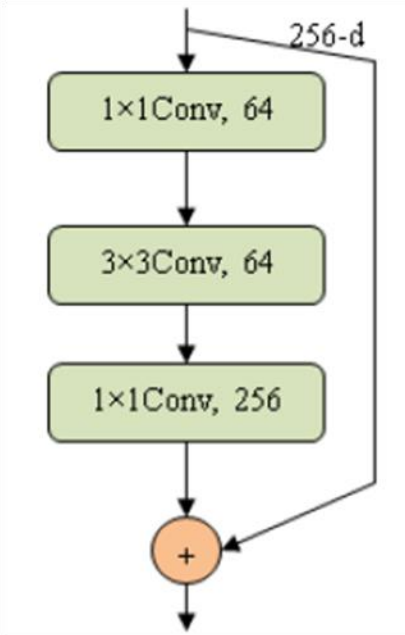

Fig. 6 Bottleneck building blocks

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we have compared the results of all of the experiments that were proposed with the intention of demonstrating the efficacy of the various CNN and deep CNN models. In environments such as Jupyter Notebook and Anaconda Prompt IDE, experiments are carried out with the assistance of deep learning packages such as Open CV [22], Numpy [23], Matplotlib [24], and sklearn [25]. In order to train and assess the suggested methods on the KTH human activity dataset, Keras [26] and TensorFlow [27] were utilized on a corei7 CPU operating at 2.6GHz, a hard disc drive with a capacity of 1 terabyte, and 8 gigabytes of random access memory.

4.1 Dataset Collection

The Royal Institute of Technology in Sweden was the organization that was responsible for the production of the KTH dataset in the year 2004 [28]. There are six different human actions that are included in this dataset. These actions are walking, jogging, running, boxing, hand clapping, and hand waving. The activities in question were carried out by a total of twenty-five different individuals in four separate scenarios. As a result, it encompasses a total of 600 video sequences, which is equal to 25 times 6 times 4. This dataset is considered to be one of the more basic possibilities for testing human activity identification algorithms because it was collected using a camera and background that remained steady during the recording process. As a result, these

movies were captured using a camera. The illustration in Figure 7 is a single picture that illustrates an example of each action that could be taken in each one of the four conceivable scenarios. The model is trained using seventy percent of the dataset, while twenty percent of the dataset is used to validate the model, and ten percent of the dataset is used to test the model.

Table 2 KTH images dataset information

S. No.	Expression Type	Total No. of Images	Training Images	Validation Images	Testing Images
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

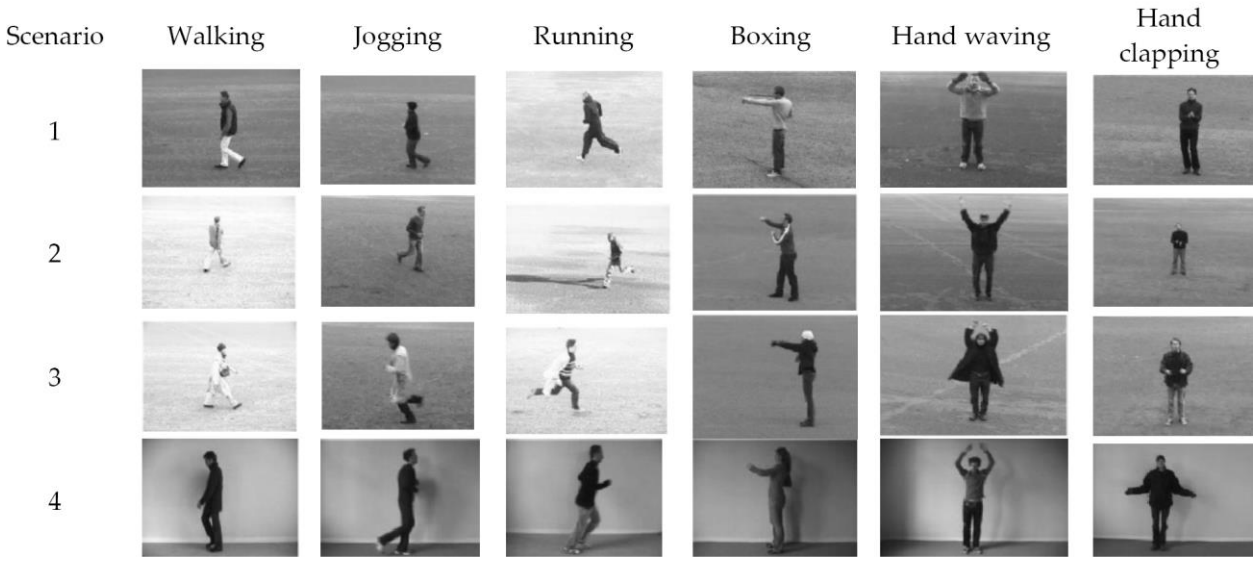

Fig. 7 One frame example of each action in KTH dataset

4.2 Evaluation Metrics

Several different performance metrics, including Precision, Recall, Accuracy, and F1-measure, are utilized in order to assess the effectiveness of the model that has been proposed. The confusion matrix, which is a table with two dimensions and is depicted in Figure 8, is utilized in the process of calculating the metrics that were previously described. When looking at this matrix, the column side has the actual values, while the row side contains the predicted values.

For the sake of this discussion, we will refer to the number of True Positives, True Negatives, False Positives, and False Negatives as TP, TN, FP, and FN individually. When the models properly forecast the positive class, the TP is an outcome that occurs. When the models properly anticipate the negative class, the TN is an outcome that occurs. One of the outcomes is the FP, which occurs when the models make an inaccurate prediction about the positive class. One of the outcomes is the FN, which occurs when the models make an inaccurate prediction about the negative class.

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

Fig. 8 Confusion matrix

Precision

Accuracy is a highly effective metric for assessing the precision of a model. The measure can be found by dividing the total number of expected positive observations by the proportion of precisely anticipated positive observations. The precision value can be determined by utilizing equation (1).

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

Rec:

One measure of accuracy is recall, or the fraction of positive observations that were correctly predicted relative to the total number of observations in the actual class. The number of cases that the model properly detects as positive is determined by using it. The calculation of recall value can be determined using equation (2).

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

Acc.

In order to find the Accuracy (Acc), one can divide the total number of samples by the number of correctly categorized data in a dataset, as indicated in equation (3).

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

F1 -measure

When comparing recall and precision, the F1-measure (harmonic mean) is a good indicator of how well they are balanced. Using the formula (4), one may determine the F1-score metric.

$$F = 2 * \frac{\text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (4)$$

4.3 Results and Discussions

We have assessed the CNN model as well as a number of different transfer learning models, including the AlexNet model and the ResNet-50 model, according to this study. In order to circumvent the issue of overfitting, Optimizers Dropout and Adam were utilized. At a maximum of fifteen epochs in duration, the CNN model and the pre-trained deep learning model were both trained and confirmed. This is demonstrated in Figure 9. Figure 10 presents the confusion matrix for the three models that are being considered. To a large extent, the KTH dataset contains classes that have been appropriately classified, resulting in satisfactory outcomes. Having poor performance in the workplace is the category that the running action falls under. Whereas the most of the hand clapping activities are misclassified as walking, the majority of the running actions are misclassified as walking and vice versa.

Table 3 Performance Comparison of proposed model with traditional CNN model

S. No.	Model	Accuracy	Precision	Recall	F1- Score
1.	Traditional CNN Model	93.35	93.66	93.35	93.28
2.	Alex Net Model	95.85	96.44	95.86	95.89
3.	ResNet-50 Model	97.85	98.17	98.86	98.89

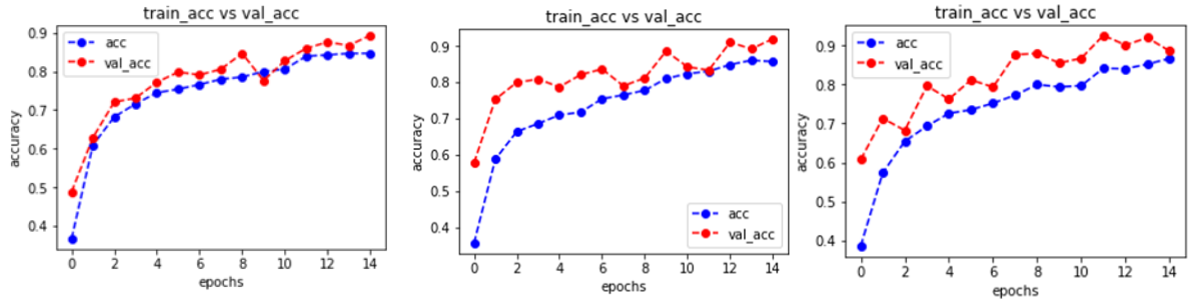

Fig. 9 Training and validation accuracy for KTH dataset with CNN, Alex Net and ResNet-50

The standard CNN as well as a variety of transfer learning approaches are analyzed and tested in this sub section using the same datasets and configurations. Figure 11, which is a comparison

chart of the various transfer learning models, corresponds to Table 3, which is including the information. Based on the results of the experiments, we are able to notice that a classification network that is based on deep learning is capable of extracting features from images, as well as performing hierarchy abstraction and classifying human activity recognition using images from the KTH dataset. AlexNet, VGG-16 Net, and ResNet-50 each report an accuracy rate of 93.35 percent, 95.85 percent, and 97.85 percent, respectively. When compared to other CNN models and transfer learning AlexNet models, the performance and classification capabilities of ResNet-50 models are superior to those of the other neural networks.

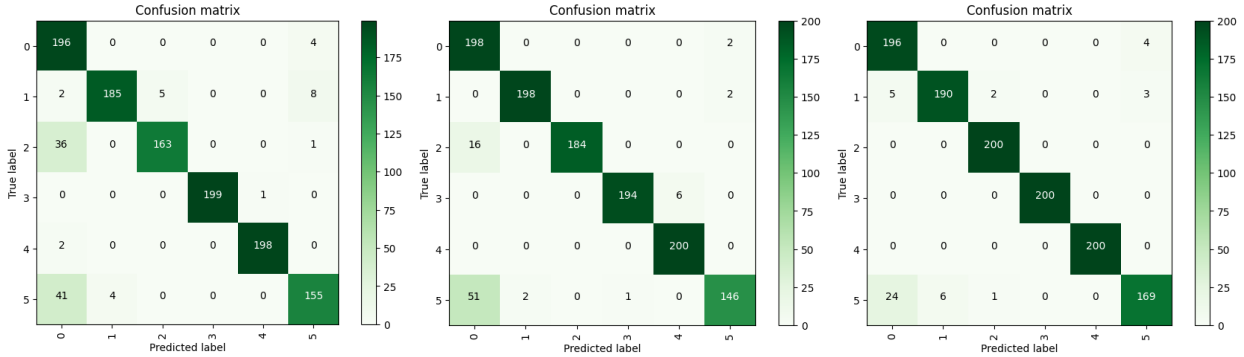

Fig. 10 Classification accuracy for KTH dataset with Traditional CNN, Alex Net and ResNet-50

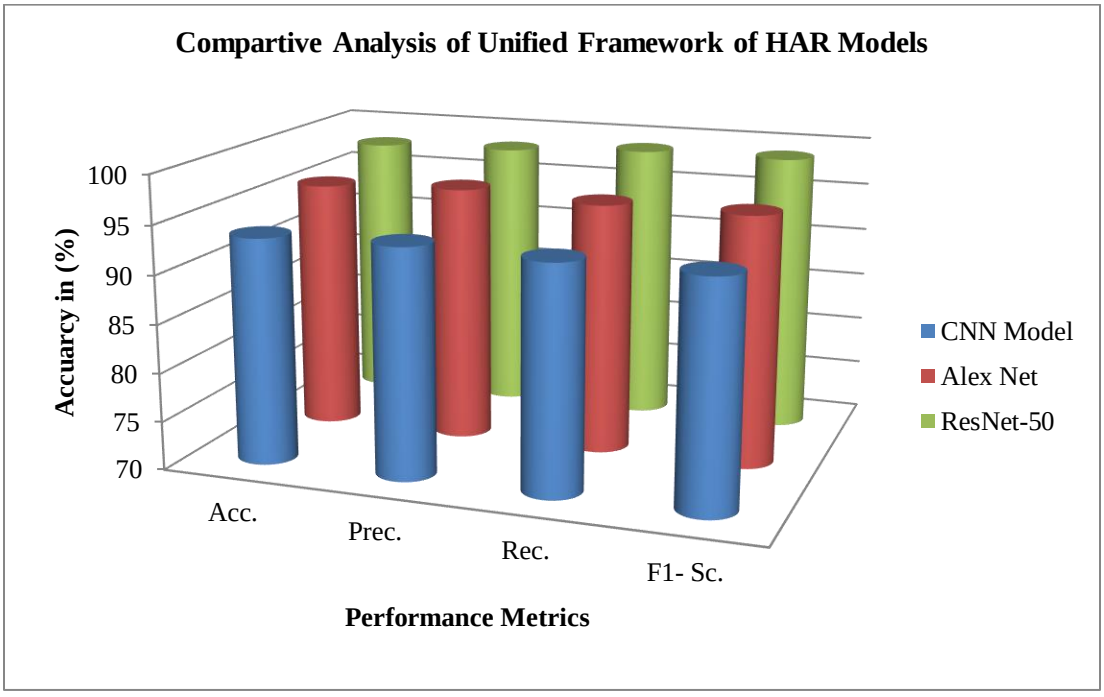

Fig. 11 Performance analysis of proposed model with traditional CNN

5 CONCLUSION

The study began with a understanding of human action or activity recognition in video sequences, as well as the primary obstacles that currently exist for action recognition in videos, associated HAR applications, and the potential manifestation of these challenges through the utilization of existing solutions. The CNN model and a number of transfer learning models, such as AlexNet and ResNet-50 models, are detailed in this study along with its application to the KTH dataset for the purpose of human activity recognition. Using the convolutional layer, the information on human activities was collected, and the soft-max classifier was used to classify the input. The KTH benchmark datasets are utilized in order to gauge the effectiveness of the models. 93.35%, 95.85%, and 97.85% are the corresponding levels of accuracy that are associated with the standard CNN, AlexNet, and ResNet-50 models.

With regard to human activity recognition, the experimental findings demonstrated superior discrimination compared to conventional CNN. Using Python 3.5 and the various library packages that are associated with it, the experiments were carried out. In the future, we intend to include our model into a human activity recognition system and in order to reduce the amount of time required for computing, we will develop it in a GPU environment. Furthermore, in order to enhance the human activity recognition system in an effective manner, we need to take the ensemble CNN model and execute it.

References

1. J. Aggarwal and M. Ryoo, “Human activity analysis: A review,” ACM Comput. Surv., vol. 43, pp. 16:1–16:43, Apr. 2011.
2. Danafar, S. and Gheissari, N. Action recognition for surveillance applications using optic flow and SVM. In Asian Conference on Computer Vision, Vol. 6(2), pp. 457–466, 2007.
3. Gorelick, Lena, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes." IEEE transactions on pattern analysis and machine intelligence, Vol. 29, No. 12, pp. 2247-2253, 2007.
4. Ankur Agarwal and Bill Triggs, “Multilevel image coding with hyper features”, International Journal of Computer Vision, Vol. 78(1), pp. 15–27, 2008.
5. A. Gilbert, J. Illingworth, and R. Bowden, “Action recognition using mined hierarchical compound features”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33(5), pp. 883–897, 2011.
6. P. Deepan, R. Santhosh Kumar, B. Rajalingam, P. Santhosh Kumar Patra and S. Ponnuthurai, "An Intelligent Robust One Dimensional HAR-CNN Model for Human Activity Recognition using Wearable Sensor Data," *2022 4th International Conference*

on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 1132-1138, doi: 10.1109/ICAC3N56670.2022.10073991.
7. Dimitris Metaxas and Shaoting Zhang, "A review of motion analysis methods for human Nonverbal Communication Computing", Image and Vision Computing, vol. 31, pp. 421-433, 2013.
8. Hbali, Y., Hbali, S., Ballihi, L., & Sadgal, M., Skeleton-based human activity recognition for elderly monitoring systems. IET Computer Vision, Vol. 12(1), pp. 16-26, 2018.
9. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
10. E. Shechtman and M. Irani, "Matching Local Self-Similarities across Images and Videos," 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383198.
11. Rodriguez, Mikel. (2013). Spatio-Temporal Maximum Average Correlation Height Templates In Action Recognition And Video Summarization.
12. Chakraborty, Bhaskar & Holte, Michael & Moeslund, Thomas & Gonzàlez, Jordi. (2012). Selective spatio-temporal interest points. Computer Vision and Image Understanding. 116. 396-410. 10.1016/j.cviu.2011.09.010.
13. Willems, Geert & Tuytelaars, Tinne & Van Gool, Luc. (2008). An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. 650-663. 10.1007/978-3-540-88688-4_48.
14. A. Laptev, J. Illingworth, and R. Bowden, “Action recognition using mined hierarchical compound features”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33(5), pp. 883–897, 2011.
15. Dalal, Navneet & Triggs, Bill & Schmid, Cordelia. (2006). Human Detection Using Oriented Histograms of Flow and Appearance. Proceedings of IEEE European Conference on Computer Vision, 2006. 3952. 428-441. 10.1007/11744047_33.
16. Gaidon, A.; Harchaoui, Z.; Schmid, C. Temporal localization of actions with actoms. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 2782–2795. [Google Scholar] [CrossRef][Green Version]
17. Thurau, C., Hlaváč, V. (2009). Recognizing Human Actions by Their Pose. In: Cremers, D., Rosenhahn, B., Yuille, A.L., Schmidt, F.R. (eds) Statistical and Geometrical Approaches to Visual Motion Analysis. Lecture Notes in Computer Science, vol 5604. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03061-1_9
18. Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P., (2015). Gradient-based learning applied to document recognition, Proceedings of the IEEE, vol. 86, pp. 2278–2324.
19. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).
20. P.Deepan and L.R. Sudha, “Comparative Analysis of Remote Sensing Images using Various Convolutional Neural Network”, EAI End. Transaction on Cognitive Communications, 2021. ISSN: 2313-4534, doi: 10.4108/eai.11-2-2021.168714.
21. Deepan P, Vidya R, Arsha Reddy M, Arul N, Ravichandran J, Dhiravidaselvi S. (2024), “A Hybrid Gabor Filter-Convolutional Neural Networks Model for Facial Emotion Recognition System”, *Indian Journal of Science and Technology*. 17 (35):3696-3703. https://doi.org/10.17485/IJST/v17i35.1998
22. Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

23. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Oliphant, T. E., Array programming with NumPy. *Nature*, *585*, 357–362, 2020. https://doi.org/10.1038/s41586-020-2649-2
24. Hunter, J. D., Matplotlib: A 2D graphics environment. Computing in Science & Engineering, Vol. 9(3), pp. 90–95, 2007.
25. https://scikit-learn.org/stable/ (2011) Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
26. https://keras.io/. (2017) Keras: The python deep learning library
27. https://www.tensorflow.org/. (2017) An open-source software library for machine intelligence
28. https://www.csc.kth.se/cvap/actions/ (2004) in *Proc. ICPR'04, Cambridge, UK.*