EMITTER International Journal of Engineering Technology
Vol. 9, No. 2, December 2021, pp. 357~376
DOI : 10.24003/emitter.v9i2.642

HActivityNet: A Deep Convolutional Neural Network for Human Activity Recognition

Md. Khaliluzzaman[1], Md. Abu Bakar Siddiq Sayem[2], Lutful Kader Misbah[3]

[1, 2, 3]Dept. of Computer Science and Engineering, International Islamic University Chittagong (IIUC), Chattogram-4318, Bangladesh
E-mail: abssayem121194@gmail.com, misbahiiuc@gmail.com
Correspondence Author: khalilcse021@gmail.com

Received September 4, 2021; Revised October 7, 2021; Accepted November 8, 2021

Abstract

Human Activity Recognition (HAR), a vast area of a computer vision research, has gained standings in recent years due to its applications in various fields. As human activity has diversification in action, interaction, and it embraces a large amount of data and powerful computational resources, it is very difficult to recognize human activities from an image. In order to solve the computational cost and vanishing gradient problem, in this work, we have proposed a revised simple convolutional neural network (CNN) model named Human Activity Recognition Network (HActivityNet) that is automatically extract and learn features and recognize activities in a rapid, precise and consistent manner. To solve the problem of imbalanced positive and negative data, we have created two datasets, one is HARDataset1 dataset which is created by extracted image frames from KTH dataset, and another one is HARDataset2 dataset prepared from activity video frames performed by us. The comprehensive experiment shows that our model performs better with respect to the present state of the art models. The proposed model attains an accuracy of 99.5% on HARDatase1 and almost 100% on HARDataset2 dataset. The proposed model also performed well on real data.

Keywords: Human activity recognition (HAR), convolutional neural network (CNN), KTH dataset, computer vision, vanishing gradient problem.

1. INTRODUCTION

Human activity recognition (HAR) is a key research field of computer vision (CV) for developing context-aware human assistance systems. Activity recognition is the task of identifying ongoing activities in image frames or videos. Human activity recognition now-a-days becomes a hotspot in the area of image processing, artificial intelligence, sign language, human-computer interaction and pervasive computing. Recognizing complex human activities has several important applications including surveillance systems for automatic recognition of suspicious activities in public areas, remote monitoring, sport play analysis, crowd behavior prediction, home based

security system, and monitoring of patient activity in hospitals. Because of real-world constraints, HAR is a challenging problem. There are so many limitations such as, background disorder, variations in view point, changes in scale and partial blockade. Besides, due to visual similarities, certain actions are challenging to differentiate, like running, jogging, walking [14]. It is also challenging to distinguish HAR from different hand gestures [38].

To overcome the challenges and improve the performance of human activity recognition, researchers develop various methods in the area of CV based on the handcraft representation as well as deep network. On that, deep network such as deep convolutional neural network plays a vital role in the area of HAR. This is because of the deep-network can efficiently handle non-linear boundaries and hence reduce the misclassification rate.

As the deep network has an efficient accuracy label in object recognition and has tremendous performance in the context of blocs of pixels. Regarding to overcome the vanishing gradient problem, in this work, we have developed a simple CNN model named Human Activity Recognition Network (HActivityNet) for HAR inspired by MiniVGGNet to enhance the recognition accuracy and decrease the training and validation loss. Though our model is developed inspired by MiniVGGNet characteristics, however, it works better than the MiniVGGNet for single and multi-view action recognition. To overcome the problem of imbalance dataset, we have built two datasets named HARDataset1 and HARDataset2. The HARDataset1 dataset created by using KTH dataset and HARDataset2 dataset is created by using real-time images.

2. RELATED WORKS

According to recent HAR research, it may be roughly classified into two classes, namely deep network-based and representation-based approaches [14]. Handcraft traits are used in representation-based approaches to classify human actions into different types. Spatiotemporal interest point extractors [15, 16], holistic representation [17-19], and motion trajectory extractors [20, 21] are some of the sub-categories. Human activity delivers information about the ongoing behavior and actions of the subjects [26]. Humans have a highly evolved visual cortex which can detect and recognize activities with ease, and requires no conscious supervision. A lot of handcraft design features such as LBPs [27], HOG [28], SIFT, and SURF [29] have been achieved excellent success in HAR. Besides traditional approaches such as k-nearest neighbor [30], SVM [31], deep architectures [32, 33] were proposed largely.

Deep learning networks, in addition to representation-based HAR approaches, can successfully handle nonlinear boundaries, lowering misclassification rates. As a result, deep learning approaches have recently gained popularity. Deep action representation (DAR) research can be divided into several groups, including transformed frames, multi-streams network, static frame learning, recurrent networks and 3D-CNN. A generative probabilistic model is established by these techniques to learn higher dimensional frame changes and thus from neighboring frames meet the

motion information to capture temporal information. Convolutional operations are done in both the temporal and spatial dimensions in 3D-CNN models using 3D cubes that are built by aggregating many synchronized frames [22]. However, due to a lack of exact spatiotemporal representation of actions and a lack of diversity in action datasets, the subsequent method fared worse than handicraft features. The temporal information was retrieved by transmuting frames to a reduced resolution to reduce computational complexity [23]. HAR's recognition ability is still comparable to both 3D-CNN and 2D-CNN on spatial video frames. This demonstrates that there is no substantial performance increase in motion information of 3D- CNN when compared to 2D-CNN. Furthermore, the verified findings of 3D-CNN are occasionally lower than those of some handcrafted representations. Following that, motion encoded RGB representations were used for action classification. Deep learning-based action recognition systems, in general, necessitate a large amount of data and high computational resources.

In recent years, the convolutional neural network, one of the most effective deep learning model types, has grown in popularity. It's a bio-inspired hierarchical multi-layered neural network that can learn visual patterns directly from the pixels of image frames [24, 25]. An artificial neural network (ANN) with numerous layers between the input and output layers is referred to as a CNN. Whether it can be a linear or non-linear relationship, it finds the exact mathematical manipulation to make the input into the output.

The Convolutional Neural Network (CNN) [34] is the most widely used technology for improving picture categorization accuracy. CNN is a worldwide utilized image processing and pattern recognition technique that is efficient and successful in recognition, identification, and classification [35]. In terms of image categorization and recognition, AlexNet performs admirably. ImageNet dataset proposed in 2012 and can classify up to1000 objects [36]. The VGGNet has up to 19 trainable layers, improved classification performance proposed in 2014 [37].

Recently, many researchers developed various deep-learning based model for HAR to improve detection accuracy and computational efficiency. Such as, in [1], authors develop a model for human action recognition using CNN to achieve human activity through user Smartphone data. The data are collected by three-axis accelerometer. Here, authors focus on the three human activities such as, sitting, jogging and walking. A body warm sensor based HAR method is proposed in [2]. Here, the authors utilized the different body worm device separately to determine the human action. Afterwards, a CNN based model is proposed in [3]. Where raw data are utilized to train and evaluate the model. These raw-data are collected from inertial sensors. Human activity detection and tracking system is proposed in [4]. The authors combine scale-invariant feature transform with a method to extract features from video sequences. In this work, authors also used optical flow computation for robust feature formation. For tracking, the Gaussian mixture model is used and CNN is used to train and evaluate the datasets.

On the MPII Human Pose Dataset, regression CNN is used for pose estimation and activity classification. MPII human pose dataset is being used by Regression CNN model which achieved near about eighty one percent accuracy. The dataset contains activity like athletics, badminton, baseball, running, and walking, which is very likely compared to our original dataset [5]. RNN fisher vectors and ST-ResNet and IDT are used on UCF101 and HMDB51 dataset respectively and got outstanding accuracy [6]. Fully Connected Networks i.e., FCNs-16 outperforms on same dataset mentioned last in [6]. In this work, UCF-101 has a similar type of accuracy which is almost same. Their activity is also matching the pattern of the previous dataset such as punch, clap, and kick [7]. CNN without Gaussian noise is introduced in this Model for Temporally Organized Joined Location Data and performed well on Cornell activity dataset. Cornell Activity dataset also uses CNN in a work which has the desired output as predicted. Cooking, talking on phone, working on a computer and similar type of activities are being used in that dataset. They basically focused on joint location data [8].

In [9], the author used the Histogram of Motion Intensity and Direction to classify the action from multiple views. Another one uses HOG model in IXMAS dataset, which is also performed better result containing activities like sit down, get up, turn around, and such kind of activities [9]. They worked on IXMAS dataset. Smartphone Data was used in real-time HAR using LSTM. Smartphone data like clapping, waving uses LSTM which also illustrate the dashing result. In previous, authors use wearable device for extracting information from running, walking shows that kind of performance using RNN [10]. Sequential deep learning model shows a greater performance on KTH1 and KTH2 dataset. Some previous work illustrates that they are performing near about ninety five percent accuracy containing dataset KTH1, KTH2, which constructs the whole body of KTH. There are two parts in these datasets. First one has the same actions three to four times, whereas the other one has one action one time only. So why, the first dataset has higher accuracy than the other one [11]. They used LSTM under a 3D-ConvNet. In [12], a CNN model with a Convolutional Auto Encoder is used on KTH dataset classifies six activities from 100 videos containing 25 different individuals. CNN extracts the feature of human activity while SVM classifies them as some category on the existing KTH dataset without dividing it like the previous one [12]. CNN and SVM were used to learn and classify respectively. Temporal activity detection is performed by ActivityNet based on Recurrent Neural Network with Challenge 2016 dataset is performs well. Activities like playing sports instrument and ice fishing are included in Activity Challenge 2016, where both CNN and LSTM model are applied to acquire about eighty seven percent accuracy [13]. Other previous works have less than ninety percent accuracy, which works using the model LSTM. In these papers, CNN have the highest result in terms of accuracy.

3. ORIGINALITY

Human activities which are composed of different and very complex actions are varying in their spatial dynamics and so often include interactions with other humans and objects. As it comprises a large amount of data and powerful computational resources, it is very difficult for a system to recognize human activity, however, that of straightforward and easy for human. According to recent research in human action recognition, different patterns with large amount of data can handle efficiently through the process of deep-learning as it handles the non-linear boundaries. Thus this process reduces the miss-classification rate. However, deep-learning neural network faces the difficulties of vanishing gradient problem, for which, the network is not able to learn the human action pattern properly. To overcome the problems of computational cost and vanishing gradient problem, in this work, a simple CNN model is proposed named as HActivityNet. Although the proposed model is simple, it works better for single and multi-viewed human action recognitions. The model actions are performed by two subtasks. In the first subtask, two conv2D neural networks are used for feature extraction from images. In the second subtask, a fully-connected layer is used followed by Softmax classifier for recognizing activity. In this work, two Datasets are used to overcome the problem of imbalance Dataset named HARDataset1 and HARDataset2. The Dataset HARDataset1 is created from KTH Dataset and HARDataset2 is created from the real-time images.

4. SYSTEM DESIGN

4.1 Datasets

Understanding the complication of the dataset is a crucial part of any learning task. In the action recognition domain, one of the challenges is insufficiency in datasets to train the model in advanced movements of human activities and imbalance in positive and negative data in the dataset for each class label. Available human activity datasets are typically too small and imbalance for CNNs to give significant results. There are some moderate size activity datasets like UCF-101 [6, 7] having 101 different activity classes and 13320 videos, THUMOS15 [8] which contains temporally untrimmed videos, KTH [12] dataset that consists of 6 activity classes with 100 videos per class.

4.1.1 KTH Dataset

Outdoors, outdoors with scale change, outdoors with varied garments, and indoors are the four situations used to build the dataset. There are four situations used to build the dataset that are outdoors, outdoors with different clothes, outdoors with scale variation, outdoors with different clothes and indoors.

The KTH database contains around 2391 sequences. A static camera takes all sequences with *25* fps frame rate over some similar backgrounds. The sequences are sampled to the resolution of 160x120 pixels, and they have four seconds longer on average.

KTH Dataset Link: http://www.nada.kth.se/cvap/actios/

4.1.2 HARDataset1

We construct our first dataset, namely HARDataset1 by converting the videos to frames taken from KTH dataset. It includes seven different types of human actions, including boxing, hand-clapping, hand-waving, jogging, running, standing, and walking, all of which are repeated multiple times by 25 subjects in four different scenarios: outdoors, outdoors with different cloths and scale variation, and indoors. This dataset contains 57441 images with seven classes. Each class contains eight thousand image frames. The outline of HARDataset1 is presented in Table 1. Figure 1 (I) shows several examples of images from the HARDataset1.

4.1.3 HARDataset2

We construct our second dataset HARDataset2 by converting videos to frames. First, we took videos of the seven activities of our own in four different angles: front, two sides, i.e., left and right, and from the top. It also contains seven types of actions, i.e., hand-clapping, boxing, running, standing, hand-waving, jogging, walking performed several times by two subjects in the indoor scenario. The second dataset contains 12046 images with seven classes. The outline of the second HARDataset2 is presented in Table 1. Some sample images from the HARDataset2 are shown in Figure 1(II).

Table 1. Outline of HARDataset1 and HARDataset2

Classes	**HARDataset1**	**HARDataset2**
	No. of Images	**No. of Images**
Boxing	8121	2184
HandClapping	8399	2424
HandWaving	8273	2892
Jogging	8292	605
Running	8015	196
Standing	8196	2709
Walking	8138	1036

4.2 Proposed Model

CNN is a worldwide utilized image processing and pattern recognition technique that is efficient and successful in recognition, identification, and classification. Here, we address the HAR problem as a multi-Label classification problem. We design a single-attribute learning Convolutional Neural Network model namely HActivityNet that explores the relationship between the attributes and predicts all the attributes at the same time.

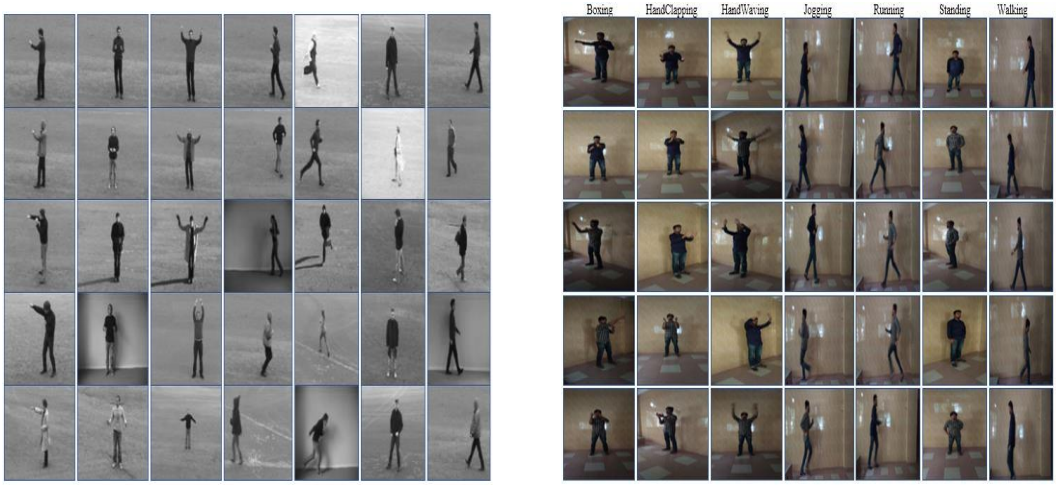


Figure 1. Sample of HARDataset1 and HARDataset2: a) boxing, b) hand clapping, c) hand waving, d) jogging, e) running, f) standing, and g) walking.

4.2.1 HActivityNet

The proposed model is designed inspired by the MiniVGGNet. However, we also concerned on the vanishing gradient problem for which the feature information are only flow in the short distance. To overcome the vanishing gradient problem, here, we have presented a simple CNN model. The proposed model has two sets of (CONV => RELU => BN) * 2 => POOL layers. Batch normalization (BN) and dropout are also included in these layer sets. Pooling layers are used in gradually reducing the input volume's spatial dimensions. For activity recognition we have considered only one fully connected (FC) layer followed by output layer. The proposed convolutional neural network's design is presented in Figure 2.

HActivityNet made up of two parts of CONV => RELU => CONV => RELU => POOL layers. That parts are preceded by the set of FC => RELU => FC => SOFTMAX layers. The first two CONV layers will each learn 64 and 32 filters, with each filter size being 3 by 3. The 3 by 3 filter size is used since the input image (48 x 48 pixels) has less pixel information. The second and third CONV layers will learn 128 and 64 filters, respectively, with each filter having a 3 by 3 filter size. A 0.25 dropout is performed after the first two convolutional layers. After that, max pooling is conducted over a 2 by 2 window with a 1 by 1 stride using the second two CONV and POOL layers. After the activations, HActivityNet features batch normalization layers, as well as dropout layers after the POOL and FC layers. In this study, just one FC layer (512) is taken into account to reduce the computational cost of the proposed model.

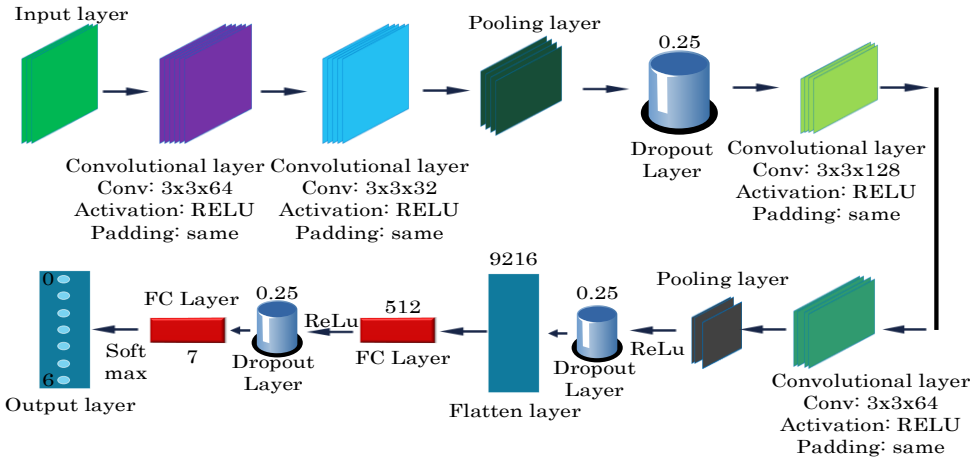


Figure 2. Proposed HActivityNet architecture for human activity recognition.

4.2.2 Data Preprocessing

Before training the model, certain pre-processing activities have to be completed. In the preprocessing step the input image is resized into 48x48 pixels. After that, the datasets images are normalized across (0, 1) range which is beneficial for the training process. After that, split the dataset in the 80:15:5 ratios for training, testing, and validation respectively. Afterward, the HARDataset1 consist 45,952 images for train and 11,489 images for test and validation. In HARDataset2 contains of 9636 images for train and 2410 images for test and validation.

4.2.3 Training and Validation

In order to expedite the training process and get better performance, we load the HActivityNet on HARDataset1. We apply the stochastic gradient descent (SGD) optimizer to train the models with 0.01 learning rate and polynomial decay as learning rate scheduler. In the training and validation process, we use 100 epochs.

5. EXPERIMENT AND ANALYSIS

5.1 Loss Function

Categorical Cross-Entropy Loss, often referred as "logarithmic/ logistic/ log loss" used for multiclass classification. In this process, firstly, predicts the binary features, and then summed the features and averaged across all examples in the dataset. We apply categorical cross-entropy loss function under SGD. The probability of each category is defined in (1).

$$Categorical\ crossentropy\ loss = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \log y_{nk} \quad (1)$$

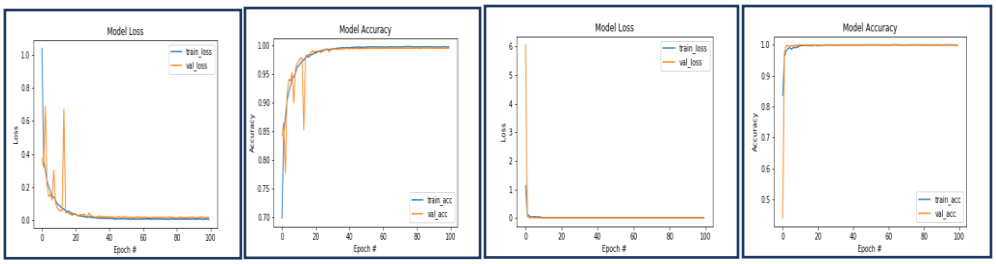
If y(n, k) is large then t(n. k) = 1; which is more probable.

$$Loss = \sum_{k=1}^{k} t^k\ logy^k \quad (2)$$

If $y(n,k)$ is more wrong, loss to be larger and if $y(n,k)$ is more right, loss to be smaller. For Exactly Right: $-1 * log\,(1) = 0.50\%$ probability on correct target: $-1 * log\,(0.5) = 0.693$. 0% probability on correct target: $-1 * log\,(0) = \infty$. The loss and accuracy curves achieved in training with the HARDataset1 and HARDataset2 dataset are shown in Figure 3.

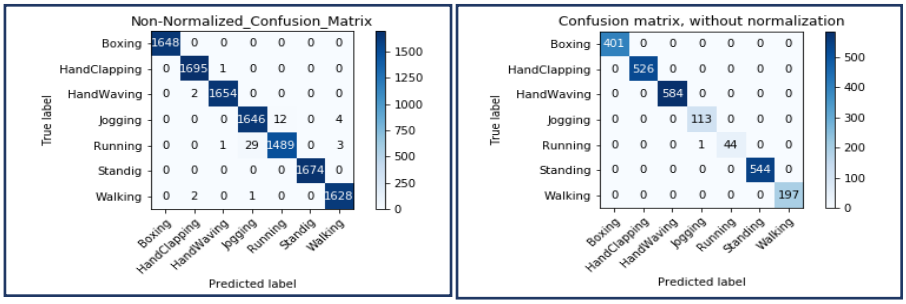
5.2 Confusion Matrix

A confusion matrix is a table that lists the actual and predicted categories in a classification system. Each row represents the projected classes, whereas each column represents the instances of an actual class in a confusion matrix. Figure. 4 show the confusion matrix for both HARDataset1 and DARDataset2.



(a) (b) (c) (d)

Figure 3. Validation loss and accuracy curve: a) and c) loss curve for HARDataset1 and HARDatase2 dataset respectively, b) and d) accuracy curve for HARDataset1 and HARDataset2 dataset respectively.



(a) (b)

Figure 4. Confusion Matrix of HARDataset1 and HARDataset2: a) Non-normalized classification matrix for HARDataset1 dataset and b) Non-normalized classification report for HARDataset2.

5.3 Classification Report

Classification report contains the overall accuracy of the experiment that includes the precision, recall, and F1-Score.

The proportion of correctly identified examples i.e., true positive (TP) and true negative (TN) divided by the total number of instances i.e., true positive (TP), true negative (TN), false positive (FP) and false negative (FN) is called accuracy. The equation (3) represents the accuracy, where, equation (4) represents the precision. Precision is defined as the percentage of relevant examples (TP) in the total number of retrieved instances (TP and FP). The

recall is the percentage of relevant instances (TP) out of the total number of wrongly retrieved instances (TP and FN). The Recall is presented in (5). Precision and Recall are harmonically combined to form the F1-Score. The F1-Score is presented in (6). Moreover, a macro-average calculates the statistic separately for each class and then averages the results. The weighted average, on the other hand, will compute the average metric by combining the contributions of all classes.

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{F1} - \text{Score} = 2\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision + Recall}}\right) \tag{6}$$

The classification report of HARDataset1 and HARDataset2 are demonstrating the overall accuracy, which is shown in Table 2. Where HARD1 represents HARDataset1 and HARD2 represents HARDataset2.

5.4 **Recognition Accuracy**

In the classification report shows the precision, recall, F1-Score and overall accuracy of training and validation of HARDataset1 and HARdataset2 dataset. The model shows satisfactory recognition results for each class labels. In HARDataset1, the model shows almost 100% accuracy for the class labels except jogging, running and walking, which are shown 99% accuracy. Whereas, in HARDataset2, the proposed model also shows almost 100% accuracy for all the class labels except jogging and running. The jogging and running show 99% accuracy. The recognition accuracy of the proposed model for HARDataset1 (HARD1) and HARDataset2 (HARD2) dataset for each class label is shown in Table 3.

Table 2. Precision, Recall, F1-Score of proposed model based on the HARDataset1 (HARD1) and HARDataset2 (HARD2) dataset

	Precision		**Recall**		**F1-Score**	
	HARD1	**HARD2**	**HARD1**	**HARD2**	**HARD1**	**HARD2**
Boxing	1.00	1.00	1.00	1.00	1.00	1.00
HandClapping	1.00	1.00	1.00	1.00	1.00	1.00
HandWaving	1.00	1.00	1.00	1.00	1.00	1.00
Jogging	0.98	0.99	0.99	1.00	0.99	1.00
Running	0.99	1.00	0.98	0.98	0.99	0.99
Standing	1.00	1.00	1.00	1.00	1.00	1.00
Walking	0.99	1.00	1.00	1.00	1.00	1.00
Accuracy					**0.9952**	**1.00**
Average	**0.9951**	**1.00**	**0.9950**	**1.00**	**0.9950**	**1.00**

Table 3. Recognition accuracy of proposed model for each class label with respect to HARDataset1 and HARDataset2

Classes	**Accuracy**		**# of samples**	
	HARD1	**HARD2**	**HARD1**	**HARD2**
Boxing	1.00	1.00	1648	401
HandClapping	1.00	1.00	1696	526
HandWaving	1.00	1.00	1656	584
Jogging	0.99	0.99	1662	113
Running	0.99	0.99	1522	45
Standing	1.00	1.00	1674	544
Walking	0.99	1.00	1631	197

5.5 Comparison of HActivityNet with MiniVGGNet and InceptionV3

The developed model is compared to MiniVGGNet and InceptionV3, two current state-of-the-art deep models. For comparison, both models are trained and evaluated with the two own created dataset HARDataset1 (HARD1) and HARDataset2 (HARD2). The comparison result based on recognition accuracy is presented in Table 4. Table 4 reveals that the developed model, i.e., HActivityNet performs better with compared to the MiniVGGNet and InceptionV3.

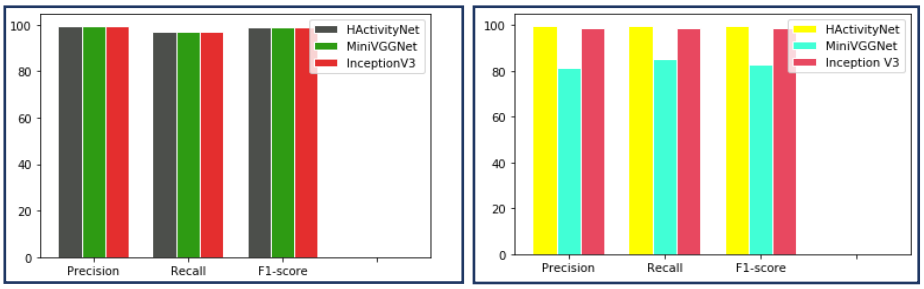
Table 4. Comparison of recognition accuracy with respect to HARDataset1 (HARD1) and HARDataset2 (HARD2)

Classes	HActivityNet		MiniVGGNet		InceptionV3	
	HARD1	HARD2	HARD1	HARD2	HARD1	HARD2
Boxing	1.00	1.00	1.00	1.00	1.00	1.00
HandClapping	1.00	1.00	1.00	1.00	1.00	1.00
HandWaving	1.00	1.00	1.00	1.00	1.00	1.00
Jogging	0.99	0.99	0.97	0.84	0.90	0.97
Running	0.99	0.99	0.98	0.99	0.91	0.94
Standing	1.00	1.00	1.00	1.00	1.00	1.00
Walking	0.99	1.00	0.99	0.99	0.97	1.00

The comparative experimental result of precision, recall, F1-Score of HActivityNet, MiniVGGNet and InceptionV3 based on the HARDataset1 and HARDataset2 dataset are presented here to show the comparison more elaborately. The results are depicted in Table 5. Table 5 shows that all models perform better for the HARDataset1 dataset. The precision, recall, F1-Score and accuracy are almost symmetric. However, for HARDataset2, which is created from the real-time images, all models are not performed well. In this dataset, the performance of MiniVGGNet is much less with compared to the other models. The bar chart of the experimental results illustrates the consecutive accuracy of the HActivityNet over the two models is shown in Figure 5.

Table 5. Precision, Recall, F1-Score of HActivityNet, MiniVGGNet and InceptionV3 based on the HARDataset1 (D1) and HARDataset2 (D2) dataset

	Filter Configuration	Dense Layers	Total No. of Parameters	Precision		Recall		F1_Score		Accuracy	
				D1	D2	D1	D2	D1	D2	D1	D2
HActivityNet	**64-32-128-64**	**512**	**4,855,783**	**0.99**	**1.00**	**0.99**	**1.00**	**0.99**	**1.00**	**0.99**	**1.00**
MiniVGGNet	32-32-64-64	512-512	4,790,503	0.97	0.81	0.97	0.85	0.97	0.83	0.97	0.98
InceptionV3	32-64-128	512-512	22,854,887	0.98	0.99	0.98	0.97	0.98	0.99	0.98	0.99



(a) (b)

Figure 5. Bar chart of the experimental result: (a) accuracy of three models with respect to HARDatase1dataset and (b) accuracy of three models with respect to HARDatase2 dataset.

We have implemented some human action recognition models which were worked on the different activity dataset. According to the results of the experimental investigation, it is reveals that, our proposed model outperforms the examined models. Table 6 demonstrates the experimental result of two models along with our proposed model, both worked on KTH Dataset. Our HARdataset1 is derived from KTH Dataset and the comparison shows that our proposed model outperforms the other two models worked on KTH Dataset.

Also a CNN model works on the Smartphone-based three-axis accelerometer dataset [1] has gained the accuracy of 91.97%. In [4], the authors proposed a model that works on the Weizmann and Kungliga Tekniska dataset has achieved the accuracy of 98.43% and 94.96% respectively. The regression CNN [5] model worked on MPII Human Pose dataset and retained 80.51% accuracy. Afterwards, RNN fisher vectors, and ST-ResNet and IDT [6] [7] both worked on UCF (101) and HMDB51 datasets and achieved the accuracy of 94.6% and 70.3% respectively. Among the others, Cornell Activity Datasets [8], IXMAS [9], Smartphone Data [10], KTH1 and KTH2 [11], KTH [12] and ActivityNet Challenge 2016 [13] gain the accuracy of 87%, 83.03%, 88.60%, 94.39%, 92.49% and 76.76% respectively. Table 7 shows a comparison of the suggested model to the current state of the art.

Table 6. Comparison accuracy of HActivityNet with some state-of-the-art models

Models	**Accuracy**
Proposed method	**HARDataset1 (99.52%)** **HARDataset2 (100%)**
[11]	KTH1 (94.39%), KTH2 (92.17%)
[12]	KTH (92.49%)

Table 7. Comparison accuracy of HActivityNet with some state-of-the-art models

Models	**Accuracy**
Proposed method	**HARDataset1 (99.52%)** **HARDataset2 (100%)**
[1]	Three-axis accelerometer dataset (91.97%)
[4]	Weizmann dataset (98.43%) Kungliga Tekniska dataset (94.96%)
[5]	MPII Human Pose (80.51%)
[6]	UCF101 (94.6%), HMDB51 (70.3%)
[7]	UCF101 (93.0%), HMDB51 (70.2%)
[8]	Cornell Activity Datasets (87%)
[9]	IXMAS (83.03%)
[10]	Smartphone Data (88.60%)
[11]	KTH1 (94.39%), KTH2 (92.17%)
[12]	KTH (92.49%)
[13]	ActivityNet Challenge 2016 (75.76%)

5.6 **Discussion**

From 11489 validation images, our model classifies correctly almost all the images. Only a few images are misclassified. As our model is trained with enough data and the model is regularized model. For that reason, the model performs better both in the HARDataset1 and HARDataset2, which are created form KTH dataset and from real-time images respectively. Some samples of the classification result are shown in Figure 7 with the corresponding confidence score.

In the testing phase, we have used 70 images from 7 activities. We have taken ten images for each dataset outside of the datasets. With HARDataset1, our model misclassifies two images that are “handwaving” is predicted as “handclapping” which is shown in Figure 8(a), and jogging is predicted as running which is shown in Figure 8(b). Furthermore, from the second dataset, HARDataset2, our model misclassified only an image that is running is

predicted as jogging which is shown in Figure 8(c). As said earlier, HAR is a hard problem due to so many limitations such as, symmetric human body structures in different activities, background disorder, variations in viewpoint and changes in scale, partial blockade, illumination causes by various lighting conditions, appearance and resolution of the frames that affect recognition accuracy.

If we see the misclassified images, the "handclapping" and "handwaving" images are almost symmetric. While considering the running, jogging and walking images, we see that the body structure of the actor is also almost identical. In outdoor pictures, we can see the shadow of the human. This situation also leads the activity as the other one. We know from the viewpoint of an observer; sometimes it is very difficult to recognize a person whether he is running and jogging. Coming to a machine, it is far more difficult due to its so many limitations. However, we are trying to overcome the limitations and make the machine more accurate over time.

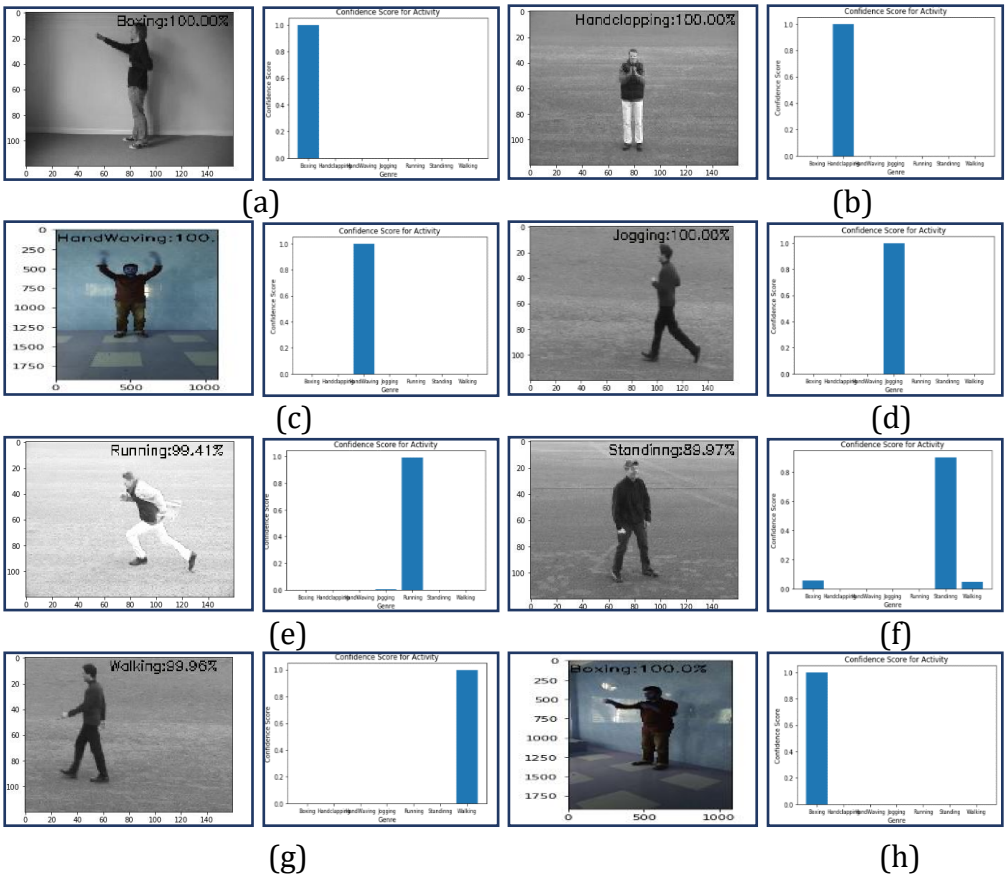
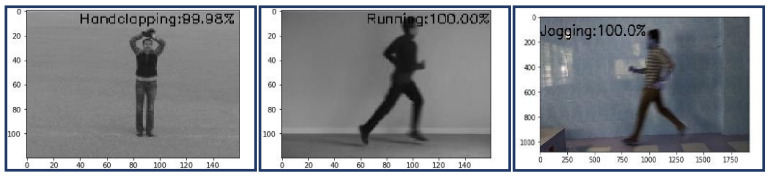


Figure 7. Classification result of some sample images: a) Boxing, b) Hand Clapping, c) Hand Waving, d) Jogging, e) Running, f) Standing, g) Walking and h) Boxing.



(a) (b) (c)

Figure 8. Some sample of misclassified images: a) misclassified HandWaving as HandClapping, b) misclassified Jogging as running, and c) misclassified running as Jogging.

Our model is trained on the homogeneous object backgrounds as both of our dataset have homogeneous object background. To check the efficiency of the proposed model, some heterogeneous object backgrounds images have been tested with the proposed model. The test result of the heterogeneous object background images is presented in Figure 9. From the experiment it is revealed that our model is able to predict the activities from the environments where backgrounds have heterogeneous objects.

Figure 9. Some experimental examples of heterogeneous object background images: a) running b) standing, c) boxing, d) handWaving, and c) Jogging.

6. CONCLUSION

In this work, a Human Activity Recognition Network (HActivityNet) is developed to recognize the rapid, precise and consistent human activity in an image. HActivityNet is a simple Convolutional Neural Network (CNN) based model developed inspired by MiniVGGNet to enhance the recognition accuracy and decrease the training and validation loss with decreasing the vanishing gradient problem. It is susceptible of accurately recognizing human activities in a variety of circumstances and perspectives. To train and evaluate the proposed model with balance positive and negative data, two different datasets, i.e., HARDataset1 and HARDataset2 on seven different activities i.e., boxing, hand-clapping, hand-waving, running, standing, jogging, and walking have been created. These datasets are also used to evaluate MiniVGGNet and InceptionV3 for human action recognition. The proposed model shows better results compare to these models. A comparison is made between the suggested model and various current state-of-the-art models, and suggested model adequate significant improvement with compared to others. The proposed model reveals the accuracy of 99.5% on HARDatase1 and almost 100% on HARDataset2.

REFERENCES

[1] Xu, W. , Pang, Y., Yang, Y., and Liu, Y., "**Human Activity Recognition Based On Convolutional Neural Network**," *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, pp. 165-170, 2018,

[2] Moya Rueda, F., Grzeszick, R., Fink, G.A., Feldhorst, S. and Ten Hompel, M., "**Convolutional neural networks for human activity recognition using body-worn sensors**," In Informatics, Vol. 5, No. 2, p. 26, 2018.

[3] Bevilacqua, A., MacDonald, K., Rangarej, A., Widjaya, V., Caulfield, B. and Kechadi, T., "**Human activity recognition with convolutional neural networks**," In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 541-552, 2018, September, Springer, Cham.

[4] Basavaiah, J. and Patil, C. M., "**Human activity detection and action recognition in videos using convolutional neural networks**," Journal of Information and Communication Technology, Vol. 19, No. 2, pp. 157-183, 2020.

[5] Bearman, A., & Dong, C. "**Human pose estimation and activity classification using convolutional neural networks**," CS231n Course Project Reports, 2015.

[6] Koohzadi, M., & Charkari, N. M. "**Survey on deep learning methods in human action recognition**," IET Computer Vision, Vol. 11, NO. 8, pp. 623-632, 2017.

[7] Yu, S., Cheng, Y., Xie, L., & Li, S. Z. "**Fully convolutional networks for action recognition,**" IET Computer Vision, Vol. 11, NO. 8, pp. 744-749, 2017.

[8] Jayabalan, A., Karunakaran, H., Murlidharan, S., &Shizume, T. "**Dynamic Action Recognition: A convolutional neural network model for temporally organized joint location data**," arXiv preprint arXiv:1612.06703, 2016.

[9] Chun, S., & Lee, C. S. "**Human action recognition using histogram of motion intensity and direction from multiple views**," IET Computer vision, Vol. 10, No. 4, pp. 250-257, 2016.

[10] Milenkoski, M., Trivodaliev, K., Kalajdziski, S., Jovanov, M., & Stojkoska, B. R. "**Real time human activity recognition on smartphones using LSTM Networks**," In 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1126-1131, 2018, May, IEEE.

[11] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A., "**Sequential deep learning for human action recognition**," In International workshop on human behavior understanding , pp. 29-39, 2011, November, Springer, Berlin, Heidelberg.

[12] Geng, C., & Song, J. "**Human action recognition based on convolutional neural networks with a convolutional auto-encoder**," In 2015 5th International Conference on Computer Sciences and Automation Engineering ICCSAE 2015. 2016, February. Atlantis Press.

[13] Montes, A., Salvador, A., Pascual, S. and Giro-i-Nieto, X., "**Temporal activity detection in untrimmed videos with recurrent neural networks**," arXiv preprint arXiv:1608.08128, 2016.

[14] Zhu, F., Shao, L., Xie, J. and Fang, Y., "**From handcrafted to learned representations for human action recognition: A surve**y," Image and Vision Computing, Vol. 55, pp.42-52, 2016.

[15] Laptev I., "**On space-time interest points**," International Journal of Computer Vision, Vol. 64, No. 2, pp. 107-23, 2005.

[16] Kovashka, A. and Grauman, K., "**Learning a hierarchy of discriminative space-time neighborhood features for human action recognition**," In 2010 IEEE computer society conference on computer vision and pattern recognition, pp. 2046-2053, IEEE, 2010.

[17] Murtaza, F., Yousaf, M.H. and Velastin, S.A., "**Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description**," IET Computer Vision, Vol. 10, No. 7, pp. 758-767, 2016.

[18] Chaaraoui, A.A., Climent-Pérez, P. and Flórez-Revuelta, F., "**Silhouette-based human action recognition using sequences of key poses**," Pattern Recognition Letters, Vol. 34, No. 15, pp. 1799-1807, 2013.

[19] Orrite, C., Rodriguez, M., Herrero, E., Rogez, G. and Velastin, S.A., "**Automatic segmentation and recognition of human actions in monocular sequences**," In 2014 22nd International Conference on Pattern Recognition, pp. 4218-4223, IEEE, 2014.

[20] Wang, H. and Schmid, C., "**Action recognition with improved trajectories**," In Proceedings of the IEEE international conference on computer vision, pp. 3551-3558, 2013.

[21] Wang, Y. and Mori, G., "**Human action recognition by semilatent topic models**," IEEE transactions on pattern analysis and machine intelligence, Vol. 31, No. 10, pp. 1762-1774, 2009.

[22] Ji, S., Xu, W., Yang, M. and Yu, K., "**3D convolutional neural networks for human action recognition**," IEEE transactions on pattern analysis and machine intelligence, Vol. 35, No. 1, pp.221-231, 2012.

[23] Memisevic, R. and Hinton, G., "**Unsupervised learning of image transformations**," In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, IEEE, 2007.

[24] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., "**Gradient-based learning applied to document recognition**," Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278–2324, 1998.

[25] LeCun, Y., Kavukcuoglu, K., Farabet, C., **Convolutional networks and applications in vision**, In IEEE International Symposium on Circuits and Systems, pp. 253–256, 2010.

[26] Clarkson, B.P., "**Life patterns: structure from wearable sensors** (Doctoral dissertation, Massachusetts Institute of Technology), 2002.

[27] Ojala, T., Pietikainen, M. and Maenpaa, T., "**Multiresolution gray-scale and rotation invariant texture classification with local binary patterns**," IEEE Transactions on pattern analysis and machine intelligence, Vol. 24, No. 7, pp. 971-987, 2002.

[28] Dalal, N. and Triggs, B., "**Histograms of oriented gradients for human detection**," In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Vol. 1, pp. 886-893, IEEE, 2005.

[29] Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., "**ORB: An efficient alternative to SIFT or SURF**," In 2011 International conference on computer vision, pp. 2564-2571, IEEE.

[30] Guo G., Wang H., Bell D., Bi Y., Greer K., "**KNN Model-Based Approach in Classification**", Meersman R., Tari Z., Schmidt D.C. (eds) On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM2003. Vol. 2888, pp: 986-996, 2003.

[31] Anagnostopoulos, G.C. "**SVM-Based Target Recognition From Synthetic Aperture Radar Images using TargetRegion Outline Descriptors**," Nonlinear Analysis: Theory, Methods & Applications, Vol. 71, Issue. 12, pp:2934–2939, 2009.

[32] YoshuaBengio, "**Learning Deep Architectures for AI**", Foundations and Trends® in Machine Learning, Vol.2, pp. 1-127, 2009.

[33] Schmidhuber, J., "**Deep learning in neural networks: An overview**," Neural Networks, Vol. 61, pp. 85 –117, 2015.

[34] Sudharshan, D.P. and Raj, S., "**Object recognition in images using convolutional neural network**," In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 718-722, IEEE, 2018.

[35] Safiyah, R. D., Rahim, Z. A., Syafiq, S., Ibrahim, Z., & Sabri, N, "**Performance Evaluation for Vision-Based Vehicle Classification Using Convolutional Neural Network**,"International Journal of Engineering and Technology (UAE), Vol. 7, pp: 86-90, 2018.

[36] Krizhevsky, A., Sutskever, I., Hinton, G.E, "**Imagenet Classification with Deep Convolutional Neural Networks**," Proceedings of the Neural Information Processing System (NIPS), Harrahs and Harveys,Lake Tahoe, NV, USA, Vol.2, pp: 1097-1105, 2012.

[37] Simonyan, K., Zisserman, A, "**Very Deep Convolutional Networks for Large-Scale Image Recognition**," Conference paper at ICLR 2015, arXiv:1409.1556.

[38] Gomathi, V. "**Indian Sign Language Recognition through Hybrid ConvNet-LSTM Networks**," EMITTER International Journal of Engineering Technology, Vol. 9, No. 1, pp. 182-203, 2021.