# HUMAN ACTIVITY RECOGNITION BY DEEP LEARNING BASED CNN

[1]S. AKHILA, [2]B. NARESH

[1,2]Assistant Professor, Dept of CSE, Megha institute of engineering and Technology for women, Ghatkesar (T.S)

***Abstract***: *This project aims to develop a model of human activities such as running, walking, jogging, clapping, hand waving, and boxing. A set of planning videos is provided, with one person running an event in each video. The movement in that particular video can be the naming of a video. This data must be discovered by release, and you can expect an input (video) tag you've never seen. Technically speaking, despite the descriptions of these actions, the copy may want to learn to distinguish between many human behaviours. Various content identification software may work in later work, such as actively monitoring objects to identify items, including a vehicle or a human from a CCTV image, and learning the movement patterns of objects. Humans can create a pattern. To guide us (people) to practice a variety of sports. In this paper, type of deep learning model convolutional neural network (CNN) is proposed for HAR that can act directly on the raw inputs. In addition, an efficient pre-training strategy has been introduced to reduce the high computational cost of kernel training to enable improved real-world applications.*

***Keywords***: *Human action recognition, deep learning, convolutional neural network.*

## I. INTRODUCTION

Due to progress on computer vision, computers improve on the resolution of some very difficult problems (such as understanding an image). Models are made where the model can predict what the image is or can detect whether or not a specific object is present in the image if an image is given to the model. These models are known as neural networks (or artificial neural networks) inspired by a human brain structure and function. Deep learning, a subfield of machine learning is the study of these neural networks, which over time have introduced several variations of these networks for various problems. For Video Recognition, this approach utilizes deep learning - in the context of a number of labelled images, a model is built so that it can generate a prediction label for a new video. Steps have been taken for execution are downloading, extracting and pre-processing a video dataset then dividing the dataset into training and testing data then creation of a neural network and train it on the training data finally testing the model on the test data [1].

In the past decade, Human Action Recognition (HAR) has become an increasingly attractive subject of study with various programs, video surveillance, virtual events, intelligent human-laptop interactions, and many others. However, correct motion recognition is very difficult due to the chaotic backgrounds, occlusions, and variations of point of view. The Humanitarian Action Report consists of several scores describing jobs that identify low-stage activities or actions. A common description of the popularity of human motion from a set of images involves two steps: 1) extracting complex hand-drawn features from raw video frames and 2) build a classifier based on these features Some of the commonly used features for human action recognition are Histogram of Oriented Gradient (HOG) [2].

In these capabilities, some of the commonly used functions of human motion reputation are directed gradient graph (HOG), optical graph flow (HOF), motion exchange patterns (MIP), spatiotemporal points of interest (STIP), bank properties and dense tracks. However, these tactics are difficult and time-consuming to extend these features to other systems. Many manual design functions are implemented through work, and exclusive tasks can also use unique capabilities. But in fact, it is very difficult to recognize the type of feature that is vital to a particular challenge, so the selection of features depends specifically on the particular difficulty. Especially for human movement recognition, sports-specific figures show many differences in appearance and movement version. It is very difficult to obtain the basic attribute of movement within the radical change of environment. Therefore, a known feature extraction technique should be proposed to mitigate the need for hand-drawn features and reduce the computational scale. CNN is a deep copy that obtains complex hierarchical functions through a convolutional operation that alternates with a subsampling process on the raw input images [3]. It is confirmed that CNN can benefit from more outstanding performance in visual target reputation responsibilities through appropriate adjustment during education. And CNN has fixed mode stability, illumination, and chaotic ambient switching. The first attempt at HAR, using CNN, turned into developing a unique 3D version of CNN that extracts features from each spatial and temporal dimension by performing 3D convolutions, thus capturing motion statistics encoded in multiple contiguous frames. The upgraded version produced more than one information channel from the input frames, and the final distinct illustration was obtained by combining data from all channels. The model progressed by occurring one-by-one partial convolution and down sampling on grayscale pixel values, horizontal gradient, vertical gradient, horizontal optical float, and vertical optical drift channels extracted from adjacent input frames using wire layers. The authors suggested using CNN's multiple decision architecture and integrating temporal facts for human action recognition into a UCF-101 database using raw video as input. Proposed a deep convolutional neural network architecture for recognizing human motions in movies using the FFA capabilities of the UCF50 database. Proposed a unique version of the dynamic neural community that can understand the dynamic patterns of visual images of human movements based on learning. Convolutional neural networks (CNN) and multiscale recurrent neural networks (MTRNN) have been added. The authors proposed a new technique that combines partial models with deep learning through natural CNN learning. Although CNN is an excellent choice for HAR, this technique still has the disadvantage that the cores/weights employed in convolution are taught via BP neural networks, which can be time-consuming. In this paper, to solve this annoyance of HAR, which is mainly dependent on CNN, a convolutional autoencoder (CAE) pre-learning method is proposed. This approach discovers excellent CNN initialization that avoids many fantastic minima of fantastically non-convex destination functions that arise in almost all deep mastering questions [4].

## II. LITERATURE SURVEY

Human activity reputation is one of the challenging and attractive topics in information technology. Typically, there are processes in the reputation of human activity, technologies based entirely on manual features, and technologies based on features extracted from a vehicle. A lot of research has been done in the first institution.

### “A Large Video Database for Human Motion Recognition “

With nearly a billion movies online considered ordinary, the popularity and search for video represent a rising new frontier in laptop fiction and clairvoyance studies. While much effort has gone into collecting and annotating huge, scalable still-image datasets containing hundreds of image categories, human motion datasets are lagging. Existing motion recognition databases are combined with ten unique action categories accumulated under fairly controlled conditions. The recent performance of these datasets is now close to its limit, so it may be desirable to design and deliver the latest standards. To address this issue, we've compiled the largest motion video database to date with 51 motion categories, totalling around 7,000 hand-drawn illustrations drawn from a branch of assets ranging from digital film to YouTube. We use this database to evaluate the overall performance of a consultant's computer vision systems to learn about procedures and to discover the strength of these methods under numerous conditions including camera movement, point of view, interesting video, and closure.

**Real-world Anomaly Detection in Surveillance Videos**

Surveillance films can capture a range of realistic aberrations. In this article, we propose anomaly analysis with the help of unconventional and daily video exploitation. To avoid annotations on syllables or anomalies in school films, which are time-consuming, we recommend examining the anomaly through the deep rating system for a few cases with the help of using low-rated school films, for example. The instructional labels (Normal or Normal) are at the video level rather than at the clip level. In our approach, we consider strange and everyday films such as briefcases and video clips as cases in multi-instance learning (MIL) and mechanistically study a deep anomaly classification model that predicts high anomalous ratings of anomalous video segments. In addition, we introduce temporal smoothness and scattering constraints within the classification loss function to improve localization of anomalies in the training context. We also present a new, first-of-its-kind, large-scale dataset of 128 hours of video. Includes 1,900 full-length, uncut real-time international surveillance movies with 13 practical anomalies, including combat, lane incidents, break-ins, entry, theft, and many more. This dataset can be used for two commits. First, preferred anomaly detection considers all anomalies in one group and all normal activities in another. Second, to expose each of the thirteen odd sports. Our experimental traces show that our MIL approach to anomaly detection achieves a significant improvement in performance compared to more advanced methods. We present the results of many of the most recent baselines of in-depth knowledge about the reputation of queer interests. The low overall performance of recognizing these known baselines shows that our data set is very challenging and opens up additional opportunities for targeted work.

## III. SYSTEM ANALYSIS

In the existing work with wearable based or non-wearable based. Wearable based HAR system make use of wearable sensors that are attached on the human body. Wearable based HAR system are intrusive in nature. Non-wearable based HAR system do not require any sensors to attach on the human or to carry any device for activity recognition. Non-wearable-based approach can be further categorized into sensor based HAR systems. Sensor based technology use RF signals from sensors, such as RFID, PIR sensors and Wifi signals to detect human activities. Sensor based HAR system are non-intrusive in nature but may not provide high accuracy. The are various disadvantages in the literature as given below.

Require the optical sensors to be attached on the human and also demand the need of multiple camera settings and Wearable dives cost are high

## IV. RESEARCH METHODOLOGY

The proposed System Vision based technology use videos, image frames from depth cameras or IR cameras to classify human activities. Video-based human activity recognition can be categorized as vision-based according to motion features. The vision-based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human. Therefore, this methodology is getting more consideration nowadays, consequently making the HAR framework simple and easy to be deployed in many applications. The most common type of deep learning method is Convolutional Neural Network (CNN). CNN are largely applied in areas related to computer vision. It consists series of convolution layers through which images are passed for processing.

The main goal of this paper is to implement human action recognition from the given videos. We segment the video footages and imported the video frames as the input data. Three deep learning methods are applied to generate feature maps for human action recognition. Throughout network training, we recognize human actions and finally export the class tags. A. CNN Model is a class of feedforward neural networks, which are principally comprised of input layer, convolutional layer, pooling layer, full connection layer, and output layer. The convolutional layer of a CNN encompasses one or more feature planes. Each feature plane is related to numerous neurons in a region, the neurons in the same plane share the same weights. The shared weights consist of network parametric set, the better weights are gained in the process of model training. By extracting local features and synthesizing them at a higher level, CNNs not only yield global features but also lessen a number of neuron nodes. At this point, the number of neurons is still very large, by setting the weight for each neuron equally, the number of network parameters will be greatly diminished. On the first convolution layer, the output is $y_m$ then the output after $k$ times of convolution operations is

$$y_k^m = \delta(\Sigma_{y_i^{n-1} \in M_k} y_i^{m-1} * W_{ik}^m + b_k^m)$$

(1)

where $\delta(.)$ is an activation function, $M_k$ is based on a layer of feature collection, $W_{ik}^m$refers to convolution kernels, $*$ means a convolution, $b_k^m$ stands for offset.

In CNNs, the pooling layer follows the convolutional layer to reduce dimensionality and increase community learning convergence. The difference is to leave out redundant capacities to avoid overfitting. Each neuron within the full communication layer is connected to all neurons in the subsequent layer. All the proximal features are combined in all communication to form the general capabilities. Each neuron in the full communication layer performs an activation function, which is transferred to the output layer. With CNN, the temporal information of a particular video cannot be used at all. The CNN model is shown in Figure 1.

The video is divided into individual frames in CNN to form a huge image dataset. This set was imported as the introduction of a single channel CNN for pre-training. The consequences of education are saved, and job sequences are established. The dataset is then imported into the community as input information. The collection of video frames is used to educate the network community. After marking, the CNN parameters are exported as spatial features of human motion trajectories.

## SYSTEM ARCHITECTURE

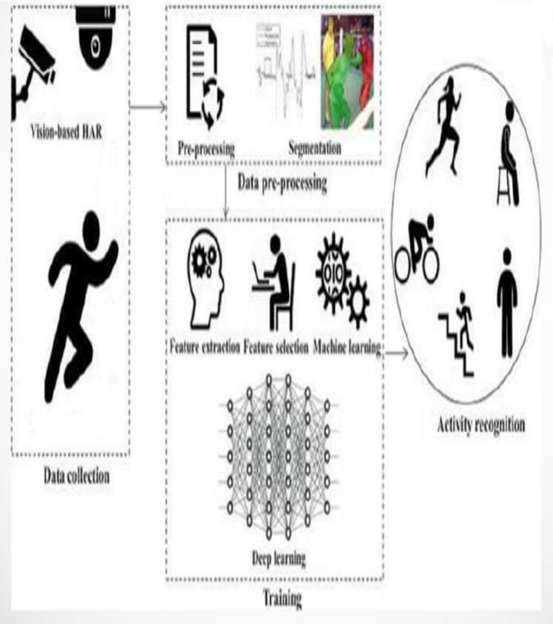


Fig.1 System architecture

## HAR System

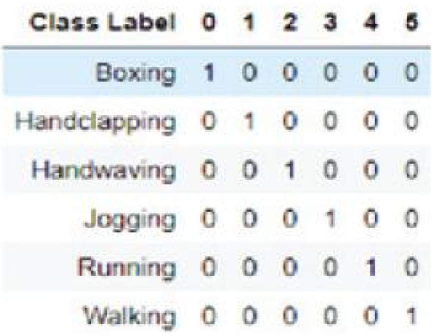
Video-based human activity recognition can be categorized as vision-based according. The vision-based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human. Therefore, this methodology is getting more consideration nowadays, consequently making the HAR framework simple and easy to be deployed in many applications. We first extracted the frames for each activity from the videos. Specifically, we use transfer learning to get deep image features and trained machine learning classifiers.

**VGG16:**

VGG16 is a convolutional neural network model. Deep Convolutional Networks for Large-Scale Image Recognition". The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous models submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU's.

**Transfer Learning:**

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems. In this post, you will discover how you can use transfer learning to speed up training and improve the performance of your deep learning model.



| Class Label | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Boxing | 1 | 0 | 0 | 0 | 0 | 0 |
| Handclapping | 0 | 1 | 0 | 0 | 0 | 0 |
| Handwaving | 0 | 0 | 1 | 0 | 0 | 0 |
| Jogging | 0 | 0 | 0 | 1 | 0 | 0 |
| Running | 0 | 0 | 0 | 0 | 1 | 0 |
| Walking | 0 | 0 | 0 | 0 | 0 | 1 |

**After One-hot Encoding**

## V. RESULTS

We must configure the parameters given below for each convolutionary layer.

**i. Filters:** The number of characteristic maps required for this convolutionary layer output.

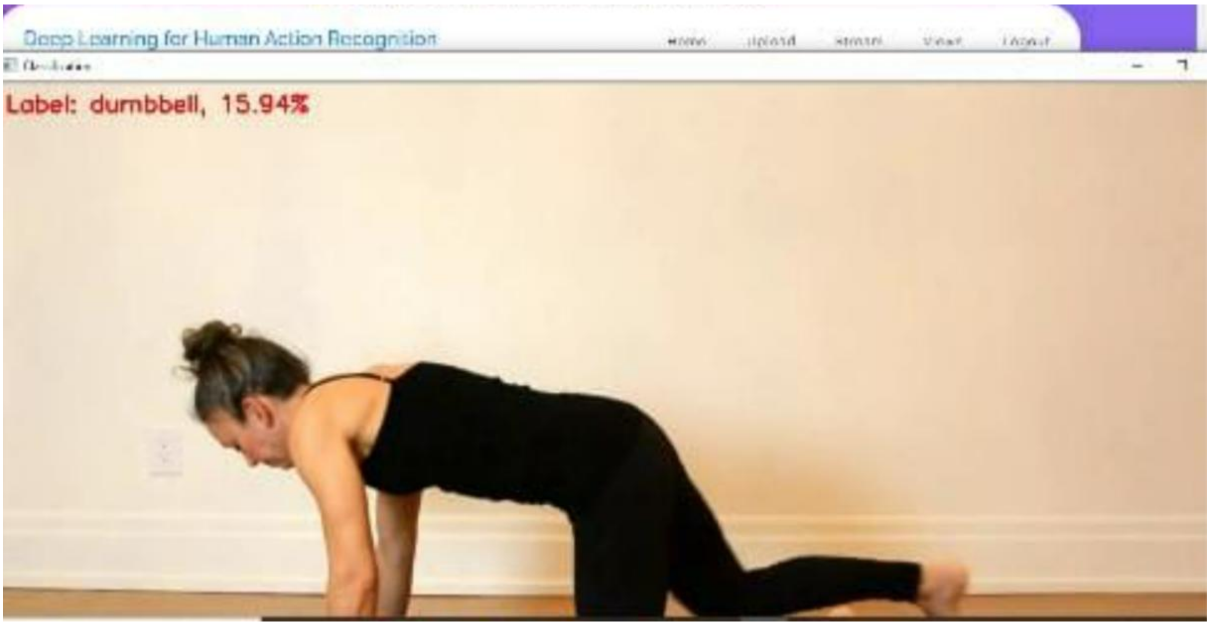
**ii. Kernel size:** window size that is to be converted into a single feature map along all axes of the input data.

**iii. Strides:** the number of pixels that should shift through the convolutionary window.

**iv. Padding:** to determine what happens on the borders– either the input is cropped (valid) or the input is padded to the same dimension (same), with zeros.

**v. Activations:** The enabling function for that layer to be used. (ReLU works best with deep neural networks due to its non-linearity and the ability to prevent the disappearance of gradients)

The video data set was loaded and the necessary Pre-processing measures were one of the most critical parts of the project. Therefore, we created a class (Videos) which was renamed (read videos)) (to play and process images. It was very difficult to create this because we worked on generalizing this feature for any type of video (not unique to that project). We used NumPy (wherever) to store and process the videos (with an additional functionality much faster than the built-in python lists). The model would face such a problem as a major challenge. The solution over such problem is obtained by applying Image AI algorithm which leads to sort the frames having human body present in selected frame.

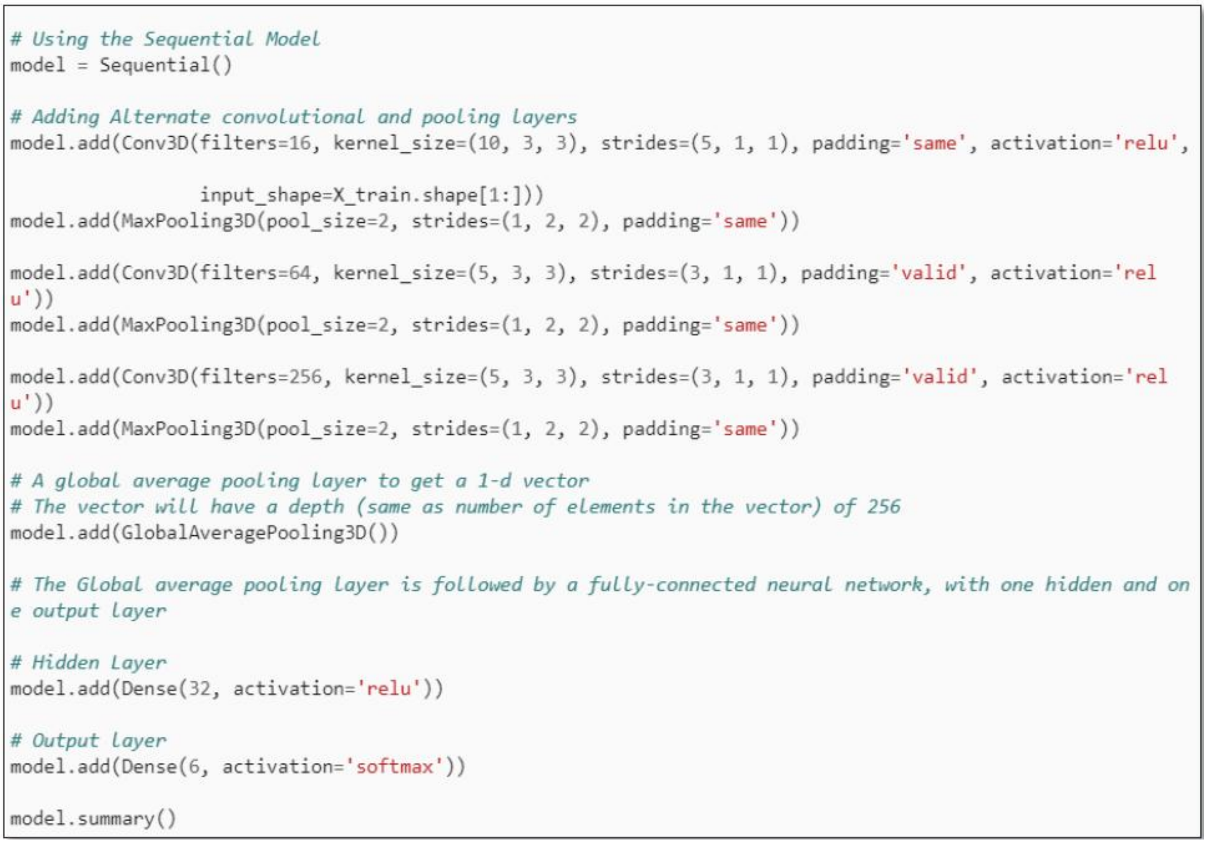


Label: dumbbell, 15.94%

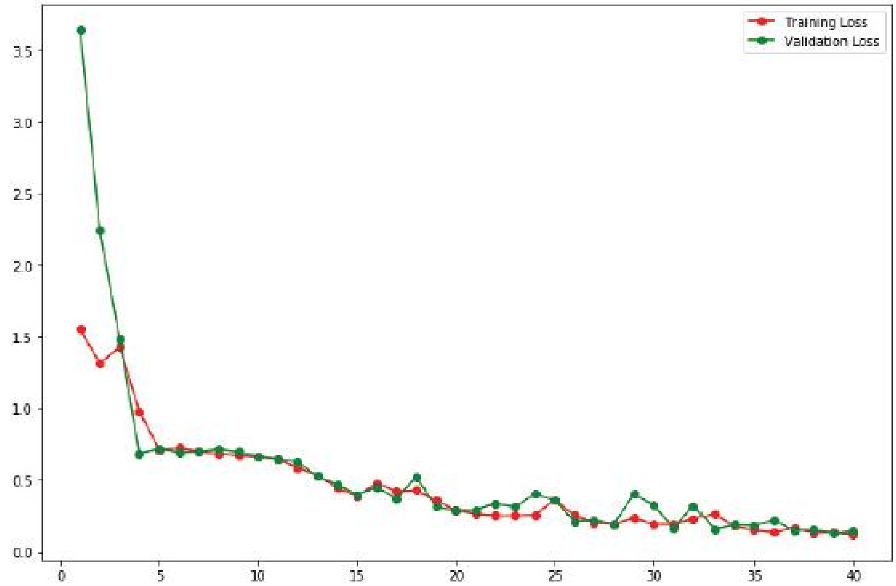
**Fig.2** label: dumbbell. 15.94%

This combination of alternating convolution and pooling layers follows a global pooling layer. If the spatial data are no longer left in the input (the input cannot be further reduced by pooling layers in the spatial dimension), the Global layer converts the input to a 1-d vector (with the same depth). This 1- dimensional vector is then used as the input for a dense neural network that is fully connected. There are several hidden layers and an output layer in the fully connected network. The activation function like SoftMax can be used for the output layer, which gives the chance that the input belongs to each class. Finally, this input is assigned to the class label with the highest probability.

The proposed model was trained on the training data for 40 epochs. The weights of the model which gave the best performance on the validation data were loaded. The model was then tested on the test data. The model gave an accuracy of 86.21% as compared to 64.5% of base paper on the test data. This model gave a higher accuracy than the previous models, using Image AI API for training. In this model, a pair of convolutional and max pooling layer was added.



```
# Using the Sequential Model
model = Sequential()

# Adding Alternate convolutional and pooling layers
model.add(Conv3D(filters=16, kernel_size=(10, 3, 3), strides=(5, 1, 1), padding='same', activation='relu',

                input_shape=X_train.shape[1:]))
model.add(MaxPooling3D(pool_size=2, strides=(1, 2, 2), padding='same'))

model.add(Conv3D(filters=64, kernel_size=(5, 3, 3), strides=(3, 1, 1), padding='valid', activation='rel
u'))
model.add(MaxPooling3D(pool_size=2, strides=(1, 2, 2), padding='same'))

model.add(Conv3D(filters=256, kernel_size=(5, 3, 3), strides=(3, 1, 1), padding='valid', activation='rel
u'))
model.add(MaxPooling3D(pool_size=2, strides=(1, 2, 2), padding='same'))

# A global average pooling layer to get a 1-d vector
# The vector will have a depth (same as number of elements in the vector) of 256
model.add(GlobalAveragePooling3D())

# The Global average pooling layer is followed by a fully-connected neural network, with one hidden and on
e output layer

# Hidden Layer
model.add(Dense(32, activation='relu'))

# Output layer
model.add(Dense(6, activation='softmax'))

model.summary()
```



,

**Fig.4** Screenshot of results produced by Model built using Keras library in Python

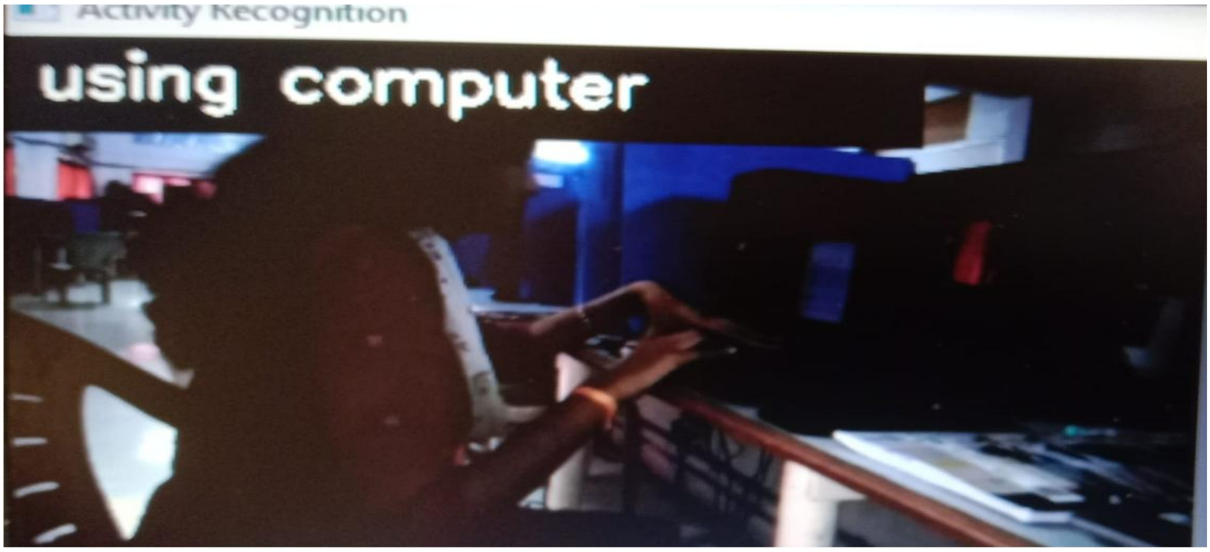


Fig.5 Activity recognition that using computer

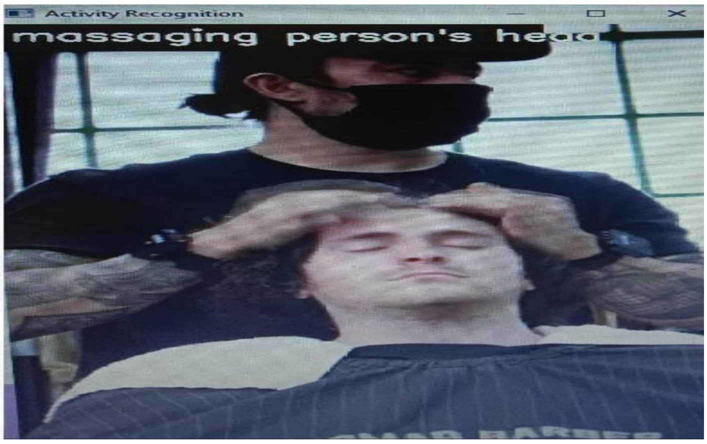


Fig.6 Activity recognition that massaging person's head

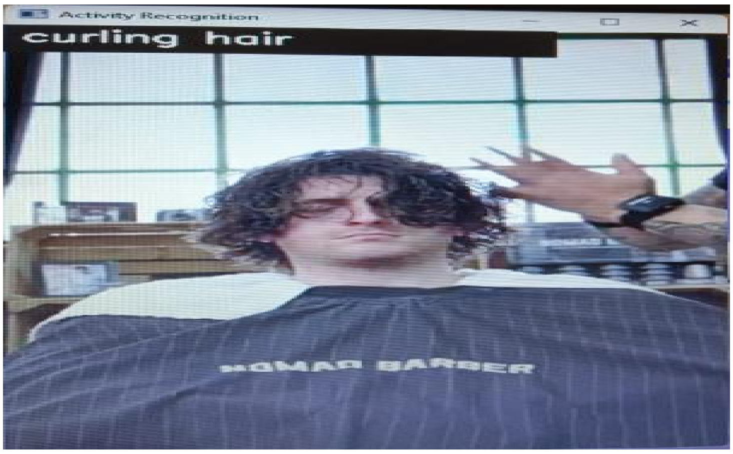


Fig.7 Activity recognition that curling hair

## VI. CONCLUSION

In this paper, we presented a CNN model for the HAR problem. We focused on a set of activities extracted from a common exercise program for fall prevention, training our model data sampled from different sensors, in order to explore the classification capabilities of each individual unit, as well as groups of units. Our experimental results indicate that convolutional models can be used to address the problem of activity recognition in the context of exercise programs.

Many standard datasets are available for video analysis to validate designed model's accuracy. In recent times it is practically possible to build a good model using very high capacity and complex libraries like Keras, Theano and Torch [6] on Python platform to make machines intelligent, the proposed model achieved nearly 20% more accuracy by Pre-processing the dataset as compared to the base model. Further work on the application of convolutional model to real-world data is recommended. More activities could be included in the workflow, and different aggregations on the activities can be tested.

## REFERENCES

[1] Learn Computer Vision Using OpenCV - With Deep Learning CNNs and RNNs | Sunila Gollapudi | Apress. .

[2] “Video Dataset Overview.”: https://www.di.ens.fr/~miech/datasetviz/.

[3] W. Sultani, C. Chen, and M. Shah, “Real-world Anomaly Detection in Surveillance Videos,” ArXiv180104264 Cs, Feb. 2019.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587756.

[5] M. Jain, MrinalJain17/Human-Activity-Recognition. 2019.

[6] “Keras vs TensorFlow vs PyTorch | Deep Learning Frameworks,” Edureka, 05-Dec-2018. https://www.edureka.co/blog/keras-vstensorflow- vs-pytorch.

7. Burns, A., Greene, B.R., McGrath, M.J., O'Shea, T.J., Kuris, B., Ayer, S.M., Stroiescu, F., Cionca, V.: ShimmerTM a wireless sensor platform for non-invasive biomedical research. IEEE Sensors Journal 10(9), 1527 { 1534 (2010). https://doi.org/10.1109/JSEN.2010.2045498

8. Cook, D., Feuz, K.D., Krishnan, N.C.: Transfer learning for activity recognition: a survey. Knowledge and Information Systems 36(3), 537{556 (Sep 2013). https://doi.org/10.1007/s10115-013-0665-3, https://doi.org/10.1007/s10115-013-0665-3

9. Godfrey, A., Conway, R., Meagher, D., Laighin, G.: Direct measurement of human movement by accelerometry 30, 1364{86 (01 2009)

10. Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: 2016 International Joint Conference on Neural Networks (IJCNN). pp. 381{388 (July 2016). https://doi.org/10.1109/IJCNN.2016.7727224.

11 Prasadu Peddi (2018), Data sharing Privacy in Mobile cloud using AES, ISSN 2319-1953, volume 7, issue 4.