Human Activity Recognition

Manidhar Kodurupaka
Computer Science
Bennett University
City, India
mk2473@bennett.edu.in

K. Natesh Reddy
Computer Science
IIIT,Nagpur
Nagpur, India
natesh1199@gmail.com

Akash Wadhwa M
Information Technology
Rajalakshmi Engineering College
Chennai, India
akashwadhwa.m.2017.it@rajalakshmi.edu.in

KONDURU CHANDRA HARSHITHA
Computer Science
S.R.K INSTITUTE OF TECHNOLOGY
Vijayawada, India
chandraharshitha51@gmail.com

Balmukund Mishra
Computer Science
Bennett University
Delhi, India
BM7477@bennett.edu.in

Shambhavi Mishra
Computer Science
Bennett University
Delhi, India
Sm7835@bennett.edu.in

***Abstract*—Human action recognition is a very important topic in computer vision due to its many applications such as video surveillance, human and machine interaction and video retrieval autonomous driving vehicle, entertainment etc.Video analysis tasks have seen great variations and it has been moving from inferring the present state to predicting the future state. This has been possible with the developments in the field of Computer Science and Machine Learning.As per vision-based action recognition on human is been varied on the different set of actions been performed,the machine needs to lean all the possible actions can be performed.It also helps in prediction of future state of the human by inferring the current action being performed by that human.Human Action Recognition can be used in numerous applications.In this report, we will be seeing the details of the work done by us and the methodology adopted for the same. We will go through the techniques used to implement a system that is suitable for human activity recognition. We will also compare the different models used for action recognition and note which one is better than other. Finally we will choose the model which provides us an better accuracy.**

I. Introduction

One of the most interesting HAR is high-level behaviour recognition a moment of body is considered an action. From the viewpoint of computer vision, recognition of action is to match the observation with previously dened patterns or algorithms and then assign it a label. Depending on complexity, human activities can be separated into four types: gestures, actions, interactions and group activities, and in many cases where human action is been used the bottom-up construction is been used. Many main components include feature extraction, action learning and classication, and action recognition and segmentation. For instance, to recognize Hand shakes activities, two person's arms and hands are rst detected and tracked to generate a spatial-temporal description of their movement. This description is compared with existing patterns in the training data to determine the action type and provide us with similar results.

Detecting features and working just on spatial region is not enough for HAR but we also need to examine features many kinds of action including local and global features based on, the methods are been represented as local and global temporal and spatial changes temporal and spatial changes, trajectory features based on key point tracking, motion changes based on depth information, and action features based on human pose changes.The result varies accordingly,and the image representation and action classification are been used separately,and the recognition problem is been divided into two parts the recognition problem into action and activity according to the complexity and different approaches need to be handled by their varying degrees of complexity.

The main purpose of this project is to build an automatic system, which when given a video, will be able to classify the action being performed in the video. To achieve this purpose, we have built a model that will be able to classify those videos. To build that model, we required a big data set that we have used to train our model and then use that model to present a solution to above classification problem. The purpose of HAR is that it can be very useful in the society it can also be operated by installing the software into a drone and that surveillance camera can recognize the activities by the human and it can check with the data sets it is been trained by and can make an alarm or even notify the nearby surroundings based on the video,and the data sets are not pre-trained data sets because a human can perform enormous number of activity and every activity can be varied the data set cannot be trained completely that is the reason we have also included live stream data sets so that whenever a drone spots any activity it will relate with the most accurate activity and work accordingly. The HAR is been trained and tested with many types of different algorithms and data sets as mentioned above, after training the data sets the algorithm which provides us with most accurate result.

II. Related Work

In recent years, crowd video analysis is done by using the crowd video classification methods. A crowd behaviour can be classified in terms of its components. [1] they have shown that

Identify applicable funding agency here. If none, delete this.

the mid-level descriptors of the groups can be used to classify the videos of crowd and analyzes. Also considering the whole crowd as a single entity similar accuracy to classify videos can be obtained. The various different tasks of detecting the group, computing their features and then combining them to define the crowd can be reduced to single step of computing the features of the crowd. The model uses traditional average/max pooling based strategies worked globally on the entire video but in an efficient feature pooling based representation of videos for action recognition, we propose a two-level video representation which takes care of global video representation and also takes care of features extracted from a set of some local snippets. In both the cases, efficient pooling techniques are applied to the difference vectors of consecutive frames. In order to obtain the snippets a self tuned spectral clustering technique has been employed and also this system is able to capture global and localized characteristics of the action under consideration and also reduces the effects due to irrelevant contents sharply. Experimental results obtained on the challenging UCF-50 and KTH datasets establish the robustness of the proposed encoding[2]. Saliency means useful. So, saliency detection means detecting the useful data. In this paper they divide the video frames into different areas according to their importance and the saliency rate, and model the static and motion information simultaneously. In this approach, the adaptive weights are learned for each class very specifically. This framework has achieved a better performance and both the streams are expected that they will improve video classification results according to our experiments done on UCF-101 and CCV datasets [3].

The Convolution RNN method considers various convolution features to recognize the emotion in wild using 1-minute gradual dataset samples. In [4] this method, training done on large (aff-wild) datasets. Using task-variant framework, it fetches all the features from high to low) and computes prediction with their mean values. Efficiency is improved by fusing of various networks and leads accurate results. It receives third rank in 1-minute emotion challenge, mainly in technologies. Mainly this method obtains second rank in technologies used category. This method produces accurate results than state-of-the art methods for visual modality. The proposed approach, further extended to improvement in arousal estimation results.

In [5] Computer vision field. many developments are implemented using large standard datasets. Mainly, in machine learning technology having large sets of open-source libraries for implementation and also less expensive hardware is possible to obtain better results in novel methodologies. By using Image net, we can use large datasets for image processing mainly in image understanding. In video classification, there is no barrier for datasets because of CCTV utilization in many places in which we need to monitor activities. In this model, they get various features from multi-label classification dataset having large amount of datasets and annotated a vocabulary of visual entities (4800) in you-tube 8m. Main, this method used labels for main topics in those videos. This method generated machine base labels with good precision and used for content based techniques. It also uses manual and automatic techniques for labels and decodes second based frame for each video. This approach, used DCNN based Image net for providing input to classification layer for training and features are compressed and also combined labels with compressed features to make further processing. This dataset having all frame level features for eight million videos and video frames ratio of 1:9. Several models are trained and evaluated various quality metrics and processes results while comparing with existing methods. While running with methods, some models gives good results in single machine with large datasets. It achieves MAP 77.6% eventhough existing reaches 53.8. Further extended in terms of diversity and scale of you tube 8m for extension in video understanding field. In this model, [6] uses temporal feature pooling (tfp) for producing high accuracy even with more number of fames having large data inputs. This technique tfp is related to number of frames. Long short term memory is less sensitive than tfp. Here, various count of frames obtains tfp that has to pool to obtain dimensions range. Long short term memory is slowly increases performance when contrast to tfp in terms of its input. so, it is not that many datasets are reason to good results, there it must be the importance various factors are reflecting their results. The vgg16 gives reasonable results with minimum (10) number of frames.

A model presents a user-independent deep learning-based approach for online human activity classification. By using Convolutional Neural Networks for local feature extraction together with simple statistical features that preserve information about the global form of time series. Furthermore, investigate the impact of time series length on the recognition accuracy and limit it up to one second that makes possible continuous real-time activity classification. The accuracy of the proposed approach is evaluated on two commonly used WISDM and UCI datasets that contain labeled accelerometer data from 36 and 30 users respectively, and in cross-dataset experiment. The results show that the proposed model demonstrates state-of-the-art performance while requiring low computational cost and no manual feature engineering [7]. With recurrent neural network (rnn) and dcnn (deep convolution neural network) is used to train with available data and also other information retrieved from various networks which similar in nature with suitable databases. To minimize loss, extra information is included in terms of its features and also performance is increased with various datasets without losing the knowledge which is learned from dataset of images. It is tested, in identifying facial expression and emotion recognition for better performance. The results are reasonably good with comparison of state of the art methods[8]. The high natural sound representations are used to capitalize for unlabled large amounts of sound data. using general synchronization in between sound and vision to represent and learn by using 2million videos which are un labelled. The benefit here is that, it acquired economically at huge dataset of data with various scales. By using student teacher training to transfer visual knowledge which is different in various vision models. It leads to sound modality with video as a bridge which is unlabeled. This method gives high semantics in siund network which is trained eventhough with our ground truth. This approach gives better results than variious sound model results in classifying object as well as screen in standard manner [9]. For image classification approaches, the convolution neural networks are establiished successful class of models. These are imroved versions of varoious methods which gives empirical results in large scale classification of videos in 1million viideos from you tube of 487 classes. It uses several methods to perform things in time domain of convolution neural network and by considering benefit of taking into spation temporal information iin local regions. It gives resonable good results with comparison to obtain better accuracy which is 63.9% mainbly in one-frame models it produces 60.9% results by using UCF 101 data set of images.

III. Simulation Part

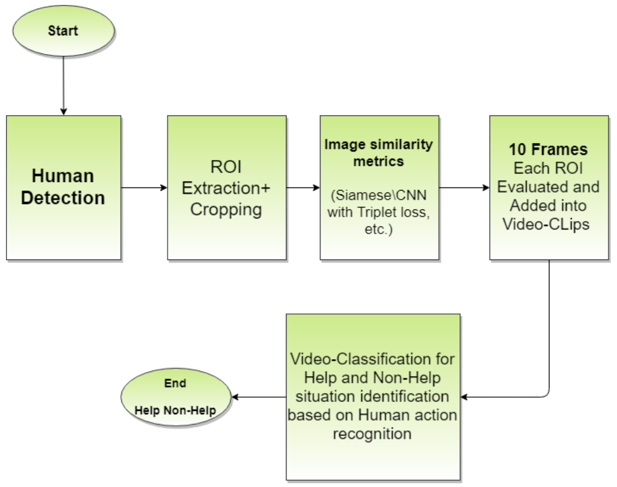

Start
Human Detection
ROI Extraction+ Cropping
Image similarity metrics
(Siamese\CNN with Triplet loss, etc.)
10 Frames
Each ROI Evaluated and Added into Video-CLips
Video-Classification for Help and Non-Help situation identification based on Human action recognition
End
Help Non-Help

Fig. 1. Complete Architecture of Search and Rescue based on Background Invariant Generalization of Deep learning models

In this above figure we have explained the complete mechanism of the human activity recognition,and the complete project is been divided into two equal half and we are using drone for this particular project as a drone is been installed with the required software to detect the human activity.The drone starts the process as it is been turned On and then it begins to search for any human and once it detects the human the software first checks for the human identity with both the data sets UCF-101 and the data set which we had created.Once the human is been identified it moves to the second stage which is ROI Extraction and cropping,basically ROI is known as Region of Interest after the human is been detected and then the camera finds for a particular region where it searches for a particular place where human is been performing an activity,activity can also be defined as a set of actions and the ROI works on that region where that particular action is been is been performed and now cropping is been activated as the region is been detected then that particular region is been cropped and moved forward for further detection of the cropped region.Now after cropping the region of interest we move forward to the image similarity matrices, over here we detect multiple human beings in a single frame and crop their region of interest and collect their activities which are being performed by them, And they are been stored in a similarity stack of matrices. The images are been classified into different categories based on their activities and this results in multiple stacks. After storing the activities we move forward to convert the images into video clips. And the procedure for converting the ROI images into video clips is been implemented by one of our algorithms. The videos are been classified into different classes on the basis of the similarity of the images which were stored in the stack.From here our second end of our project is been initiated, after storing the video clips in different classes and each class consist of similar videos, now with the help of these classes we can train our model to learn these videos in many different algorithms and check for which provides us with a better accuracy, and the data sets which are taken are UCF-101 and Help Non-Help video clips. All the videos are based on human detection and recognition. Now we have got the complete flow of how to proceed with the HAR detection and recognition. Now we can check the pro's and con's of each algorithm in detail.

We are moving forward with four algorithms, They are as follows -

1) 3D ResNet
2) CNN
3) 3D CNN
4) RNN and CNN

A. 3D ResNet

The reason for this investigation is to decide if current video datasets have adequate information for preparing exceptionally profound convolutional neural systems (CNNs) with spatio-transient three-dimensional (3D) bits. As of late, the exhibition levels of 3D CNNs in the field of activity acknowledgment have improved essentially. Notwithstanding, until this point, regular research has just investigated moderately shallow 3D models. We analyze the designs of different 3D CNNs from moderately shallow to profound ones on current video datasets. In view of the aftereffects of those investigations, the accompanying ends could be gotten: (I) ResNet-18 preparing brought about critical overfitting for UCF-101, HMDB-51, and ActivityNet however not for Kinetics. (ii) The Kinetics dataset has adequate information for preparing of profound 3D CNNs, and empowers preparing of up to 152 ResNets layers, strikingly like 2D ResNets on ImageNet. ResNeXt-101 accomplished 78.4% normal precision on the Kinetics test set. (iii) Kinetics pretrained straightforward 3D designs outflanks complex 2D structures, and the pretrained ResNeXt-101 accomplished 94.5% and 70.2% on UCF-101 and HMDB-51, separately. The utilization of 2D CNNs prepared on ImageNet has delivered critical advancement in different assignments in picture. We accept that utilizing profound 3D CNNs together with Kinetics will follow the fruitful history of 2D CNNs and

ImageNet, and animate advances in PC vision for recordings. The codes and pretrained models utilized in this investigation are openly accessible.

B. *Convolutional Neural Network*

The focal point of this calculation is to viably use profound Convolutional Neural Networks (CNNs) to propel occasion location, where just casing level static descriptors can be separated by the current CNN toolboxs. This paper makes two commitments to the surmising of CNN video portrayal. Initially, while normal pooling and max pooling have for some time been the standard ways to deal with conglomerating outline level static highlights, we show that presentation can be fundamentally improved by exploiting a suitable encoding technique. Second, we propose utilizing a lot of idle idea descriptors as the edge descriptor, which improves visual data while keeping it computationally moderate. The coordination of the two commitments brings about another cutting edge execution in occasion identification over the biggest video datasets. Contrasted with improved Dense Trajectories, which has been perceived as the best video portrayal for occasion recognition, our new portrayal improves the Mean Average Precision (mAP) from 27.6% to 36.8% for the TRECVID MEDTest 14 dataset and from 34.0% to 44.6% for the TRECVID MEDTest 13 dataset.

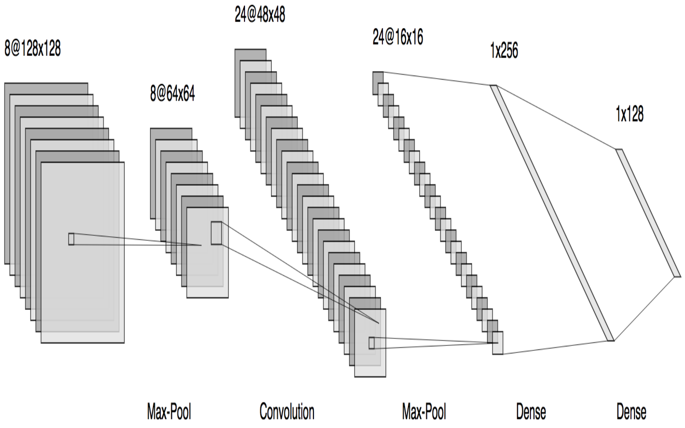


Fig. 2. CNN Model

C. *3D CNN*

We pre-prepared a solidified 2D CNN N1 and haphazardly introduced a 3D CNN N2 utilizing the strategy. It ought to be noticed that in this methodology it is essential to accurately instate the loads of the system. Trials have demonstrated that with standard introduction of the loads of the system, the misfortune won't diminish. We at that point made a two-stream coordinate with N1 and N2 as independent streams, linked last FC layers of the two systems which is thusly associated with two FC layers with 512 and 128 sizes (fc1 , fc2) and to the last parallel classifier layer. We utilized a basic parallel (0/1) coordinating classifier: given a couple of X outlines — to choose whether the sets have a place with a similar class or not. NB: unique strategy use connection of N1 and N2 FC layers, yet our investigations have demonstrated that total contrast works better. An incredible bit of leeway of this technique is that it doesn't require the nearness of named information.

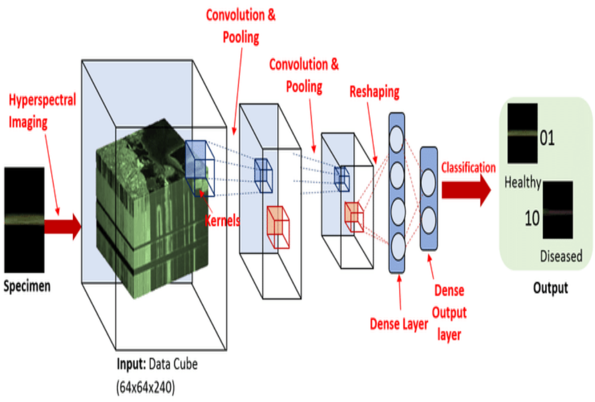


Fig. 3. 3D Model

As we referenced before, one of the issues in the activity acknowledgment task is a current irregularity in datasets: in various datasets similar classes can be marked in an unexpected way. Standard datasets (UCF-101, Kinetics, Help Non-Help, and so forth.) use recordings rejected from video stages (Youtube, Flickr, Vine, and so forth.), or from motion picture scenes.

We physically picked various classes from the UCF-101 and Help Non-Help datasets. With this strategy we got around 200 classes of human day by day activities. You can see the perception of our outcomes in Figure 4. A few activities covered. For instance, 'high fiving' and 'shaking hands' wound up in a similar class, yet in general the system arranged activities particularly.

D. *CNN and RNN*

CNN can straightforwardly recognize the visual example from the first picture and it needs next to no pretreatment work [2]. CNN can at the same time concentrate and train an assortment of highlights of PC vision and with the removed highlights, CNN can copy the procedure of individuals' acknowledgment of the photos. At this point, CNN has been effectively applied to written by hand character acknowledgment [3], face acknowledgment [4] and so forth. In late year, Google's GoogLeNet [5] makes the nature of picture acknowledgment and item location demonstrated at a sensational pace. In this paper we will portray how our model handle the above issues to perceive activities. We depict how our model progressively pools convolutional highlights and how to utilizing these successions to digest the highlights for

understanding the recordings.
We have tried different initiation nets, leftover nets and plain nets, and have watched steady wonders. The all out profound neural systems for activity acknowledgment is appeared in Figure 2 which contains the accompanying advances:

1) use different deep convolution neural networks to extract various features of the images;
2) reshape the feature matrix to sequence;
3) set the feature matrix as the input of multilayered recurrent neural networks;
4) use softmax regression classifier to classify the videos.

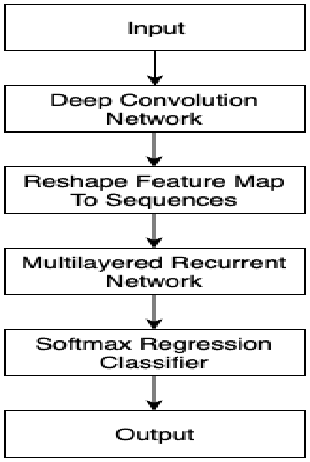

Input
Deep Convolution Network
Reshape Feature Map To Sequences
Multilayered Recurrent Network
Softmax Regression Classifier
Output

Fig. 4. The process of the model for action recognition

Convolutional neural networks We extract the last convolutional layer obtained by inputting the video frames through kinds of GoogLeNets. The last convolution layer has K convolution maps and the features form a feature cube with the shape of W*H*K. After convolutional neural network, we use a down sampling method such as average pooling with kernel size W* H by padding method of 'VALID' to make sure the feature cube to 1-dimensional vector(1*1*K). The 3-dimensional matrix can easily reshape to 1-dimensional vector, and the vector is the feature of the picture.
As for the detail of convolutional neural networks, we use a variety kind of residual and inception blocks. All the convolutions not marked with 'V' in the figures are same-padded meaning that their output size is same as the original input. And if the kernel size not equals to 1*1, we use the ReLU activation function to the layer. The 1*1 convolution without activation is used for controlling the filter's size to fit the depth of input and output.

The LSTM has muddled elements that enable it to effortlessly "retain" data for an all-encompassing number of time steps. The "long haul" memory is put away in a vector of memory cell. A LSTM cell contains an info entryway, an overlook door and a yield door.

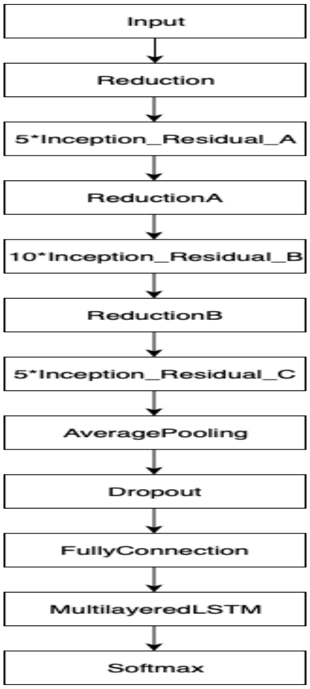


Fig. 5. Overall architecture of our model.

The LSTM cell used in this paper. Each gate's activation function is sigmoid function. By a lot of practical evidence, in fact, we can't get high precision with single layer LSTM cells based RNN. So we choose to use multilayered RNN which is shown in Figure 8, and we denote the state of last LSTM cell in last layer as the final state, which is used to classify actions.

E. Data Set

As we have understood all the four concepts completely, now we can move forward to understand which data sets can be used to train and test the data set,in our algorithm we are using two types of data sets UCF-101 and the help non-help data set which we had created using the drone.
The dataset used in building our model is UCF101. UCF101 is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories. This data set is an extension of UCF50 data set which has 50 action categories.
With 13320 videos from 101 action categories, UCF101 gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc, it is the most challenging data set to date. As most of the available action recognition data sets are not realistic and are staged by actors, UCF101 aims to encourage further research into action recognition by learning and exploring new realistic action categories.
The videos in 101 action categories are grouped into 25 groups, where each group can consist of 4-7 videos of an action. The videos from the same group may share some com-

mon features, such as similar background, similar viewpoint, etc.
The action categories can be divided into five types:

1) Human-Object Interaction
2) Body Motion Only
3) Human-Human Interaction
4) Playing Musical Instruments
5) Sports.

In the fig 6 we can have a view of the image as their are humans been detected and they are waving their hand.

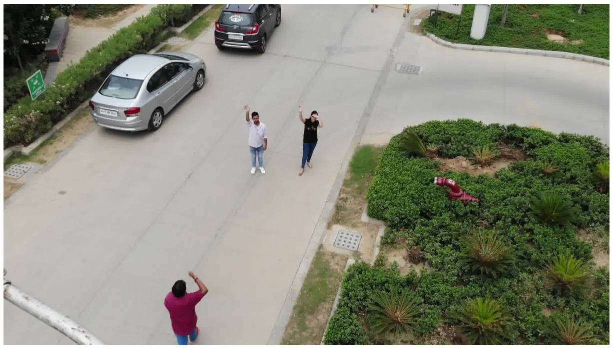


Fig. 6. Extracted Frames

After detecting multiple humans in a single frame we crop down the images by using the ROI concept and we can see that one human is not performing any activity and another human is performing a set of activities.

Fig. 7. Stacked Frames

Now after identifying the humans it displays how much accurate is the human being been detected.

IV. Results

After running the HAR in different algorithms we have trained and tested different data sets including UCF-101 and the Help, Non-help data set which is been created. And we gain the final accuracy of both the data set by following all the algorithms.

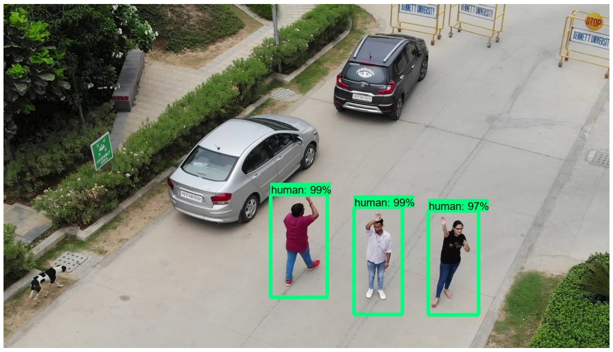


Fig. 8. Predicted result

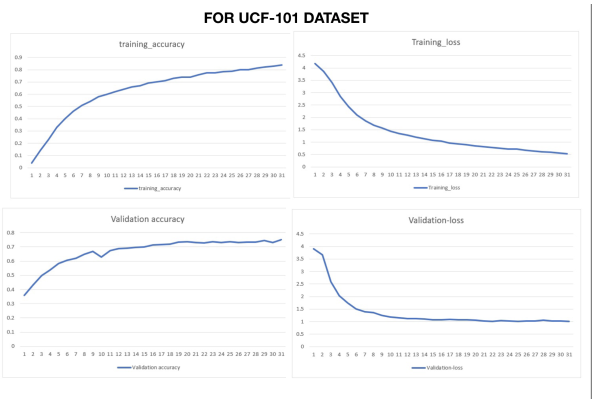


Fig. 9. Accuracy Graph on UCF-101 Data Set

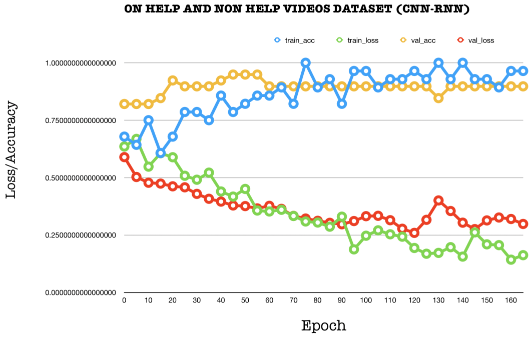
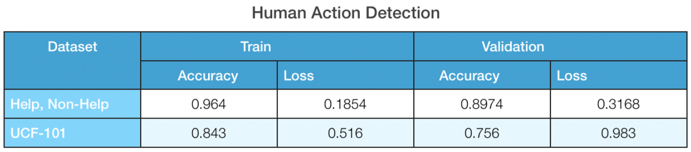


Fig. 10. Accuracy Graph on Help Non-Help Data Set



Human Action Detection

Dataset	Train		Validation	
	Accuracy	Loss	Accuracy	Loss
Help, Non-Help	0.964	0.1854	0.8974	0.3168
UCF-101	0.843	0.516	0.756	0.983

Fig. 11. Accuracy table

V. Conclusion

With the model and architecture we have designed, the project works fine with the dataset we have created using drones.

This can be further taken into real-time deployment of this model as the accuracy and prediction was satisfied and could be useful during the disasters.

The CNN-RNN algorithm provides us with an accuracy of 96.4

References

[1] Zhongwen Xu, Yi Yang, Alex G. Hauptmann; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1798-1807

[2] Chen Zhao, JunGang Han, Xuebin Xu Xi'an, CNN and RNN Based Neural Networks for Action Recognition, Xi'an, Shanxi 710121, China, awp4211@gmail.com

[3] Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh, Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, Submitted on 27 Nov 2017 (v1), last revised 2 Apr 2018 (this version, v2).

[4] Kirill Zhingalov, Machine Learning Specialist at Neurodata Lab, Elizaveta Zaitseva, SMM Specialist at Neurodata Lab. , Real-time Action Recognition using a 3D CNN, unpublished.

[5] Aman Pandey (2017csb1127) Amit Srivastava (2017csb1189)IIT Ropar,Punjab, Human Action Recognition

[6] Dimitrios Kollias and Stefanos Zafeiriou quot;Exploiting multi-CNN features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Datasetquot; oct 2019.