NVEO
Natural Volatiles &
Essential Oils

Deep Learning-Based Real-Time Multiple-Person Action Recognition System

[1]A.Sivasangari,[2]P.Ajitha, [3]R.M.Gomathi, , [4].Ananthi,[5]Nirmalrani

[1]Associate Professor, Sathyabama Institute of Science and Technology,Chennai.
[2]Asst Professor, Sathyabama Institute of Science and Technology,Chennai.
[3] Asst Professor, Sathyabama Institute of Science and Technology,Chennai.
[4] Asst Professor, Sathyabama Institute of Science and Technology,Chennai.
[5] Associate Professor, Sathyabama Institute of Science and Technology,Chennai.
*[1]sivasangarikavya@gmail.com

Abstract:
Recognition of human activity has attracted growing interest from researchers, mainly because of its potential in areas such as surveillance videos, robotics, interaction with human computers, user interface design, and multimedia. Since it includes details about a person's identity, personality, and psychological state, it is difficult to remove. The human ability to recognise another person's behavior is a primary subject of study in the scientific fields of computer vision and machine learning. People's actions are determined by their behaviors, which makes assessing the problem of determining the underlying action extremely difficult. The proposed system would identify and track a large number of people arriving on the scene before recognizing their behavior. Due to the improved resolution of the video frames, when persons in the scene become too far away from the camera, we build a zoom-in feature to produce more suitable action identification results.The work proposed is based on the behavior recognition model of Deep Belief Networks and Deep Boltzmann Machines.

Keywords: *Machine Learning, Deep Belief Network, Deep Bottzmann Machines*

1 Introduction

As imaging technology improves and the camera system improves, new approaches to activity recognition emerge. By examining both representative conventional and state-of-the-art literature, this study aims to provide a comprehensive overview of video-based human behaviour identification, as well as an overview of various approaches and their developments. The underlying hierarchical structure of human behaviours reveals their various levels, which can be thought of as a three-level categorization. Second, there is an atomic element at the most fundamental level, and it is these simple actions that make up more complex human activities. Finally, dynamic interactions form the top level.An action is a series of human body movements and these movements appear to be sequential. Intuitively, behaviour recognition means remembering a series of acts. Action recognition refers to action recognition from the perspective of computer vision.Learning a series of video sequences to classify the sequence of movements associated with a specific action and to predict a potential action based on the associated movements using the information gained.

In the classification of human behaviour in recordings, several essential techniques have been used. For several years, kernel-based classifiers have been widely used in a variety of applications, often resulting in major improvements in presentation. Because of its consistent performance across a variety of tasks, SVM is the most commonly used algorithm for high-level classification of video events. The core components of a HAR framework are feature extraction, behaviour modelling, and identification. In a robust HAR scheme, different inputs can be implemented. Body joints can produce stronger features than features of the entire body, which may represent more robust HAR.

Some techniques combine the technique of hand crafted characteristics and deep learning algorithms, an example is suggested. Feature descriptors based on learning appear to be more beneficial than handcrafted feature descriptors. The explanation is that descriptors focused on learning have the potential to learn additional features that handcrafted function descriptors can not encode.

We consider the possibility and the difference in the frames (video) of using features from the frames. We introduce the behaviour recognition model of Deep Belief Networks and Deep Boltzmann Machines on the basis of this consideration.Chapter 2 is a discussion of related jobs. Chapter 3 addresses the different steps in the proposed algorithm that are involved. Chapter 4 demonstrates us the experimental results and a brief conclusion is offered to us in Chapter 5.

2 Related Work

The In [1] M.Z.Uddin and J.Kim proposed a HAR Methodology which consists of acquisition of video Body parts segmentation by Random forests, generation of features, and DBNN modelling. Shape of the body is extracted from the depth image. Also stated that HAR methodology can be effectively used in many smart applications like health care system to monitor the activities of a human which will improve the users life quality.

An accurate background is created by Background Estimation module. Extracted a object from the video using OS model. FE model is used to extract the feature like shape, histogram etc. Finally using the Deep learning classifier event is classified as normal and abnormal[2].

In [3] Roy et.al proposed an approach is that deep learning helps to achieve high precision, while the probabilistic model that can be interpreted makes the method explainable.A new tractable dynamic probabilistic modelling method called dynamic cutset networks is proposed and demonstrated that the accuracy estimation is substantially improved.

In[4] Rahul kavi et.al proposed a ConvNet LSTM architecture. Data is collected from multiple camera to get a better accuracy. This proposed methodology did not require any background subtraction .

In [5] A.Ullah et.al proposed a Activity Recognition approach for Industrial Surveillance. Salient features are selected using the CNN algorithm. Temporal optical flow feature is used to represent the selected features. Then multilayer LSTM is proposed to learn the long term sequences for activity recognition.

In [6] M.Z.Uddin et. al proposed an approach for activity recognition using robust translation and 5 h55541uman activity RNN model is proposed. This proposed system is explored in more real time activities.

In [7] M. Ehatisham-Ul-Haq et al proposed a feasible feature level multimodal fusion methodology for a strong human action recognition, where the data's are collected from the multiple sensors like RGB camera, wearable sensors and Depth sensor. Support vector machine and K-nearest neighbour classifiers are used to train and testthe fusion model for human activity recognition.

S. N. Gowda suggested combining two deep belief networks for recognising human behaviour in [8]. In this case, the Weber descriptor is used to extract motion characteristics. He suggested the Local binary patterns descriptor to remove the features from the images. As a consequence, this approach assists in the encoding of temporal and special information from different images or motion in the picture[11,12,13,14].

Shugang Zhang emphasisedbehaviour identification and classification methods in this paper in [9]. Jen-Kai Tsai et al. proposed a deep learning-based multiple-person action recognition method in[10], which is used in various real-time surveillance applications. This proposed project would detect and monitor the actions of several individuals in the picture, whoever they are. The best results where there were patients with both sets of data combining[15,16,17,18].

3 Proposed Work

Deep neural networks are special structure since they are composed of set of hidden elements in between the input layers and output layers that is very complex and very large structure. These hidden elements must compose of minimum two layers to be called a deep neural network. Due to their structure, deep neural networks can able to efficiently identify patterns compared to shallow networks. Without code of precises rules for each task, deep neural networks categorize the data based on input according to trained labeled data.

One kind of deep neural network with multiple layers of hidden elements are called Deep Belief Networks (DBN). This is kind of productive graphical system and it contains two phases. The first phase of DBN is pre-train phase which comprises several layers of Restricted Boltzmann networks. The second of DBN is Fine-Tune phase which is a feed forward neural network is used to adjust weights of networks. This graphical structure will learn to retrieve a deep hierarchical structure of the training data.

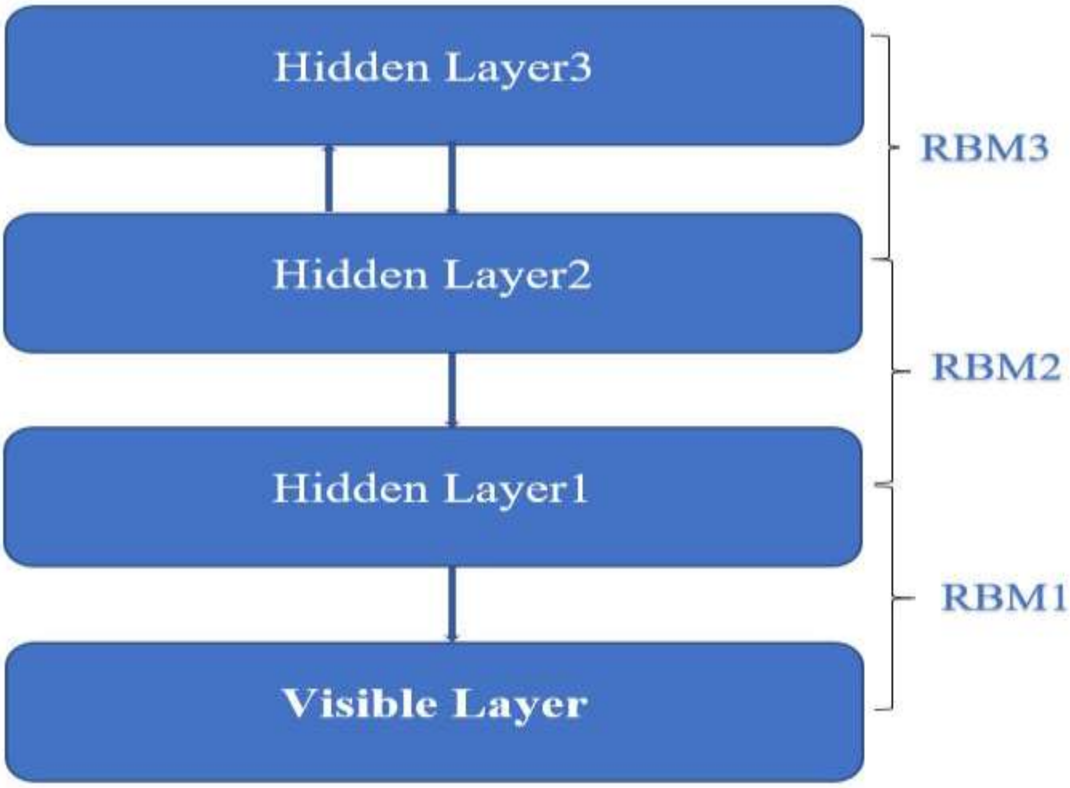

Fig 1. Structure of DBM

DBN's structure is depicted in Figure 1. One input layer, three hidden layers, and one output layer are all present. RBM is used to set up the network. RBM is useful for unsupervised learning and may also help to avoid local optimum errors. The network is initialised using a greedy layer-by-layer technique.

After the RBM1 weights have been trained, the hidden layer1 weights are represented by h1. After the RBM2 weights have been trained, the weights of the hidden layer2 are computed. After the RBM3 Weights have been trained, the hidden layer3 weights are represented by h3.

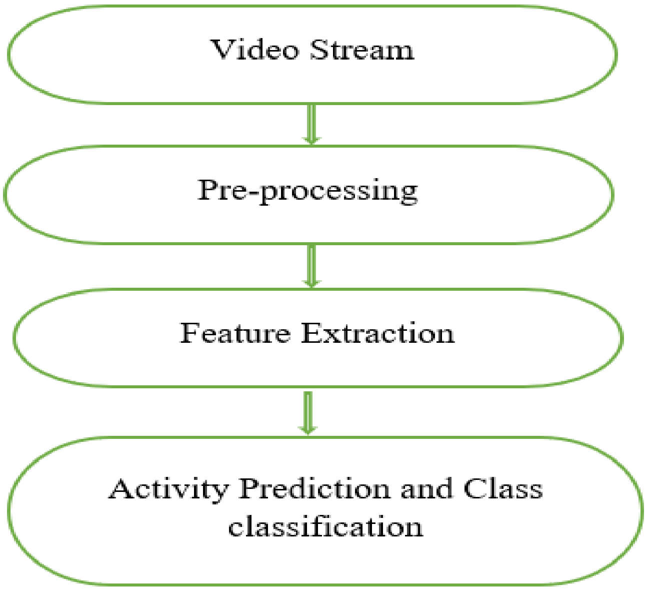

Fig 2. Steps for Human Activity Recognition

The sequence of learning steps for prediction and classification of human activity is depicted in Fig.2. Initially, the extracted video frame from video streaming is needs to be pre-processed before going to predicting activities of a human. Each RGB color imaged frame is converted into YUV model. In YUV model, Y component contains grayscale information or luminance, the U channel contains the chromatic data and the V component contains color data. D-Image or difference image is determined by finding the absolute value of the variation among successive video frames. In D-Image, motion elements are spotlighted.

Feature Extraction:

Features are points of interest in video frames. The most important characteristics of points of interest are: Saliency, Repeatability, Locality, Numerous and Efficiency. Repetitive structure less patches are very difficult to detect always. Large contrast gradients patches are easy to detect. So, our proposed approach mostly considers shape-related and Histogram-related features.

Shape-related features are taken by putting boundary for any object in x and y directions. The smallest boundary which contains maximum pixel which is enough to predict or identify content of images or objects. From the x, y values shape and size of the object is determined.

$$R^{W}_{x,y} = \frac{T^{W}_{P}}{T_{P}} \qquad (1)$$

$$R^{B}_{x,y} = \frac{T^{B}_{P}}{T_{P}} \qquad (2)$$

Where, TP represents total number of pixels, represents total number black pixels, i represents total number of white pixels. R and R represents number of repetitions of black pixels and white pixels.

Using the shape related features detect human objects in given frame and find out the same human object in how many subsequent video frames. Once find out the human object that will be pass it in to RBM layers to compute human activities for that it will select the feature maps that contains only human objects. For selected Feature maps it will determine the global mean. After that find the histogram value by finding histogram peak value in the particular time interval. Histogram value is computed by

$$H_i = \frac{H_P}{T_P} \qquad (3)$$

Where Hi is representing histogram value for intensity i. HP is histogram peak value for intensity value i. Those feature maps with global mean is zero and also with histogram value is null is selected for activity detection.

Human Activity Detection:

Deep Belief Network is used to predict human activity and classify that activities. Based on the selected feature maps machine is trained and classify the input features using multiple layers of RBM. The visible layer of RBM is represented by V vector and hidden layers are represented by vector H. The Energy function of Multiple layers of RBM is expressed as follows:

$$E(V,H) = -\sum_{i=1,j=1}^{i=x,j=y} v_i W_{i,j} h_j - \sum_{i=1}^{i=x} a_i v_i - \sum_{j=1}^{j=y} b_j h_j \qquad ($$

Bias value of visible layer is expressed as a, bias value of hidden layer is showed as b and weighted value among visible and hidden unit is expressed as W. The conditional distribution for one layer to another layer is described as follows:

$$P(v_i) = \sigma(a_i + \sum_{j=1}^{j=y} W_{i,j} h_j \qquad (5)$$

$$P(h_i) = \sigma(b_i + \sum_{j=1}^{j=y} W_{i,j} v_j \qquad (6)$$

is used to compute the output of RBM in the range of [0,1]. Where, (x) is expressed as

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad (7)$$

In the pre training phase of DBN is constructed using above sequence of steps using training set. In this multilayer architecture every layer of RBM, the input layer gets initialized. The output of one layer is goes to input for the next level layer.

Algorithm: Pre-training phase of DBN:

1. First the visible unit V is assigned as training.

2. Upgrade the visible unit and hidden unit by using equation (5) & (6) respectively.

3. Stream line the visible and hidden units by repeating step2.

4. Carryout the weight improvements and bias value of both the layers.

$$\delta(W_{i,j}) = < v_i h_j \quad _{data} - < v_i h_j >_{reconstruction} \qquad (8)$$

The next phase of DBN is fine-tune phase. The classifier is trained as 12 human activities as 12 class labels based on hand position, head position, leg position, standing and sitting and pending sequences. Using back propagation approach the human activities are categorized as any one of the 12 classes. Our proposed model attained greater than 90% of accuracy to categorize human activity recognition.

- LSSVR solves linear equations instead of a quadratic programming problem.

4 Performance

the data and responding to the steps that are necessary to put the transaction data into a format usable for processing may result from inspecting the computer to read the information from. In this section, using distinct recognition accu-racy, metrics including confusion matrix, overall accuracy, and classical accuracy graph, our suggested activity recognition approach is experimentally evaluated. The proposed method is tested by various benchmark datasets, including UCF101, UCF50, YouTube and Holly wood Actions respectively.The deep learning toolbox "Caffe" is used for the extraction of temporal optical flow features using the CNN model FlowNet2, while the "Ten-sorflow" deep learning system is used for the implementation of multilayer LSTM. Experiments were carried out using a stratified 60 percent sample for preparation for all five datasets, 20 percent for validation and 20 percent for research.Recognition of activity has been represented as a challenging task in the literature survey for UCF101 dataset. The reason for this is the video content reappear as a original activity. With a total of 12,220 videos picked from YouTube, the overall dataset has 90 groups. The proposed method has been compared with the task identification method like LSTM, Hierarchical Clustering and Long Term memory Regularization.

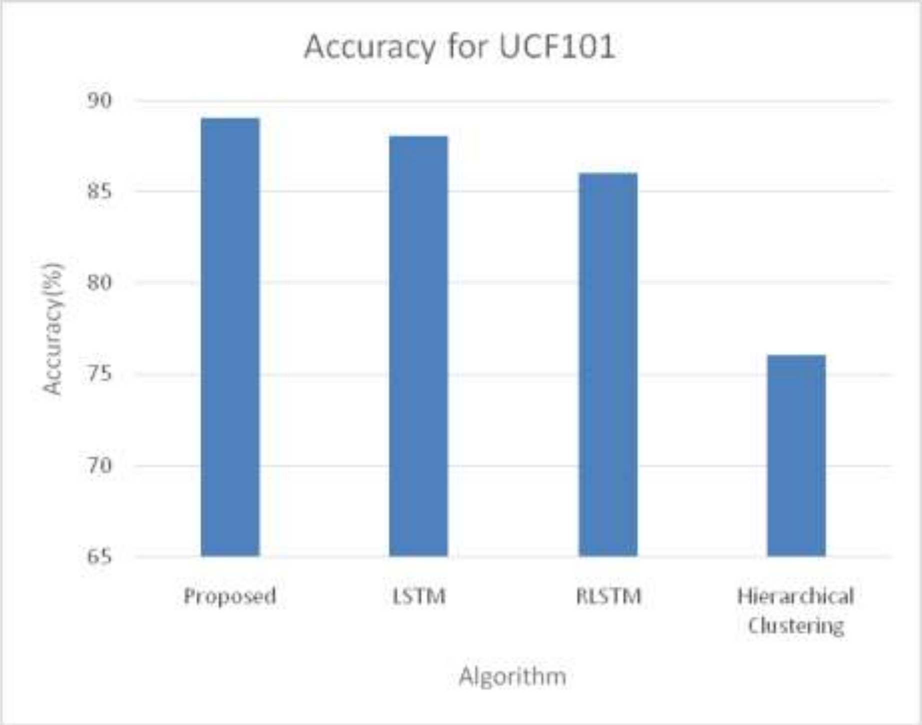

Fig.3. Accuracy for UCF101 Dataset

In order to predict the various action of human behaviour, UCF50 plays a major role among the various dataset as specified in the literature. The UCF50 as the name specifies has 50 categories of action. In few categories same action shown in different ways. The proposed method has been compared with the task identification method like LSTM, Hierarchical Clustering and Long Term memory Regularization. The accuracy obtained is shown in the Fig 4.

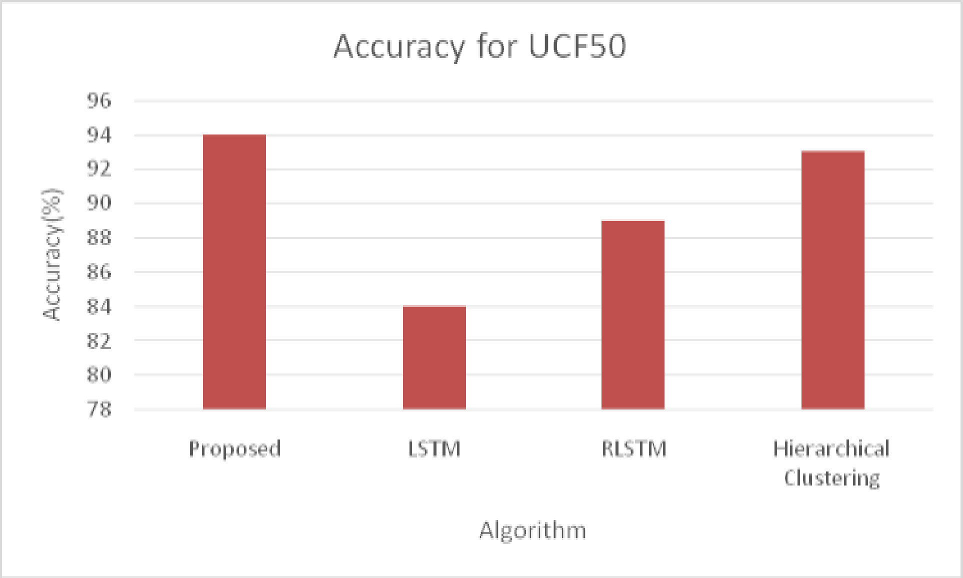

Fig4. Accuracy for UCF50 Dataset

Due to very complex dynamic and static camera images, the YouTube action dataset is difficult. There are several sports videos compiled from YouTube and other videos. The accuracy with other methods using this dataset are given in Fig 5.

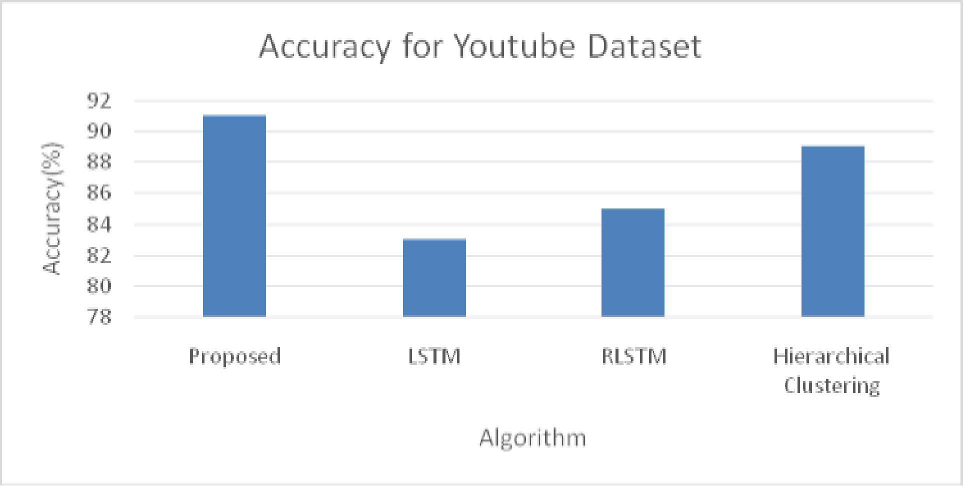

Fig.5. Accuracy for YouTube Data

Twelve different human actions has been represented in the Hollywood dataset. To identify the behaviour the literature survey inferred that the Hollywood dataset is the best one as it offers 12 human activities in a clear manner. Sixty Nine movies from Hollywood with video clips of AVI format of 810 has been stored in this dataset. The accuracy for various category of Hollywood collection dataset using the proposed approach is shown in the fig.6.

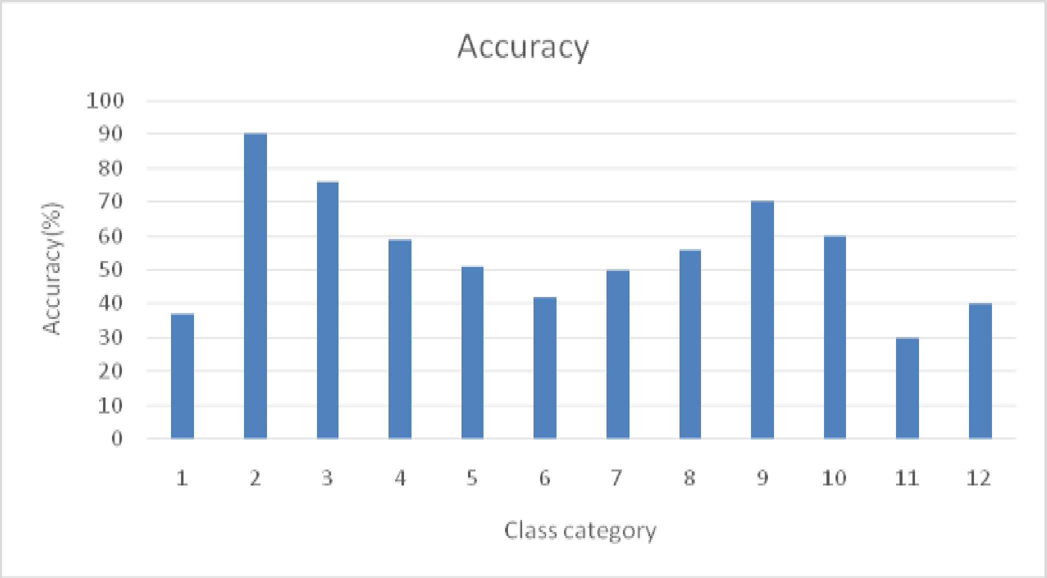

Fig.6. Accuracy for Category of Hollywood Action Data

Since several flap actions are there within the group, it is one of the most difficultdataset to identify the activity. In the dataset the Walking activity is performed near to standing it is very difficult to predict the accurate activity. Even though combination of Two dimension and three dimension convolution neural network is having greater accuracy compared with our approach, recovery of task in an industrial set up can be handled easily in our proposed approach. The proposed approach is the appropriate method for industrial systems focusing on viability and precision of the implementation. With the dataset our proposed approach is compared with the various approaches like LSTM, Hierarchical Clustering and Long Term memory Regularization. The accuracy is shown in fig.7

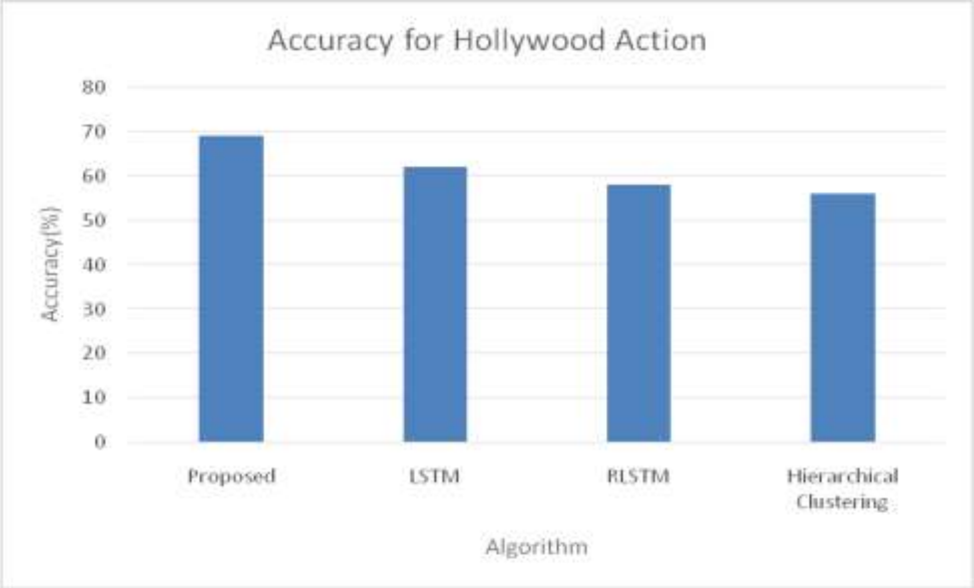

Fig.7. Accuracy for Hollywood Action Data

The four dataset statistics check that our method has enhanced performance on four datasets and that the quality of most datasets is uniformly superior across all categories.

Conclusion:

The models were trained and assessed on the public UCF-ARG dataset. The size of human patches varies in the same movie depending on the height of the moving airborne platform and the diverse viewpoints of humans, which is the most challenging component of this dataset. The suggested gadget can recognize persons without the need for a manual detection threshold to select the one with the highest true positive rate. The proposed methodology can be used in the future for a variety of application situations to enable smart surveillance, such as the detection of suspicious behavior.

References:

1. Md. Zia Uddin and Jaehyoun Kim, "A Robust Approach for Human Activity Recognition Using 3-D Body Joint Motion Features with Deep Belief Network", KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 11, NO. 2, Feb. 2017, ISSN : 1976-7277, https://doi.org/10.3837/tiis.2017.02.028.
2. Revathi, A.R., Kumar, D. An efficient system for anomaly detection using deep learning classifier. SIViP 11, 291–299 (2017). https://doi.org/10.1007/s11760-016-0935-0
3. Roy, Chiradeep, Mahesh Shanbhag, M. Nourani, Tahrima Rahman, Samia Kabir, V. Gogate, N. Ruozzi and Eric D. Ragan. "Explainable Activity Recognition in Videos." IUI Workshops (2019).
4. Kavi, Rahul, V. Kulathumani, Fnu Rohit and V. Kecojevic. "Multi-view fusion for activity recognition using deep neural networks." (2016).
5. A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik and V. H. C. de Albuquerque, "Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM," in IEEE Transactions on Industrial Electronics, vol. 66, no. 12, pp. 9692-9702, Dec. 2019, doi: 10.1109/TIE.2018.2881943.
6. M. Z. Uddin, W. Khaksar and J. Torresen, "Activity Recognition Using Deep Recurrent Neural Network on Translation and Scale-Invariant Features," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 475-479, doi: 10.1109/ICIP.2018.8451319.
7. M. Ehatisham-Ul-Haq et al., "Robust Human Activity Recognition Using Multimodal Feature-Level Fusion," in IEEE Access, vol. 7, pp. 60736-60751, 2019, doi: 10.1109/ACCESS.2019.2913393.
8. S. N. Gowda, "Human Activity Recognition Using Combinatorial Deep Belief Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, 2017, pp. 1589-1594, doi: 10.1109/CVPRW.2017.203.
9. Shugang Zhang, Zhiqiang Wei, JieNie, Lei Huang, Shuang Wang, and Zhen Li, "A Review on Human Activity Recognition Using Vision-Based Method", Hindawi Journal of Healthcare Engineering Volume 2017, Article ID 3090343, 31 pages https://doi.org/10.1155/2017/3090343
10. Tsai, Jen-Kai; Hsu, Chen-Chien; Wang, Wei-Yen; Huang, Shao-Kang. 2020. "Deep Learning-Based Real-Time Multiple-Person Action Recognition System." Sensors 20, no. 17: 4758.
11. Nagarajan, G., R. I. Minu, and A. Jayanthiladevi. "Brain computer interface for smart hardware device." International Journal of RF Technologies 10, no. 3-4 (2019): 131-139.
12. Nirmalraj, S., and G. Nagarajan. "An adaptive fusion of infrared and visible image based on learning of sparse fuzzy cognitive maps on compressive sensing." Journal of Ambient Intelligence and Humanized Computing (2019): 1-11.
13. Nirmalraj, S., and G. Nagarajan. "Biomedical image compression using fuzzy transform and deterministic binary compressive sensing matrix." Journal of Ambient Intelligence and Humanized Computing 12, no. 6 (2021): 5733-5741.
14. Nagarajan, G., Ravi, C.N., Vasanth, K., Immanuel, D.G. and Jebaseelan, S.S., 2016. Dual converter multimotor drive for hybrid permanent magnet synchronous in hybrid electric vehicle. In

Proceedings of the International Conference on Soft Computing Systems (pp. 237-249). Springer, New Delhi.

15. Minu, R., Nagarajan, G., Suresh, A. and Devi, J.A., 2016. Cognitive computational semantic for high resolution image interpretation using artificial neural network. BIOMEDICAL RESEARCH-INDIA, 27, pp.S306-S309.
16. Vasanth, K., V. Elanangai, S. Saravanan, and G. Nagarajan. "FSM-based VLSI architecture for the 3× 3 window-based DBUTMPF algorithm." In Proceedings of the International Conference on Soft Computing Systems, pp. 235-247. Springer, New Delhi, 2016.
17. Nagarajan, G. and Minu, R.I., 2016. Multimodal fuzzy ontology creation and knowledge information retrieval. In Proceedings of the International Conference on Soft Computing Systems (pp. 697-706). Springer, New Delhi.
18. Indra, Minu Rajasekaran, Nagarajan Govindan, Ravi Kumar Divakarla Naga Satya, and Sundarsingh Jebaseelan Somasundram David Thanasingh. "Fuzzy rule based ontology reasoning." Journal of Ambient Intelligence and Humanized Computing 12, no. 6 (2021): 6029-6035.