IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115205

HUMAN ACTIVITY RECOGNITION IN REAL TIMEUSING DEEP LEARNING

AZHAGUMEENATCHI.C[1], DURGA DEVI.R[2], KAREESHINI.S[3], SARANYA.B[4], SANGEETHAPRIYA.J[5]

[1]Information Technology, Saranathan College of Engineering, Trichy, India

[2]Assistant Professor, Information Technology Saranathan College of Engineering, Trichy, India

Abstract: In today's world, Human ActivityRecognition [HAR] plays a critical role in 'human- to-human' interaction. HAR displays and provides the identification of a human as well as the action done by that human, which is tough to recognize. Due to the high processing time, deep learning techniques such as CNN and LSTM cannot be used,instead we will apply transfer learning to recognize human activities. For many computer vision-based applications, such as video surveillance, criminal investigations, and sports applications, human action recognition is one of the difficult issues. Using the similarities between each pair of frames, each extracted sub-unit is further separated into frames that represent action. We will detect the action by comparing the generated HOG to the existing HOGs in the training phase, which represents all the HOGs of many actions using a dataset, utilising the Histogram of the Oriented Gradient (HOG) of the Temporal Difference Map (TDMap) of the frames.

1. INTRODUCTION

The goal of Human Action Recognition (HAR) is to comprehend a person's action and assign a label to each and every action. It has a wide range of applications, and as a result, it has gotten a lot of attention in the subject of "computer vision." Human activity recognition [HAR] is a function in Human Computer Interaction [HCI] that identifies various actions to aid computerizedoperations. HAR is also useful in a variety of security surveillance and CCTV applications, as well as sports forecasting.[3]

Human activities/actions refer to all activities done by people in order to achieve their own objectives. In real life, certain human actions are carried out for monetary benefit, while others are carried out for personal satisfaction. Since Real- Time action has high difficulties in extraction of landmarks due to the speed of the movement/action is not definite all the transfer learning algorithms cannot be used only some of them can be used.[5] This system has made sure to compare all the algorithms and find out which algorithm is bestsuited and has the best accuracy. Combining two algorithms to make an hybrid system might be useful to disregard the background recognization ofthe grids.

2. EXISTING SYSTEM

In the training phase, most extant actionrecognition systems use pertained weights of various AI architectures for every visual representation of video frames, which has a significant impact on the feature discrepancy that is identified, such as the separation between visual and temporal indications[1]. The major issues that has not been resolved are the background grids being identified as a voxel which might end upbeing a wrong prediction. As the training increases the accuracy increases so using a large dataset can be useful but data loading is difficult which is alsoa disadvantage[13].

Disadvantages: System's overall performance is poor. The system's processing speed is also slow[8]. This system is difficult to deploy in real- world circumstances.

3. PROPOSED SYSTEM

This is all about a technique that has been utilized to predict using 16 different types of action classes. Each class includes a total of 1000 photosto train with. The MideaPIPE program is a deeplearning tool for extracting human body landmarks. This project recognizes human actions such as walking, running, sleeping, fighting, and so on. The entire process entails supplying a video/image input, turning it to frames, passing it through modules that detect the activity, and eventually producing a result in the form of a label, which is then transformed into a speech output[15]. This method is vision-based, meaning it recognizes human activity from video or photos. In this scenario, the system uses a camera to collect data and recognize activities. Smartphones that recognize human activities are also a common data source.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115205

The method of analysing human motion using computer and machine vision technology is known as human activity recognition, or HAR. Sensors record human motion, which can beinterpreted as activities, gestures, or behaviours. The movement data is subsequently converted into action commands, which are then executed and analysed by computers[11]. The inputs from theraw sensors of the human activity recognition dataset are fed into HAR models, and the output isa forecast of the user's action/activities.

The step-by-step process of human activityrecognition follows the steps below:

1. At each moment of the video the system identifies the position of the human in space using image triangulation.
2. This spatial position is then matched with an empty 3D grid.
3. Then one voxel is projected onto each of the images.
4. Pixels from the image areas where the human ispresent is transformed into a voxel[6].
5. The neural network then processess the datafrom all voxels and the outputs the estimated positions of human's landmark on the original 3D grid.

MOTION DETECTION

A "background-subtraction-based" methodis utilized to start the recognition process and ensure proper detection of the human body moving in the scene. One of the most widely utilized techniques in motion detection is background subtraction. This method's major role is background modeling, which entails removing the input video's unmodified pixels, referred to as voxels and regions. Fig .I illustrates how detection is carried out.

The computation of a moving object during the time of the input video is represented by a binary image which is created using a selected threshold.

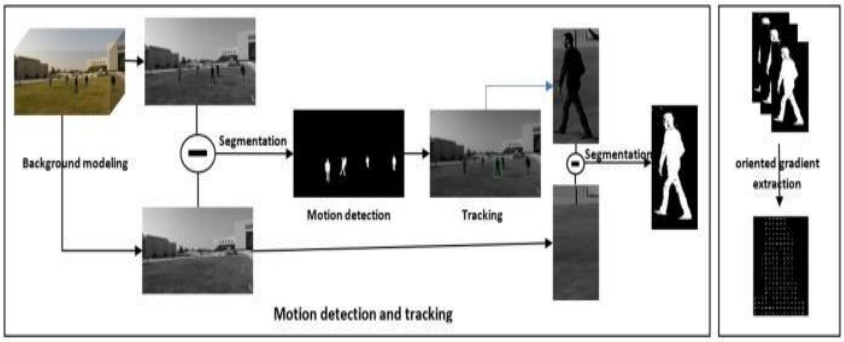

Fig.I Detection of human action

DATA PREPARATION

The training video is first pre-processed to select the landmark locations of the human bodies. The Histogram of oriented gradient (HOG) has been used to identify features, and the Temporal Difference Map (TDMap) is computed between each and every frame (each pair of successive frames). TDMap is used to create a motion history of a person's movements across all grids inthe regions.

We use HOG to extract information from theoriented gradient representation about thelandmarks of each action. Then, for each action, theHistogram of Oriented Gradient (HOG) is compared, and the relevant set of histograms is collected. For each activity, the same procedure is followed.

ACTION RECOGNITION

HOG BASED RECOGNITION - In the case of human interactions, we train with images from the UT-Interaction and INRIA XMAS (IXMAS) datasets. The histogram of the oriented gradient of the temporal difference map between each pair of photos is produced or computed after the detection of human body and extraction of landmarks for each detected individual. The histogram of each observed action is shown in Fig.II

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115205

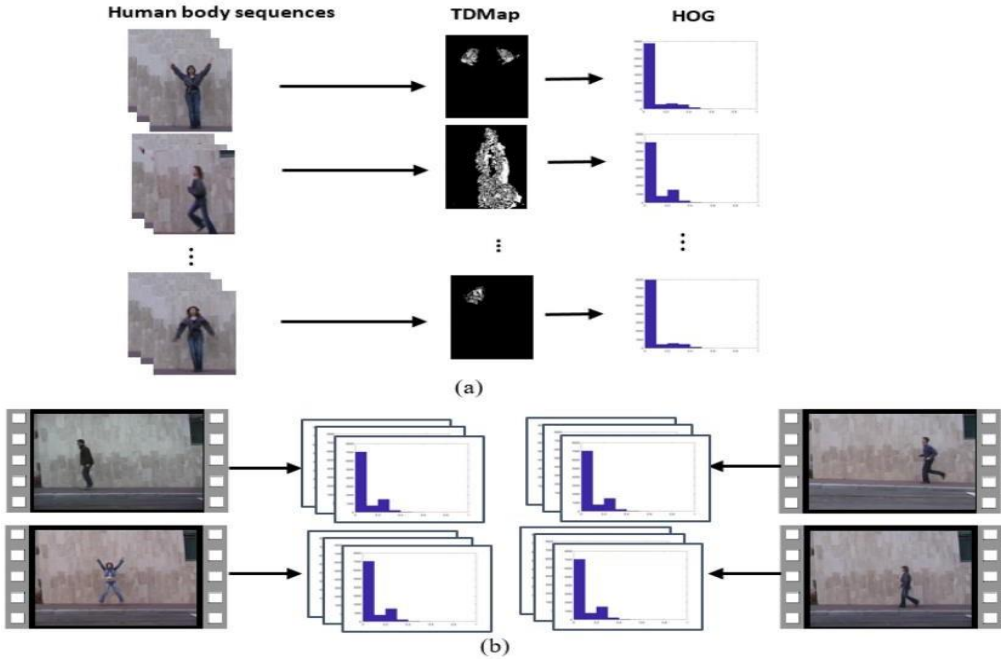

Fig.II HOGs of each action

SUMMARIZATION OF ACTION

After completing the recognition of action, the summary is finished by documentingeach action taken by a person. The procedure of differentiating the frames of each action from the sequence of actions performed by a person is known as detection. As a result, a sequence from each frame represents a person's summarized actions. Selecting a frame to represent each action can also be used to conduct the summarization. A summary is a collection of action labels from a video in which a person does something specific. The action is summarized in Fig. III

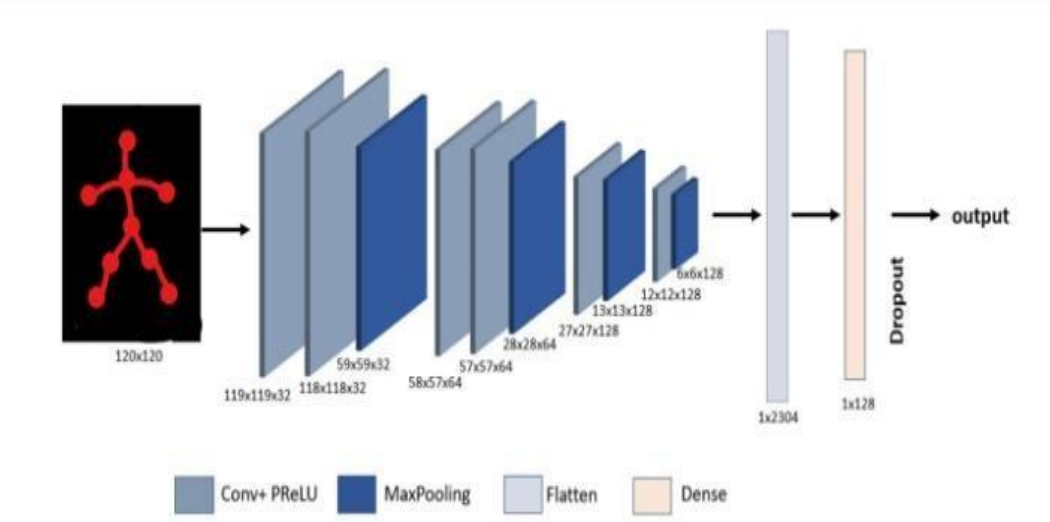

Fig.III Summarization of action

4. IMPLEMENTATION

A dataset must be loaded into the system for this process to take place. The training is split into 16 different classes, each with 1000 images. Once the input image/video has been loaded into the Home page, it is transformed into frames and preprocessed to match the training dataset in order to detect the action. When a highly similar activity is identified, it is presented as a main output in the form of a label, which is then transformed into a vocal output.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115205

5. EXPERIMENTAL RESULTS

After the continuous training of epochs, the output is obtained it is attached. Once the training is completed we can start testing the system by providing an input like an image or a video.

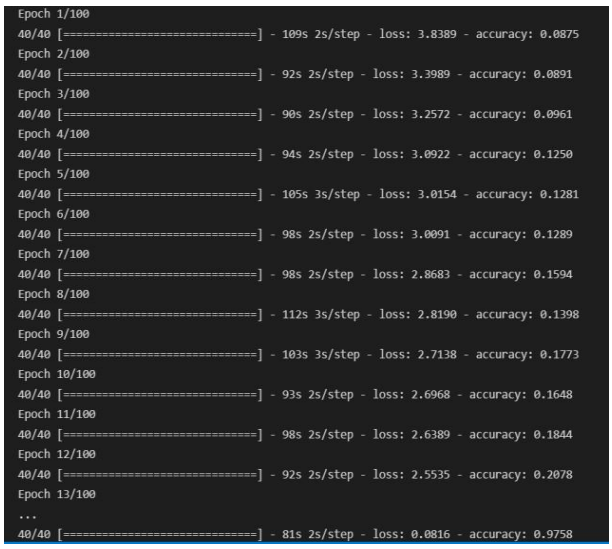

Fig.IV Training

We use Google Collaborative for training the model. We had trained upto 140 epoches with an accuracy of 96%. Fig.IV shows the training of dataset and the epochs during training session.

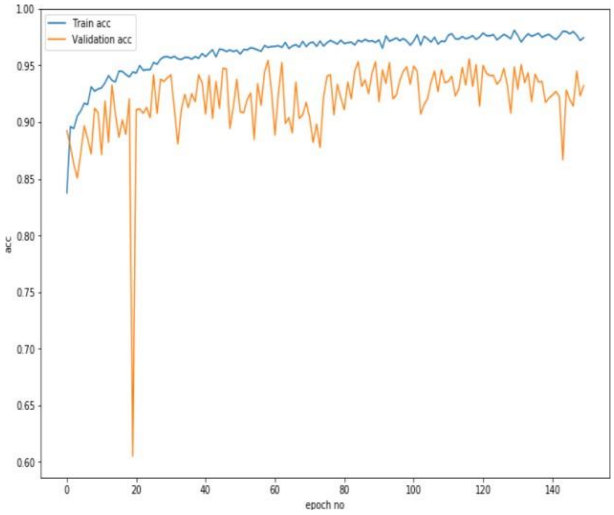

Fig.V Accuracy of Detection

Fig.V shows the accuracy of the detection.
To identify the action, the given input image/video is successively matched with the dataset. The TTS function converts the identified action from label (text) to voice (speech).

In Machine Learning (ML), the TTS [Text To Speech] function is typically used to convert text input into speech format.The primary output is the action's label, and thefinal output is an audio of the activity. The label, which is the detected output of the action done, is shown in Fig.VI.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115205

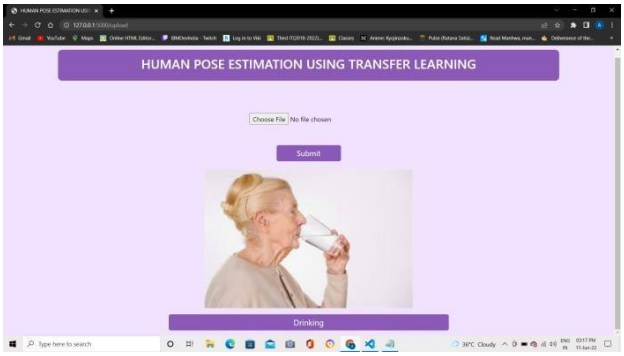

Fig.VI Label of the action detected

6. CONCLUSION

This model uses Random Forest and Grid Search together which has the highest accuracy rate while comparing with a lot of other algorithms like decision tree, k-nearest and 4 more. First and foremost, the confusion matrix for all the algorithms was found then all the algorithms were compared and the best one was found which isbeing used. Here a lot of graphs were also used to identify features that are suitable for the action detection. This model makes sure of having thehighest accuracy by plotting accuracy loss and validation graphs so that it will be more useful to know the accuracy and loss during training period. Since the training has been done with 1000 images for each class, which is 16 in total, which in turn increases the accuracy as the training period increases. The image is loaded as an Input then it is preprocessed into frames and grid to extract landmarks/features. Once extraction is done the Landmark are matched with the trained dataset to identify the highly possible action that resembles/matches the action performed. Once matched the corresponding label is given as a text output which is further converted into voice output. Human Action Recognition is one of the highly interested field which can be incorporated with any computer-vision based problems/systems. This system is to be enhanced further with more featuresand more accurate models.

REFERENCES

[1] A.D. Antar, M. Ahmed, M.A.R. Ahad, Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review, in: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition, IcIVPR, IEEE, 2019.

[2] B. Ali, et al., A volunteer supported fogcomputing environment for delay-sensitive IoTapplications, IEEE Internet Things J. (2020).

[3] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, ACM Comput. Surv. 43 (3) (2011) 1–43.

[4] M. Alazab, et al., Intelligent mobile malware detection using permission requests and api calls, Future Gener. Comput. Syst. 107 (2020) 509–521.

[5] M. Baccouche, et al., Sequential deep learning for human action recognition, in: International Workshop on Human Behavior Understanding,Springer, 2011.

[6] K.A. da Costa, et al., Internet of things: A survey on machine learning-based intrusion detection approaches, Comput. Netw. 151 (2019) 147–157.

[7] C. Dai, et al., Human behavior deep recognition architecture for smart city applications in the 5G environment, IEEE Netw. 33 (5) (2019) 206–211.

[8] C. Dai, X. Liu, J. Lai, Human action recognition using two-stream attention based LSTM networks, Appl. Soft Comput. 86 (2020) 105820.

[9] M. Elhoseny, et al., A hybrid model of internet of things and cloud computing to manage big data in health system

[10] R. Girdhar, et al., Actionvlad: Learning spatio- temporal aggregation for action classification. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[11] R. Hou, C. Chen, M. Shah, An end-to-end 3d convolutional neural network for action detection and segmentation in videos, 2017, arXiv preprint arXiv:1712.01111.

[12] Y.-L. Hsueh, W.-N. Lie, G.-Y. Guo, Human behavior recognition from multiview videos,Inform. Sci. (2020).

[13] H. Kwon, et al., First person action recognition via two-stream convnet with long-term fusion pooling, Pattern Recognit. Lett. 112 (2018) 161–167.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022
DOI: 10.17148/IJARCCE.2022.115205

[14] R. Khemchandani, S. Sharma, Robust least squares twin support vector machine for human activity recognition, Appl. Soft Comput. 47 (2016) 33–46.
[15] A. Keshavarzian, S. Sharifian, S. Seyedin, Modified deep residual network architecture deployed on serverless framework of IoT platform based on human activity recognition application, Future Gener. Comput. Syst. 101 (2019) 14–28.
[16] Y. Li, et al., Spatiotemporal interest point detector exploiting appearance and motion- variation information, J. Electron. Imaging 28 (3) (2019) 033002.

