Human Action Recognition Using ConvLSTM

Praveen S - 191EC251
Department of Electronics and Communication
National Institue of Technology, Karnataka
Surathkal, India
Email: praveensri.191ec251@nitk.edu.in

Y.S.Ch. Gowtham - 191EC158
Department of Electronics and Communication
National Institue of Technology, Karnataka
Surathkal, India
Email: yarlagaddasuryachandragowtham.191ec158@nitk.edu.in

***Abstract*— Human Activity Recognition is a real life problem which on solving unlocks many applications in real life. Our attempt is to recognise human activities by taking a video and extract RGB information using OpenCV, CNN skeletal information using media pipe and finally recognize the activity. The objective can be achieved only by using OpenCV but our incorporated the idea of extracting the skeletal information and passing it into neural network so that efficiency will be improved and the model will turn more robust. Here, the model can recognise only 4 actions because of computational power constraint with us but this can be extended. HAR can be done in many ways, after exploration we finalised to integrate 2 neural networks in which one is trained from the features extracted through CNN from RGB video, and other is trained with skeletal data that is extracted from the video using mediapipe. Atlast we connect these 2 neural networks using attention module to get the final prediction with an accuracy greater than 95%.**

Keywords— Action Recognition, Media-pipe, OpenCV, CNN.

I. Introduction (*Heading 1*)

Human Action Recognition (HAR) is basically concerned about identifying a specific action being done by a person based on sensor data. It aims to recognise activities from a series of observations based on action done by person and the environment beside. Human action recognition plays an important role in human-human interaction & interpersonal relations. Human action recognition is very vital and difficult research problem in the field of computer vision . Recognising activity of human beings has become very popular these days.

The basic idea is to understand human behaviour and assign a label to it concluding what is that activity. The actions which our model can recognise depends on what actions we wish to train on. So, we have to have an idea about what actions our model need to predict and we have to train using those datasets only. HAR can be done by using sensors or accelerometer data, here we have extracted frames from video, extracted RGB information using one neural network and passed it into other neural network and finally connected these 2 networks to get a final prediction.

Motivation describes the applications where we got motivated to do this work. Related work is the survey of the works in the past related to Human Activity Recognition. And Methodology is the section where the entire description of the procedure we followed to do the project. Discussion and Conclusions include the data-set description we followed and Results we achieved. Limitations section gives the idea on what limitations our idea works. Conclusions of the work done are included in the conclusions section. Then there at the end References are placed which helped us to refer that relates the work.

II. MOTIVATION

- It provides information about identity of person, their personality & their psychological state.
- It can be used to detect person's emotion (whether he is interested in listening to class or whether he is sleepy).
- It can be used to detect sleepy drivers & alert them, thereby saving lives.
- It can be used to detect if any restricted activities are done in some places (like playing in restricted areas).

III. RELATED WORK

A. Abbreviations and Acronyms

B. Equations

IV. METHODOLOGY

A. Background

LSTM : Long Short Term Memory is a sequential network. It is one of the forms of Recurrent Neural Network (RNN). LSTM network consists of three parts, and each component performs different functions. Forget Gate, Input Gate, and Forget Gate are the three components present in the LSTM network.

LSTM IMAGE

Forget Gate :The first step of the LSTM network is to decide whether we should keep the information from the last hidden state. Forget Gate has a sigmoid function takes current state and previously hidden state as inputs and generates output(0 to 1) which will undergo point ny point multiplication with cell state.

EQUATION

Input Gate: The input gate performs the following operations to update the cell status. The current state and previously hidden state are passed into a sigmoid function and a tanh function simultaneously. The sigmoid will transform the values to be in between 0 and 1 on basis of importance . The tanh function will create a vector with all the possible values between -1 and 1. The output values generated form the activation functions will undergo point-by-point multiplication and update the cell state.

EQUATRION

Output Gate: The output gate determines the value of the next hidden state. First, the current state and previous hidden state are passed into a sigmoid function. Then the new cell state generated from the cell state is passed through the tanh function. Both these outputs are multiplied point-by-point. Based upon the final value, the network decides which information the hidden state should carry. This hidden state is used for prediction.

EQUATION

ConvLSTM:

The work is to predict human activities, for that we need to recognise the sequence of postures. To do this the model should be trained with dataset, by extracting both RGB features and Skeleton data from the frames of the video clips. The Fig. 1 is the architecture of the work human activity recognition. The detailed process is as follows:

A. Extracting Data

Used the UCF-50 dataset which has the 50 classes of activities and each class has an average of 130 videos. Considered 4 classes for the implementation in view of the resource limit, where the same architecture can be extended to the other classes as well. Each Video clip is in the format of avi files and duration is about an average of 10 sec. For extracting RGB features from the videos, CV2 is used to extract each frame from the video and processed to the size of 64x64. Now this frame data is in the format of numpy array.

V. DISCUSSIONS AND CONCLUSIONS

A. Dataset description

UCF-50 dataset has been used which contains video samples of many activities done by humans. It has videos for 50 different activities each with 140 videos approximately. We selected only Basketball, Biking, Golfswing, Kayaking. Each class has around 140 video samples which captures the activity done by different persons from different angles so that all possibilities will be covered and the model wouldn’t be overfitted. So, model has been trained using 560 video samples (approximately) each of 10 seconds.

REFERENCES