EasyChair Preprint

№ 7377

Real Time Human Activity Recognition with Video Classification

S Janhavi and Chandra Sekhar Malepati

January 27, 2022

REAL TIME HUMAN ACTIVITY RECOGNITION WITH VIDEO CLASSIFICATION

S Janhavi
Department of Computer Science
Presidency University
Banglore, India
Email id:
janhavisathya.1998@gmail.com

Malepati Chandra Sekhar
Department of Computer Science
Presidency University
Banglore, India
Email id:
mchandrasekhar@presidencyuniversity.in

***Abstract*— Recognition of the human activity is a very broad area of study that will aim to identify the specific movement or the action of the person. Human activity recognition is a very vast area of research and exploration This is a type of time series classification problem where data from a series of time steps is needed to properly classify the current activity. Activities are actions such as walking, eating, sleeping, reading newspapers, talking, jumping, standing, drinking and sitting. Recognizing human activity, or HAR, is a very difficult task of classifying the data points. To put this in simple words, the action of classification or prediction the activity or the action will be performed by someone is known as the activity recognition. The issue which arises here is, in the action Recognition, you will actually need the series of the knowledge-based points for the prediction of the action which is being performed accurately. So, the action-based Recognition would be in a form of the statistic classification with the draw back where it is likely that the data from the series of the timesteps to properly classify the action which is being performed properly. It also involves the prediction of the movement of an individual data and will involve the deep domain expertise and different methods from the signal method to engineer choices properly from the information therefore on the pursuit a machine learning model. Recently, the deep learning models will appreciate the convolutional neural networks and the continual neural networks which have shown the capable and even showed the successful progressive results by automatic learning options from the raw sensing element data.**

Keywords— CNN, Real Time, Video, Human action, ML, AI.

I. INTRODUCTION

Artificial intelligence (AI) is the knowledge interpretation done by the machine. AI inventions explains each field of interest as a gaining information of "intelligent agents": any particular system which will identify the surrounding environment and performs related actions which will increase the chances of achieving its aim. Some use "artificial intelligence" to describe the machines that will resemble the "mental" activities of human which is related to the mind, like "dancing" and "learning".

AI developed applications consist of very advance web search engines such as google, and many recommended applications (used by Amazon, Netflix and YouTube), and speech-based recognition (such as Google Assistant and Alexa), and automotive based applications (e.g.., Tesla), to give the automated decisions. Competing at higher levels of the game programs (such as the chess). As machines are emerging rapidly, activities which need "intelligence" are deleted from AI, something known as the result of AI. For example., visual based recognition is deleted from being a part of AI, and as emerged to become a new technology.

Video Editing is the function of producing a video-related label in view of its frames. A good video quality separator that not only provides accurate frame labels, but also defines the entire video when considering the features and annotations of the various frames in the video. Videos can be understood as a series of individual images; therefore, many in-depth reading professionals can quickly manage video segmentation as making image editing a total of N times, where N is the total number of frames in the video.

The human ability to recognize the activities of another person can be an important topic of study in the scientific fields of computer vision and machine learning. once one tries to recognize human activities, it is necessary to process the kinetic states of a person, in order for the pc to actually recognize this activity. advanced activities are often reduced to other less complicated activities, which are generally easier to recognize. Usually, excessive detection of objects in a scene can make it easier to perceive human activities as it can offer useful data about that event. separate the components of the image which are invariant over time (background) from moving or dynamic objects (foreground). Human tracking, wherever the system detects human movement over time, human activity associated with the detection of nursing objects when the system is ready to locate an action in an image.

II. EXISTING SYSTEM

Most of the existing approaches represent human activities as a set of visual features extracted from video sequences or still images and recognize the underlying activity label using several classification models in controlled environments. The dataset considered are generic. However, these limitations constitute an unrealistic scenario that does not cover real-world situations and does not address the specifications for an ideal human activity dataset

Observe the human activities which are performed like the studying and eating. We can use the wearable device sensors, which will collect the testing and the training data. After collecting the data, it will send this data to the record output. Further in the exit feature, the pre-processing of the data takes place from the external sensors and the functions which are from the processed data such as upscale, lie, location etc. According to known work the model works. HMM analyzes and evaluates different provinces based on different human activities. Functional analysis can be done with temporary patterns. The performance of the improved model is tested in the test phase. An important strategy Sentence separation is widely used in analyzing human activities. HMM therefore issues self-awareness and self-discovery.

This shall represent the working of HMM model, and the Artificial Neural Network model and the Dictionary Learning Algorithm for the human activity recognition. The present

methods which are mentioned here are very useful and very effective in observing the human activity

A. Related Works

Neural Networks are used to monitor a person's daily activities. A major challenge in designing a HAR-enabled neural network is finding the number of hidden layers. A 2-layer server network which could be accessed in the HAAR mode. And also, the Concurrent Neural Network and the Recurrent Neural Networks, are the most common in-depth learning strategies can be used to address the HAR problem. In the Big Data era, where various devices can connect independently through network and cloud services, the smartphone has many sensors that can detect data about everything around it. This makes identing-process (AR) applications and behaviors aware of the context. We use an algorithm to predict human activity based on collected sensory data. Also, Principal Component Analysis (PCA) is used in 561 databases. PCA reduced data size from 561 to 50, reducing data weight. Therefore, many important features are identified in 561 aspects.

The Dictionary Learning Algorithm plays a key role in signal processing and machine learning. This algorithm will also work well for both the offline and online categories. The offline categories include events such as collecting the data from the sensors and then processing it. An online forum updates data whenever a new signal arrives. In order to extract time series data, two methods are used: structural and mathematical. Structurally, it defines related to data and statistics, the Fourier and Wavelet modifications reflect features with plurality. Represents the stages of flexibility and data processing compared to MOD, K-SVD. In real time, job recognition does not work.

It is very thankful to the pre-processing of the efficiency and the flexibility of this models which are improved. A very productive thesaurus learning algorithm have been proposed for overcoming the problems and then provide the effective solution for this. Again, by going through the same models which have been mentioned below, the human live activities can be predicted easily.

- Vector Support Equipment (SVM)
- Condition-Based Learning (IBL)
- Bayesian ways
- A combination of class dividers
- Decision trees

The wearable device sensors will provide the very inexpensive, visible function, and a straightforward answer for the human activity recognition. It may be used in several cases such as in the surveillance systems and the identification systems.

III. PROPOSED SYSTEM

A. Objective

Building an AI based model which can work on the 3D CNN architecture and can recognize the human activities and also yield better accuracy than any other model in a short period of time and can be implemented on the real time video data also

B. Overview

The dataset which is used in the current project is the UCF100 dataset which contains a pair of human action recognition data of 100 action categories, which includes the real time videos extracted from the you tube. The present dataset used is the YouTube Human Action (UCF100) data extension with the 100 stages of action.

All the human activity recognition data sets are not the original real one and they will be performed by the actors. In the present data set we are using; the major focus is giving a clear view of a systemic view of action data recognition which will include the real time videos that are taken from the YouTube site. The collection of this dataset has been very difficult and challenging for many of the reasons as to the wide variation of camera movement, location and location, object scale, vision, integrated background, lighting conditions, etc. In the available 100 categories of action, the video clips have been grouped into the 25 categories, where each of them contains content of each higher than the 4 action clips. Videos which are in the same circle could share the similar features, like they belong to the same or similar person, and they may belong to the same background, the they may also belong to the similar views, and more.

After the preparation of data and feature extraction now, drop the data into neural network. The training procedure isas follows.Now that the need for video editing models is identified to solve problems of the Personal Identification, let us now discuss some of the basic and the logical way of editing the videos. The model will learn to differentiate between the two similar actions in the natural context.

.

C. Tools and Methodologies used

1) Programming language: Python is compatible in all the environent especially for the AI and ML Algorithms

Problem here is that the model does not always be fully confident of the prediction of each video frame, so these predictions will automatically change very quickly and very smoothly because model do not identify the flow of entire video.

One of the simple solutions to this kind of problem like instead of dividing and showing the results of a single framework, limiting the frames eliminates the dose. After the determination of the value of n, use something which is very simple such as a medium moving or rolling in middle to get this result.

2) Single Frame CNN: It is already found out that the most easy and basic implementation of the video segmentation always is to use a photo segmentation network. Now, we will use the image separation model throughout the video frame and measure all the chances of getting the final vector. This method defines well, and we are getting a chance to use it here. One of the many challenges for training video designers is finding a way to feed videos on the network. Since video is a fixed sequence of frames, we can simply remove frames and insert them into a 3D tensor. But the number of frames can vary from video to video which may prevent us from packing them in batches (unless we use pads). Alternatively, we can

save video frames for a limited time until the maximum number of frames is reached. In this example we will do the following:

3) Late Fusion: The Late Fusion model will the help of two different networks of the same frame (as described above, to the end of the presented convolutional layer and the variables are divided into the different frames and will connect the streams to the first fully. Integrated layer, then, none of the single cell tower can be able to detect any of the movement, while the fully integrated layer will be able to calculate the characters of the global movement by comparing the output to both towers.
The Fusion layer is also used for the integration of output of the different enormous networks operating in remote frames. It is usually used in the form of a large compound, in the middle, or in a flat manner.

4) Early Fusion: Early Fusion Extension integrates information into the timeline window quickly pixel density. The following can be done by adjusting the filters on the first initial layer of the convolution in the single base framework model in the enlarging it to the 11 × 3 × 11 × t pixels, where t will be a temporary measure. Pre-connected and the directed connection to the pixel data which will allow the used network to locate the location accurately direction of movement and speed.

5) Using CNN and LSTMs: We use the convolutional architecture of VGG16 proposed by Zisserman et al because it has obtained excellent results in ImageNet Large-Scale.Visual Recognition Challenge (ILSVRC- 2014) Separation activities and space.We have seen the emergence of the top 10 horizontal combinations to find the vector size 7 × 7 × 512 = 25088 to provide neural network LSTM.

6) Using Pose Detection and LSTM: Pose estimation is a computer diagnostic method that predicts and traces the location of a person or object. This is done by looking at the combination of the shape of the person / object. Pose measurement is a method of computer vision to track the movements of a person or object. This is usually done by finding important points for the given items. Based on these key points we can compare movements and different shapes and draw details. The pipeline contains both the Pose Detection model and the LSTM model..
• Our model accepts video inputs, duplicates Frame and uses Pose Detection Model to get key points across the framework.
• The key point results are then linked to a size 32 website, which works with a sliding window model.
• The contents of the bath are finally sent to our expert LSTM model for action identification.
• The actions of our target pipe are described in the video and are shown as a result.

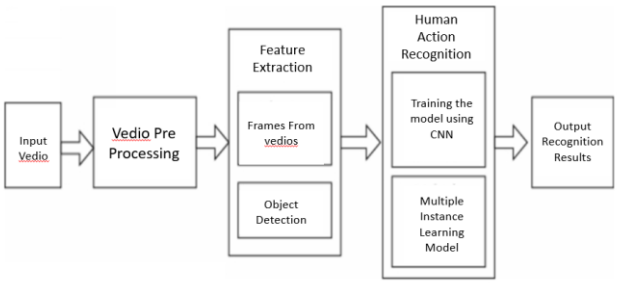
7) Using Optical Flow and CNNs: In this method, two identical communication methods are taken into the account. The high system is also known as the Spatial Stream. It will take one of the frames from the video and will use a lot of the CNN clues to it, based on the location information it will make a prediction. The provided stream below is called Temporal stream which captures all the visual flow of the adjacent structure after merging using the pre-assemble method and uses movement information to make predictions. Finally, a rating on both predicted opportunities was made to determine the latest.

8) Using SlowFast Networks: One of the streams works with video with minimal adjustment compared to others. All temporary and local operations are performed on a single network..High-quality live streaming, called slow-moving branch, operates with low-quality video frame and there are many channels in all layers to process the details of each frame. On the other hand, the streaming below, which is also meant as fast branch, have lower number of channels and gets operated with the high-quality temporary basic version of a same piece of video.

9*) Using 3D CNN's/Slow Fusions:* The Slow Fusion model is a balanced mix between the two approaches that gradually integrates temporary information across the network so that the top layers can access continuous global information both local and temporary. This is done by expanding the connectivity of all convolutional layers over time and making temporary changes over local changes to calculate performance. In this present model which we will be using, is the first initial layer of the convolution which is then extended to apply for each of the template filter template where T = 4 to a 10-frame input clip using a valid step 2 stride and to produce 4 responses at a particular period of time. The second one and the third layers which are above the repeat in this process with the middle level filters T = 2 and step 2. Hence, the third convo layer will be able to access the information in all the 10 input frames

D. *Working of the proposed system*

The website contains videos that are categorized into different action categories, such as cricket, piercing, cycling, etc. This database is often used to create action viewers, which is a video sharing app. Video contains ordered frames. Keras is a very effective and easy-to-implement and use the python library for the open-source development and the testing of in-depth reading models. It integrates Theano and TensorFlow mathematical libraries and allows you to define and train neural network models with a few lines of code.



The steps which are involved in the building of model is as follows:
1. Downloading the dataset and Extracting the Dataset: Download the required dataset which is used to build the

model such as the UCF dataset which contains the dataset of actions

2.* Visualization of the Data along with its Labels*: Pick some of the random clips of videos for each of the database class and show them, this gives a clear picture of how the database is visible.

3. Reading the dataset and Pre-process the Dataset: We shall be using the component format for training on the videos site, need to preview the website first.

Extraction, Resizing and Normalizing the Frames

Now creating a function which will eliminate the number of frames for each of the videos during the performance of the other pre-processing tasks such as resizing images and making them more standard.

4. Splitting the Data set into the Train and the Test Set: Now have to consider two identical blank members, one which contains all the images. The second one which contains all of the class labels in one particular format along with the hot code. Let us now split the data for building a training, as well as the test set. Then need to rush the data before splitting.

5 Construction of the Model: Use the plotting model function to check the architecture of the final end model. This will really help when building a very complex network, and to make sure building the network in the right way.

6. Compiling and Training the Model: Let us start the training of model. Prior to that integrating the model is very important

Rate your professional model in feature testing and label sets. Now you have to save your model for future use ***2.***

7. Plotting the Model's Loss and Accuracy Curves: Let us visualize the losses with the accuracy of the curves.

8. Making the Predictions with the Model*:* Now it is time to test the working using some other videos.

pictures, and text.

• Work in tandem with all other MLOps features.

E. System Implementation:

System Requirements

Requirement	Minimum	Recommended
RAM	4 GB of free RAM	8 GB of total system
CPU	Any modern CPU	Multi-core CPU
Monitor correction	1024x768	1920 × 1080

Disk space: :

Minimum :2.5 GB and extra 1 GB of storage

Recommended: SSD drive with the least 5 GB of space

Operating system:

Minimum: The following 64-bit versions have been officially released:

Windows 8 or later

macOS 10.14 or later

Any Linux distribution which will support the Gnome, the KDE, or the Unity DE. PyCharm is not only available in the other Linux distributions, such as the RHEL6 or the CentOS6, which will not include the GLIBC 2.14 or later.

The previously released versions which will not be supported.

Recommended: Latest 64-bit version of Windows, macOS, or Linux

JavaScript should be enabled to implement the PyCharm as JetBrains Runtime will be integrated with the IDE (based on JRE 11).

Python 2: install the updated version of 2.7

Python 3: from the version of 3.6 to the version of 3.11

Subversion

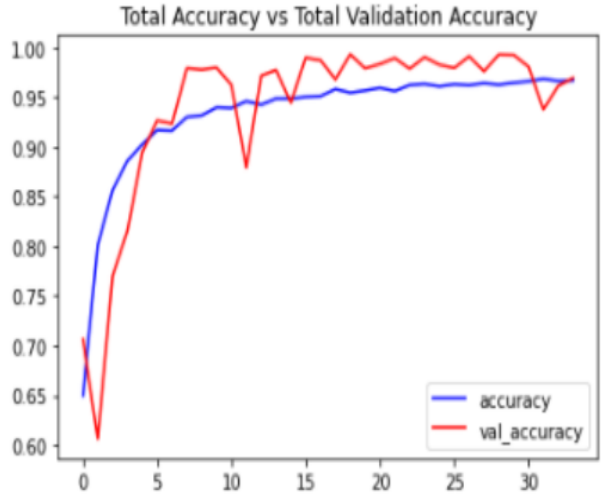
Subversion is also a version control system that keeps track of individual changes whiledeveloping the source code. Some of the example scripts still depend on this package.

IV. EXPERIMENT RESULTS AND DISCUSSIONS

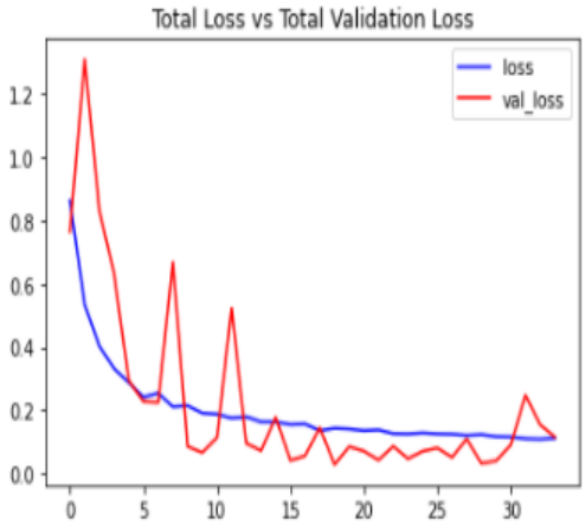
A. Training the model:

```
Epoch 1/50
5120/5120 [==============================] - 27s 4ms/step - loss: 1.0299 - accuracy: 0.5714 - val_loss: 0.7648 - val_accuracy: 0.7066
Epoch 2/50
5120/5120 [==============================] - 19s 4ms/step - loss: 0.5767 - accuracy: 0.7827 - val_loss: 1.3107 - val_accuracy: 0.6062
Epoch 3/50
5120/5120 [==============================] - 19s 4ms/step - loss: 0.4272 - accuracy: 0.8468 - val_loss: 0.8311 - val_accuracy: 0.7703
Epoch 4/50
5120/5120 [==============================] - 18s 4ms/step - loss: 0.3333 - accuracy: 0.8820 - val_loss: 0.6321 - val_accuracy: 0.8156
Epoch 5/50
5120/5120 [==============================] - 19s 4ms/step - loss: 0.2993 - accuracy: 0.8973 - val_loss: 0.2933 - val_accuracy: 0.8945
Epoch 6/50
5120/5120 [==============================] - 19s 4ms/step - loss: 0.2470 - accuracy: 0.9129 - val_loss: 0.2271 - val_accuracy: 0.9266
Epoch 7/50
5120/5120 [==============================] - 19s 4ms/step - loss: 0.2652 - accuracy: 0.9140 - val_loss: 0.2241 - val_accuracy: 0.9232
Epoch 8/50
5120/5120 [==============================] - 19s 4ms/step - loss: 0.2108 - accuracy: 0.9319 - val_loss: 0.6683 - val_accuracy: 0.9791
```

B. Accuracy graph of CNN Model



C. Loss Graph of CNN Model



D. Single Frame predictions

E. Multiple Frame Predictions

V. CONCLUSION

Recognition of human hobby could be an immense region of take a glance at those goals to understand the actual motion or movement of someone. this is often a form of your time assortment sort problem whereby statistics from a sequence of time steps is needed to nicely classify the up-to-date movement. Recognizing human hobby, or HAR, is likewise a troublesome venture of classifying statistics factors. To positioned for lightly, the venture of classification or prediction of the movement is accomplished via means of means of a person is mentioned as hobby reputation. We may also additionally have a problem tight here: but is that this fully absolutely special from a standard Classification venture. The problem right here is, in act Recognition, you merely need a sequence of experience factors to expect the movement being accomplished correctly. So, act Recognition may well be a form of data point sort shy away whereby you would like statistics from a sequence of timesteps to nicely classify the movement being accomplished. once making an attempt to know human sports, one ought to decide the kinetic states of a person, so as that the laptop will with success understand this hobby.

So, for the future work we will run the model using more live dataset and get the result. Also, we will explore different activation functions and optimizer and different new advanced approaches to improve the accuracy of the models. Future work may also include the implementation of this model for useful application like monitoring the activities, predicting danger, etc.

Future works

The model can be used to detect the anomalies and monitoring the human activities and can also be used in the field of advanced robotics to train the robots for performing the human actions

VI. REFERENCES

[1] N. Käse, M. Babaee and G. Rigoll, "Multi-view human activity recognition using motion frequency," 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3963-3967, doi: 10.1109/ICIP.2017.8297026.

[2] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan and M. Zaharadeen, "Automated daily human activity recognition for video surveillance using neural network," 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), 2017, pp. 1-5, doi: 10.1109/ICSIMA.2017.8312024.

[3] M. M. Hossain Shuvo, N. Ahmed, K. Nouduri and K. Palaniappan, "A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network," 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2020, pp. 1-5, doi: 10.1109/AIPR50011.2020.9425332.

[4] S. M. Kwon et al., "Demo: Hands-Free Human Activity Recognition Using Millimeter-Wave Sensors," 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), 2019, pp. 1-2, doi: 10.1109/DySPAN.2019.8935665.

[5] A. Bagate and M. Shah, "Human Activity Recognition using RGB-D Sensors," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 902-905, doi: 10.1109/ICCS45141.2019.9065460.

[6] Pubali De, Amitava Chatterjee, and Anjan Rakshit (2018), 'Recognition of Human Behavior for Assisted Living Using Dictionary Learning Approach', IEEE Sensors Journal, pp. 2434-2441.

[7] Uriel Martinez-Hernandez, Imran Mahmood, and Abbas A. Dehghani Sanij (2018),'Simultaneous Bayesian Recognition of Locomotion and Gait Phases With Wearable Sensors', IEEE Sensors Journal, pp. 1282- 1290

[8] Akram Bayat, Marc Pomplun, Duc A. Tran (2014), 'A

Study on Human Activity Recognition Using Accelerometer Data from Smart phones', Elsevier International Conference on Mobile Systems and Pervasive Computing, pp. 450-457.
[9] Bo Tang, Jin Xu, Haibo Heand Hong Man (2017),'ADL Active Dictionary Learning for Sparse Representation', IEEE Sensors journal, pp. 2723-2729.
[10] Christian Debes, Andreas Merentitis, Sergey Sukhanov, Maria Niessen, Nicolaos Frangiadakis, and Alexander Bauer (2016), 'Monitoring Activities of Daily Living in Smart Homes', IEEE Signal processing, pp.84-94.
[11] Meisam razaviyayn, hung-wei tseng, zhi-quan luo (2014), 'Dictionary learning for sparse representation complexity and algorithms', IEEE international conference, pp.5247-5251.
[12] Chao Wang, Siwen Chen, Yanwei Yang, Feng Hu, Fugang Liu, and Jie Wu (2018), 'Literature Review on Wireless Sensing-Wi-Fi Signal-Based Recognition of Human Activities' International conference on control systems, pp.203-222.
[13] Yan Wang, Shuang Cang, Hongnian Yu1, 'A noncontactsensor surveillance system towards assisting independent living for older people', Research gate, December 2017.
[14] Kai-Chun Liu, Chien-Yi Yen, Li-Han Chang, Chia-Yeh Hsieh and Chia-Tai Chan (2017), 'Wearable SensorBased Activity Recognition for Housekeeping Task', IEEE sensors journal, pp.67-70
[15] Soumalya Sen,Moloy Dhar, Susrut Banerjee. Implementation of Human Action Recognition using Image Parsing Techniques Yu Kong, Yun Fu.VOL. 13, NO. 9, SEPTEMBER 2018-IEEE. Human Action Recognition and Prediction: A Survey Sci. Technol., vol. 26, no. 2, pp. 239–246, Mar. 2011.