USING CNN AND TRANSFER LEARNING TO RECOGNIZE HUMAN ACTIVITY BASED ON VISION

Sk. Yasmin Sulthana[1], S.Ruchika [2], M. Varalakshmi [3], D.Khadar Vali [4], B.Uday Sankar [5]
[1] *Asst. Professor, Krishna Chaitanya Institute of Technology & Sciences , Markapur, A.P, India*
[2,3,4,5] *Scholar, Krishna Chaitanya Institute of Technology & Sciences , Markapur, India*

ABSTRACT:

With the advent of the Internet of Things(IoT), there have been significant advancements in the area of human activity recognition (HAR)in recent years. HAR is applicable to wider application such as elderly care, anomalous behaviour detection and surveillance system. Several machine learning algorithms have been employed to predict the activities performed by the human in an environment. However, traditional machine learning approaches have been outperformed by feature engineering methods which can select an optimal set of features. On the contrary, it is known that deep learning models such as Convolutional Neural Networks (CNN) can extract feature and reduce the computational cost automatically. In this paper, we use CNN model to predict human activities from Image Dataset model. Specifically, we employ transfer learning to get deep image features and trained machine learning classifiers. Our experimental results showed the accuracy of 96.95%using VGG-16. Our experimental results also confirmed the high performance of VGG-16 as compared to rest of the applied CNN models.

Keywords: CNN, Transfer Learning,VGG16 , HAR

[1] INTRODUCTION

Human activity recognition (HAR) is an active research area because of its applications in elderly care, automated homes and surveillance system. Several studies has-been done on human activity recognition in the past. Some of the existing work are either wearable based or non-wearable based. Wearable based HAR system make use of wearable sensors that are attached on the human body. Wearable based HAR system are intrusive in nature. Non-wearable based HAR system do not require any sensors to attach on the human or to carry any device for activity recognition. Non-wearable based approach can be further categorized into sensor based and vision-based HAR systems. Sensor based technology use RF signals from sensors, such as RFID, PIR sensors and

Wi- Fi signals to detect human activities. Vision based technology use videos, image frames from depth cameras or IR cameras to classify human activities. Sensor based HAR system are non-intrusive in nature but may not provide high accuracy. Therefore, vision-based human activity recognition system has gained significant interest in the present time. Recognizing human activities fromthestreamingvideoischallenging.Video-basedhumanactivityrecognitioncanbecategorised as marker-based and vision-based according to motion features . Marker-based method make use of optic wearable marker based motion capture (MoCap) framework. It can accurately capture complex human motions but this approach has some disadvantages. It require the optical sensors to be attached on the human and also demand the need of multiple camera settings. Whereas, the vision based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human.

Most of the vision-based HAR systems proposed in the literature used traditional machine learning algorithms for activity recognition. However, traditional machine learningmethodshavebeenoutperformedbydeeplearningmethodsinrecenttime.Themostcommon type of deep learning method is Convolutional Neural Network (CNN). CNN are largely applied in areas related to computer vision. It consists series of convolution layers through which images are passed for processing. In this paper, we use CNN to recognize human activities from Wiezmann Dataset. We first extracted the frames for each activities from the videos. Specifically, we use transfer learning to get deep image features and trained machine learning classifiers. We applied 3 different CNN models to classify activities and compared our results with the existing works on the same dataset.

In summary, the main contributions of our work are as follows:

1) We applied three different CNN models to classify human recognition activities and we showed the accuracy of 96.95% using VGG-16.
2) We used transfer learning to leverage the knowledge gained from large-scale dataset such as Image Net to the human activity recognition dataset.

Hidden Markov Model (HMMs) methods have been largely used as the recognition techniques in the past because of its capability of temporal pattern decoding. However, researchers are more interested in using deep learning techniques because of its ability to automatically extract the features and learn deep pattern structures. Deep learning method shave clearly ruled out traditional classification methods in the domain of computer vision. Deep learning techniques have been largely employed recently in the domain of computervisionandhaveachievedtremendousresults.Therefore,video-basedhumanactivityrecognition using deep learning models have gained a lot of interest in recent years. Zhuetal. Proposed an action classification method by adding a mixed-norm regularization function to a deep LSTM network. One of the most popular deep learning method sin frames/image processing is Convolutional Neural Network (CNN). Wang et al. applied CNN to RGB and depth frames to automatically extract the features. The obtained features were passed through a fully connected neural network and achieved an improved accuracy Jietal. proposed a 3DCNN model which perform s 3D convolutions and extract spatial and temporal features bycapturingthemotioninformationforactivityrecognition.Simonyanetal.introducedConvNet, a two-stream convolution layer architecture that could achieve good results despite of limited training data. Khaire et al. proposed a model that train convents from RGB-D dataset and combined the soft max scores from depth, motion and skeleton images at the classification level to identify the activities. Karpathy et al. proposed the extension of CNN architecture in the first convolution layers over a 40 video chunk. Similarly, Tran et al. used a deep 3DCNN architecture (quiet similar to VGGnet)that utilize spatio temporal convolutions and pooling in all yers to improve the accuracy of the model. In comparison, we are more interested to explore how transfer learning can be leveraged with CNN models on benchmark data set to improve classification accuracy.

Depression is also common at all stages of dementia. It occurs in about 20– 40% of PwD. Identifying depression in PwD can be difficult. To date, there is no single test or questionnaire to detect the depression due to the complexities and multifaceted nature of the condition. The common approach to monitor and manage the above-mentioned behavioural symptoms s via direct observation by caregivers, family members and health care professionals. These technologies could be adapted into the early detection of behavioural symptoms that would aid caregivers and guide the headway of tailored interventions. Most ofthe above-mentioned technologies often use on body bio-sensing devices (e.g. actigraphs, accelerometers, biomarkersandbiopatches)formeasuringsignalslinkingbehaviouralsymptoms. Monitoring and recognition of aggression and depression using such systems isstill very much in its infancy. This could be due to the challenge faced by the researchers todevelopstandardalgorithmsthatcanadequatelyandconciselyrecognizebehaviouralsymptoms. In this paper, we propose a novel method for recognizing behavioural symptoms involving aggression and depression. The proposed approach benefits from the power of transfer learning(TL)byusingappearancefeaturesasdeepCNNfeatures,whichareextractedfromvariousstate-of-the-artdeepmodels(e.g.VGG16,Inception-V3andInception ResNet V2).Wealsoexplorethevariouslevelofabstractionbyexploringdifferentextraction points in a given CNN model (e.g. VGG16). This work includes the following novel contributions:

• To our knowledge, we are the first to report vision based recognition of behavioral symptoms(aggressive, depressive, happy and neutral)in PwD.

• We demonstrate the effectiveness of TL using different state-of-the-art deep CNN models for recognizing behavioural symptoms in PwD. We evaluate various combinations of deep CNN features using SVM.

• We introduce a novel image dataset to advance video based surveillance research for behaviour recognition.

Human action and behaviour recognition has many potential applications including intelligent surveillance, assistive technologies, robotics and human-computer interaction Many of these approaches explore the spatial configuration of body parts and hand-object interactions that often require body parts and/or object detector. These CNN models are trained and evaluated on very large and highly diverse datasets often consisting human-human, human-objects and human animals interactions. In contrast, the targeted behavioural symptoms are often expressed via body language (e.g. gestures) and facial expression, and usually a hard problem for a machine to differentiate various symptoms shown by the same person. It is also known as fine-grained recognition. Deep CNN models are comprised of multiple layers to learn representation of images/videos with multiple levels of abstractions through a hierarchical learning process. Such models learn from very general (e.g. Gaborfilters, edges, color blobs) to task-specific features as we move from first-layer to the last-layer. Thus, these models are explored for TL in solving visual recognition tasks. In TL, abase network is trained on a base dataset. Then, the learned features (e.g.weights) are adapted, or transferred toa second target network/model to be trained on a target dataset. This would work if the learned features are task-independent, which means they are suitable for both base and target task. More recently, it has been shown that it is possible to obtain state-of-the-art results using TL. This suggests the layers of deep models do indeed learn features that are fairly general. In this paper, we explorestrategiestostrengthenthisgeneralizability.Automaticmonitoringofthebehaviouralsymptom sisoftenbasedonwearable sensors.

[2] LITERATURE SURVEY

Although smoking prevalence is declining in many countries, smoking related health problems still leads the preventable causes of death in the world. Several smoking intervention mechanisms have been introduced to help smoking cessation. However, these method sare in efficient since they lack in providing real time personalized intervention messages to the smoking addicted users. To address this challenge, the first step is to build an automated smoking behavior detection system. In this study, we propose an accelerate meter sensor based non-invasive and automated framework for smoking behavior detection. We built a prototype device to collect data from several participants performing smoking and other five confounding activities. We used three different classifiers to compare activity detection performance using the extracted features from accelerometer data. Our evaluation demonstrates that the proposed approach is able to classify smoking activity among the confounding activities with high accuracy. The proposed system shows the potential for developing are al time automated smoking activity detection and intervention framework. Ruan Understanding and recognizing human activities is a fundamental research topic for a wide range of important applications such as fall detection and remote health monitoring and intervention. Despite active research in human activity recognition over the past years, existing approaches based on computer vision or wearable sensor technologies presents Several significant issues such as privacy (e.g., using video camera to monitor the elderly at home) and practicality (e.g., not possible for an older person with dementia to remember wearing devices). In this paper, we present a low-cost, unobtrusive, and robust system that supports independent living of older people. The system interprets what a person is doing by deciphering signal fluctuations using radio-frequency identification (RFID) technology and machine learning algorithms. To deal with noisy, streaming, and unstable RFID signals, we develop a compressive sensing, dictionary-based approach that can learn a set of compact and informative dictionaries of activities using an unsupervised subspace decomposition Our approach achieves efficient and robust activity recognition via a more compact and robustrepresentationofactivities.Extensiveexperimentsconductedinareal-liferesidentialenvironment demonstrate that our proposed system offers a good overall performance and shows the promising practical potential to underpin the applications for the independent living of the elderly.

This project presents an approach to recognize human activities using body poses estimated from RGB-Ddata.Wefocusonrecognizingcomplexactivitiescomposedofsequential or simultaneous atomic actions characterized by body motions of a single actor. We tackle this problem by introducing a hierarchical compositional model that operates at three levels of abstraction. At the lowest level, geometric and motion descriptors are used to learn a dictionary of body poses. At the intermediate level, sparse compositions of these body poses are used to obtain meaningful representations for atomic human actions. Finally, at the highest level, spatial and temporal compositions of these atomic actions are used to represent complex human activities.

The explosion of image data on the Internet has the potential of oster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harnessed and organized remains a critical problem. We introduce here a new database called “Image Net”, a large-scale ontology of images built upon the backbone of the Word Net structure. Image Net aims to populate the majority of the 80,000 synsets of Word Net with an average of 500-1000clean and full resolution images. This will result in tens of millions of annotated images organized by the semantic hierarchy of Word Net. This paper offers a detailed analysis of Image Net in its current state: 12 subtrees with 5247 synsets and 3.2 million images in total. We show that Image Net is much larger in scale and diversity and much more accurate thanthecurrentimagedatasets.Constructingsuchalarge-scaledatabaseisachallenging task.

With an aging population that continues to grow, dementia is a major global health concern. It is a syndrome in which there is a deterioration in memory, thinking, behaviour and the ability to perform activities of daily living. Depression and aggressive behaviour are the most upsetting and challenging symptoms of dementia. Automatic recognition of these behaviours would not only be useful to alert family members and caregivers, but also helpful in planning and managing daily activities of people with dementia (PwD).

In this work, we propose a vision-based approach that unifies transfer learning and deep convolution neural network(CNN)for the effective recognition of behavioural symptoms. The proposed method is evaluated on a newly created dataset, which is based on the dementia story line in ITVs Emmer dale episode.

In the existing work with wearable based or non-wearable based. Wearable based HAR system make use of wearable sensors that are attached on the human body. Wearable based HAR system are intrusive in nature. Non-wearable based HAR system do not require any sensors to attach on the human or to carry any device for activity recognition. Non-wearable based approach can be further categorized into sensor based HAR systems. Sensor based technology use RF signals from sensors, such as RFID, PIR sensors and Wifi signals to detect human activities. Sensor based HAR system are non-intrusive in nature but may not provide high accuracy. In the disadvantages of existing system are the following

- Require the optical sensors to be attached on the human and also demand the need of multiple camera settings.
- Wearable dives cost are high.
- Algorithm: Marker based motion Capture (MoCap) Framework.

In the proposed System Vision based technology use videos, image frames from depth cameras or IR cameras to classify human activities. Video-based human activity recognition can be categorized as vision-based according to motion features. The vision based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human. Therefore, this methodology is getting more consideration nowadays,consequentlymakingtheHARframeworksimpleandeasytobedeployedinmanyapplications.Themostcommon type of deep learning method is Convolutional Neural Network(CNN).CNN are largely applied in areas related to computer vision. The Advantages of proposed system are

- We use CNN to recognize human activities action recognition kinetics data set.
- We use transfer learning to get deep image features and trained machine learning classifiers.
- Does not require the user to carry any devices or to attach any sensors on the human

- Algorithm: Convolutional Neural Networks(CNN),VGG-16(also called OxfordNet)

[3] SYSTEM ARCHITECTURE

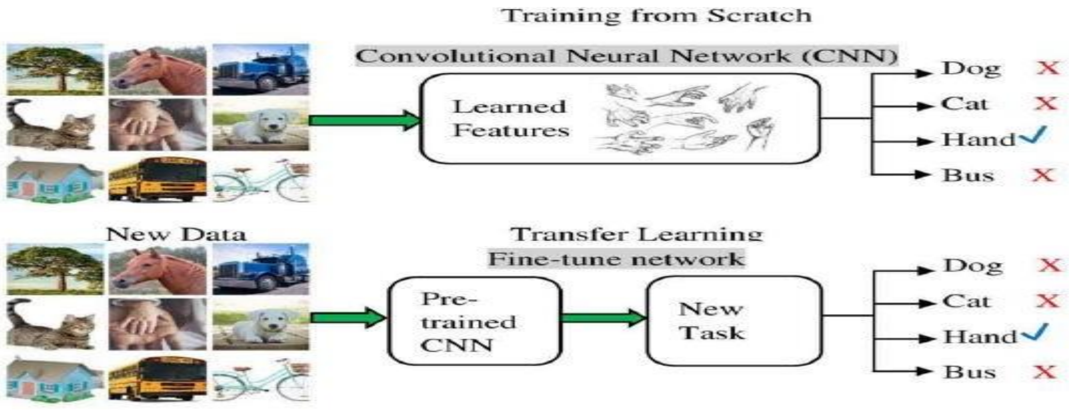


Fig. 1 Recognizes the human hand

Fig.2 Flow Chart Diagram

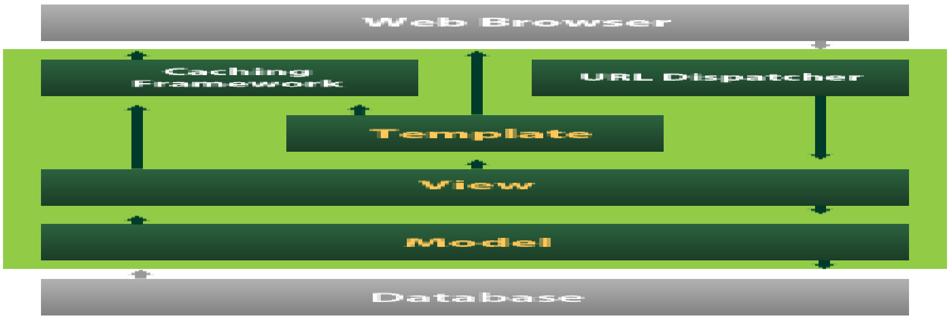


Fig. 3 Data base Architecture

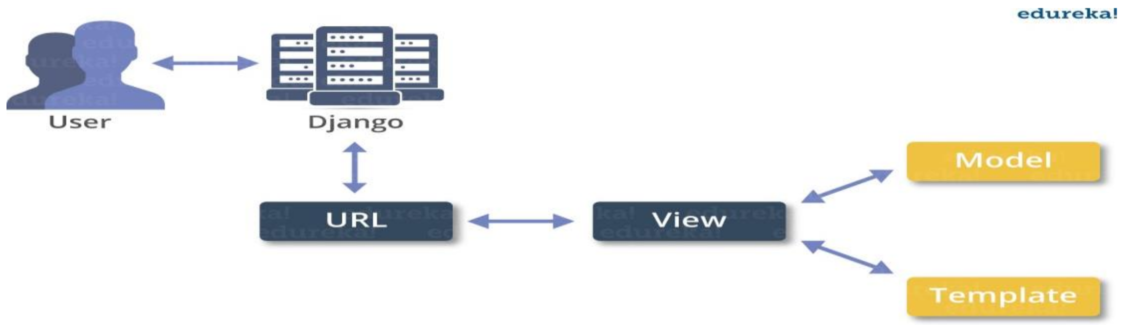


Fig.4 Django model diagram

[4] IMPLEMENTATION

4.1 Modules Description

i) User: The User can start the project by running mainrun.py file. User has to give – input(Video file path).The open cv class Video Capture(0) means primary camera of the system, Video Capture(1) means secondary camera of the system. Video Capture (Video file path) means without camera we can load the video file from the disk. Vgg16, Vgg19 has program it logically configured. User can change the model selection in the code and can run in multiple ways.

ii) HAR System: Video-based human activity recognition can be categorized as vision-based according. The vision based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human. Therefore, this methodology is getting more consideration nowadays, consequently making the HAR framework simple and easy to be deployed in many applications. We first extracted the frames for each activities from the videos. Specifically, we use transfer learning to get deep image features and trained machine learning classifiers.

iii) VGG16: VGG16 is a convolution neural network model. Deep Convolutional Networks forLarge-ScaleImageRecognition".Themodelachieves92.7%top-5testaccuracyinImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous model submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolution layer, respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU's**.**

iv)Transfer Learning: Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems. In this post, you will discover how you can use transfer learning to speed up training and improve the performance of your deep learning model.

4.2 Screenshots

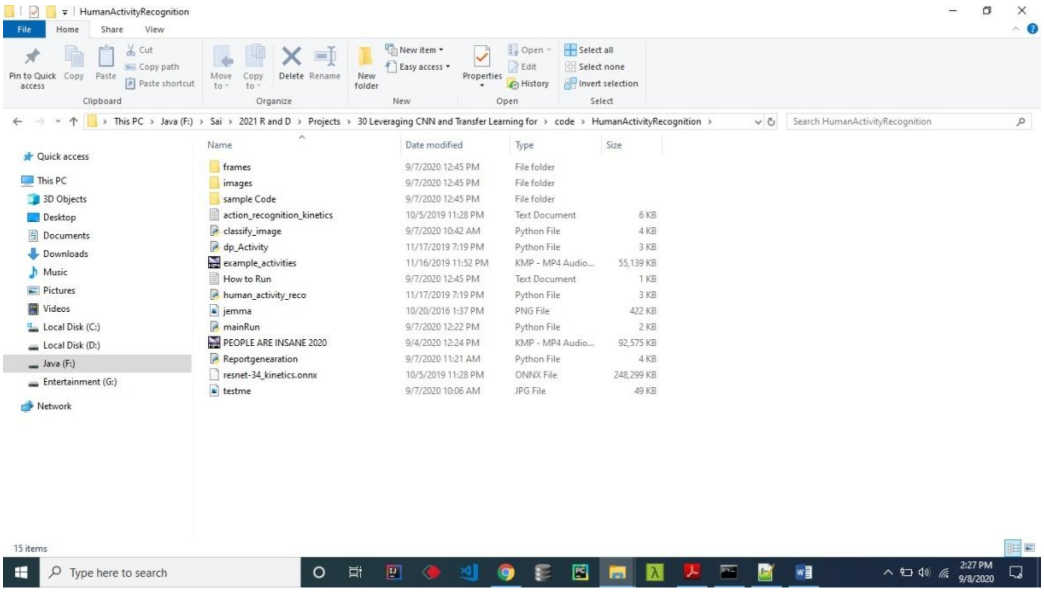


Fig. 5 : Get the file path in the project file location folder.

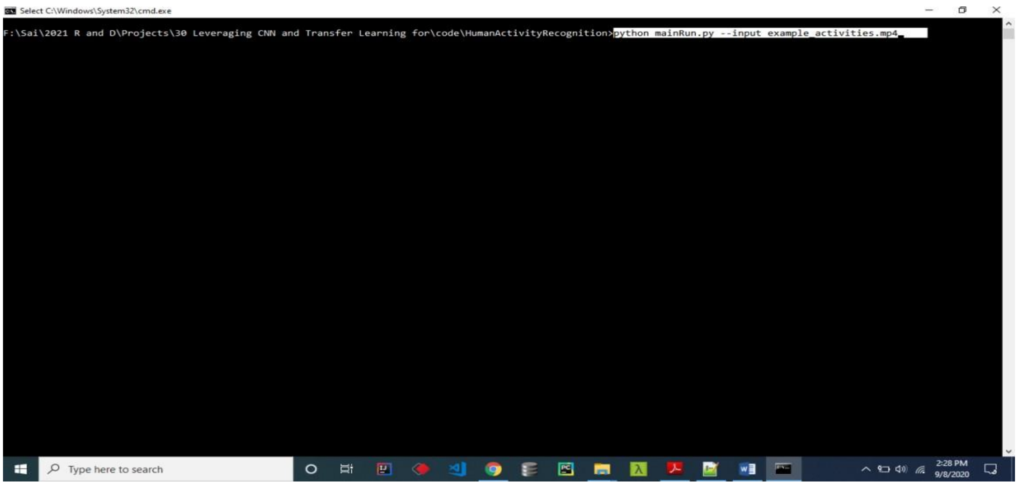


Fig. 6 : Run the main program.

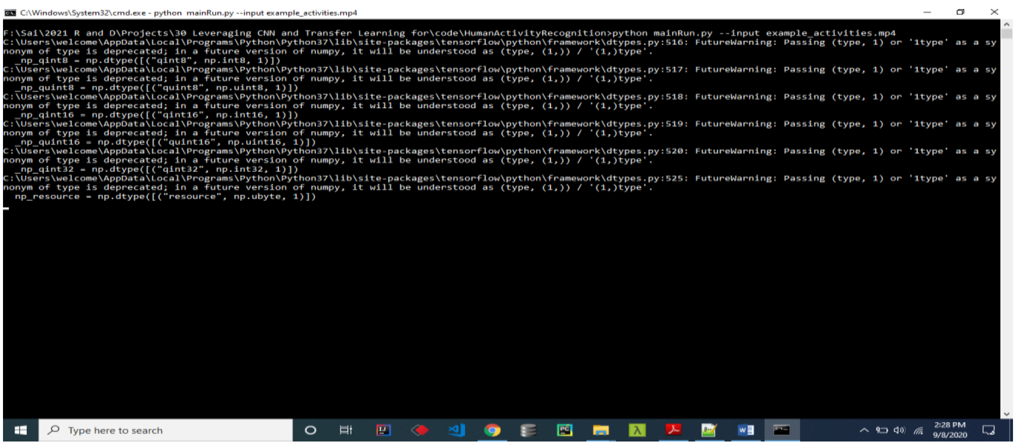


Fig. 7 : Loading Tensor flow Libraries

Tensor flow describes the activity names of each frame divide from the video file

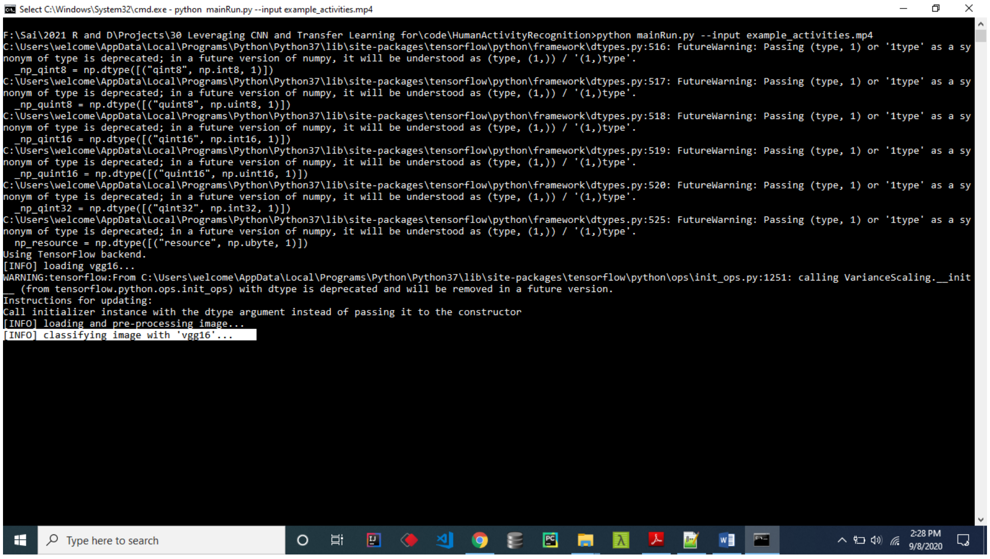


Fig. 8: Classification with vgg16

It gives the accuracy level of the activity from the given video file.

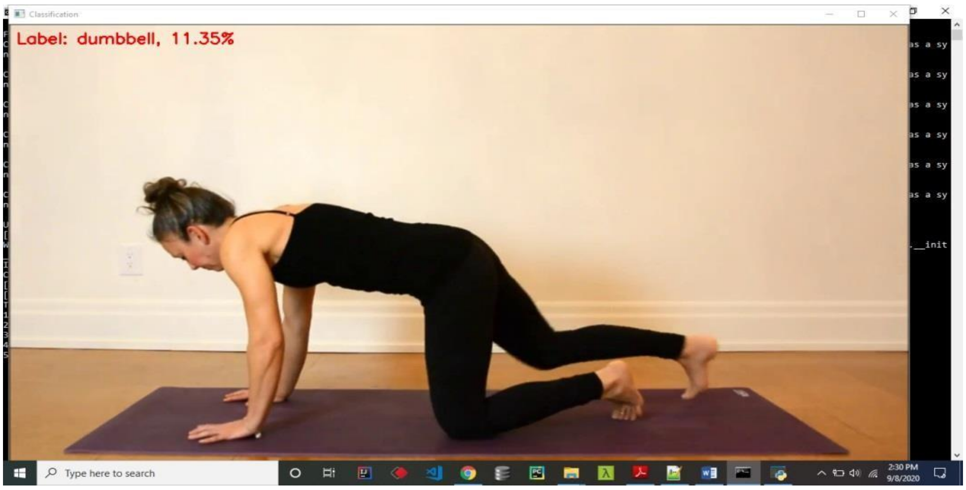


Fig. 9 : Get Image label

It gives the label names for the frames

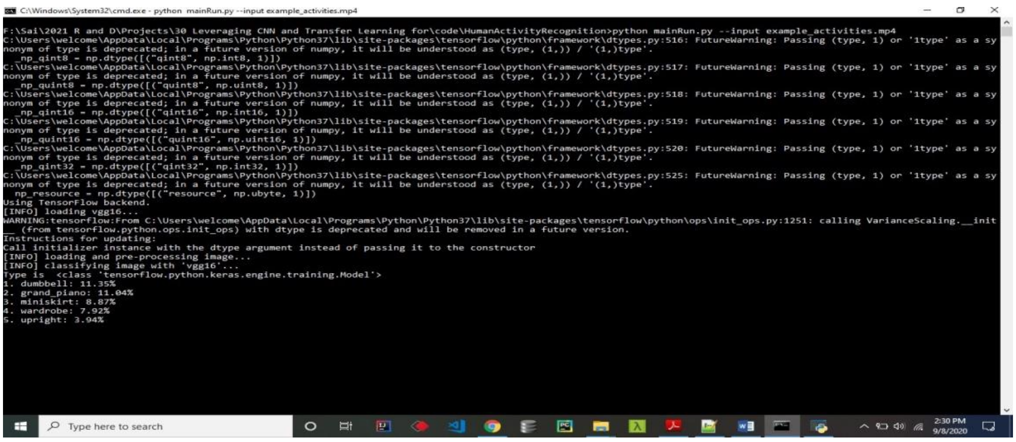


Fig.10: Result from image

The model detects the activity of the frames

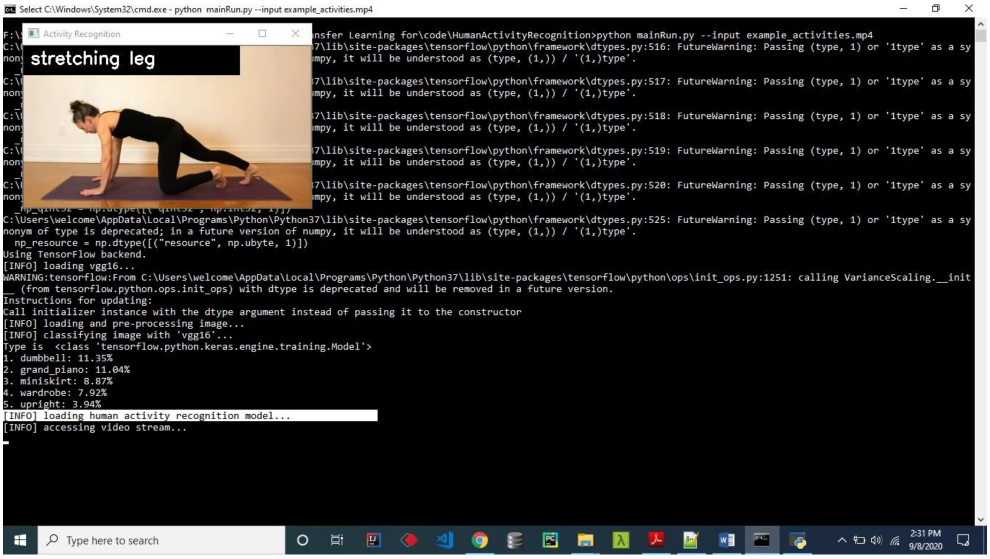
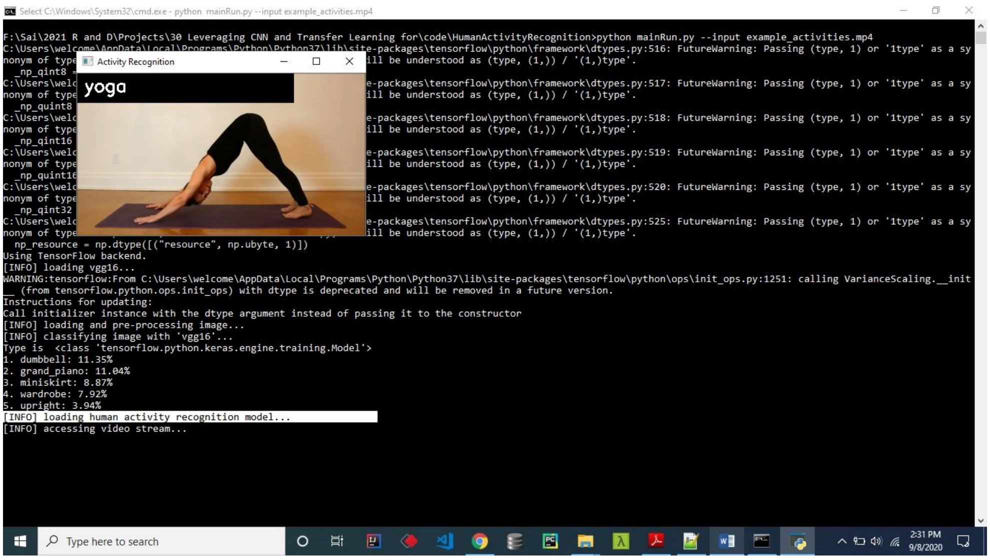


Fig. 11: Loading model HAR



The result gives the activity name as stretching leg.

Fig. 12: Result 1

Gives the name of next activity

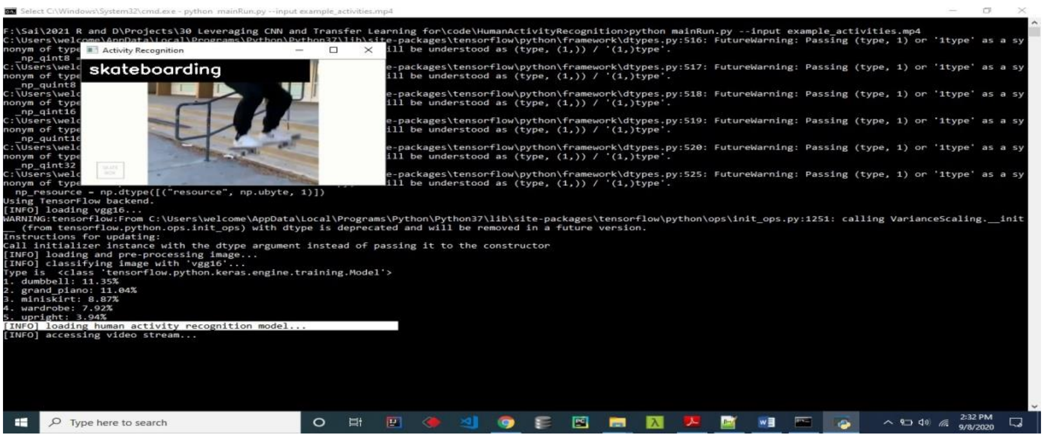


Fig. 13: Result 2

Gives the name of another activity.

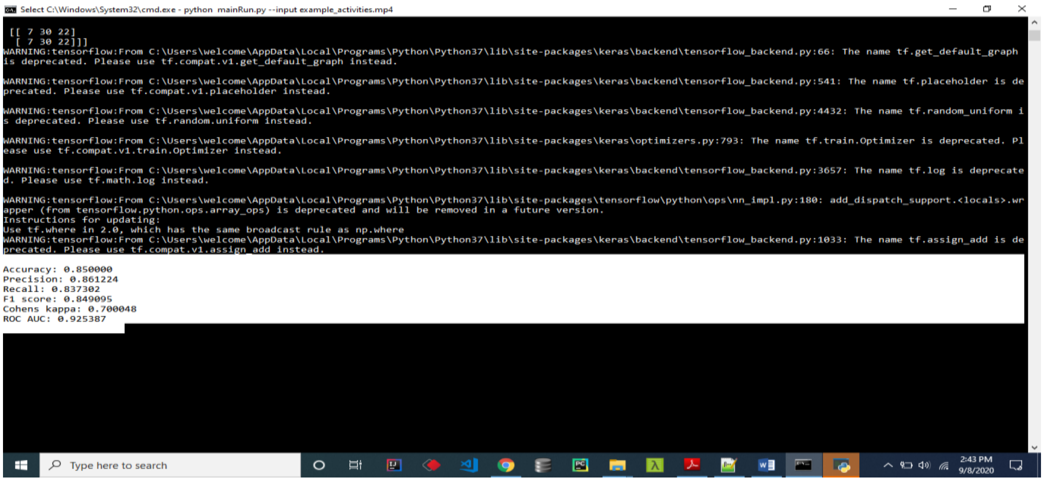


Fig.14 : Accuracy

Here we can get the how much accurate levels we obtained from the video files.

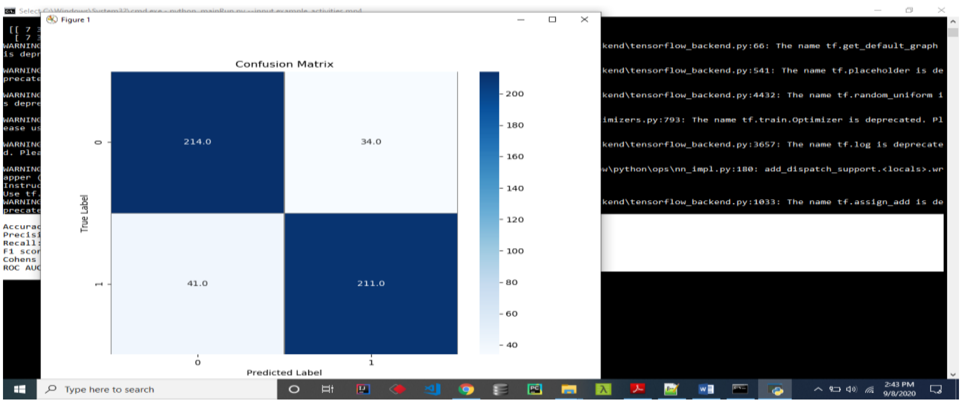


Fig. 15 : Confusion Matrix

Confusion matrix gives the performance of a classification. How fast it detects the activity name from the given input video file. This table shows the result of the past test cases.

[5] CONCLUSION

We used CNN models to predict the human activities from Wiezmann Dataset. We experimented with 3different Convolutional Neural Networks(CNN)for activity recognition. We have employed transfer learning to get the deep image features and trained machine learning classifiers. Our experimental results showed the accuracy of 96.95% usingVGG-16 with the implementation of transfer learning. Our experimental results showed thatVGG-16 outperformed other CNN models in terms of feature extraction. Our experimental results with transfer learning technique also showed highperformanceofVGG-16ascomparedtostate-of-the-artmethods. In future, we aim to extend this study by developing the context-aware recognition system to classify human activities. Also, we will extend our work to recognise complex human activities such as cooking, reading books, and watching TV.

References

[1] B. Bhandari, J. Lu, X. Zheng, S. Rajasegarar, and C. Karmakar, "Noninvasive sensorbasedautomatedsmokingactivitydetection,"inPro-ceedingsofAnnualInternationalConference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017,pp.845–848.
[2] L.Yao,Q.Z.Sheng,X.Li,T.Gu,M.Tan,X.Wang,S.Wang,andW.Ruan,"Compressiverepresentation for device-free activity recognition with passiverfid signal strength,"IEEE Transactions on MobileComputing,vol.17,no.2,pp.293–306,2018.
[3] I. Lillo, J. C. Niebles, and A. Soto, "Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos," Image and Vision Computing, vol. 59, pp.63–75,2017.
[4] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature elearning for skeleton based action recognition using regularized deep lstm networks," in Thirtieth AAAI Conference onArtificialIntelligence,2016.
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke,andA.Rabinovich,"Goingdeeperwithconvolutions,"inProceedingsoftheIEEEconference oncomputervisionandpatternrecognition,2015,pp.1–9.
[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scalehierarchical image database," in Proceedings of IEEE Conference on Computer Vision and PatternRecognition,June2009,pp.248–255.
[7] A. Jalal, N. Sarif, J. T. Kim, and T.-S. Kim, "Human activity recognition via recognizedbody parts of human depth silhouettes for residents monitoring services at smart home,"Indoorandbuiltenvironment,vol.22,no.1,pp.271–279,2013.
[8] K.SimonyanandA.Zisserman,"Two-streamconvolutionalnetworksforactionrecognition in videos," in Advances in neural information processing systems, 2014, pp. 568–576.
[9] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," inProceedingsoftheIEEEinternationalconferenceoncomputervision,2015,pp.1080–1088.
[10] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-streamconvnets,"arXivpreprintarXiv:1507.02159,2015.
[11] D.Tran,L.Bourdev,R.Fergus,L.Torresani,andM.Paluri,"Deepend2endvoxel2voxel prediction," in Proceedings of the IEEE conference on computer vision andpatternrecognitionworkshops,2016,pp.17–24.
[12] Y Sri Lalitha, SK Althaf Hussain Basha, Ayesha Mariyam, S Viswanadha Raju, "A Brief Research On Deep Learning Models", International Journal of Computer Engineering and Applications, Volume 13, Issue 6, November 2020, ISSN2321-3469.
[13] Sk. Althaf Hussain Basha, Sreedhar Jinka,Baijnath Kaushik, D.Praveen Kumar, A Jagan, "NLP: Context Free Grammars and Parse Trees for Disambiguiting Telugu Language Sentences", International Journal of Scientific Research in Computer Science, Engineering and Information Technology Volume 2 , Issue 7 , pp.332-337, 2017, ISSN : 2456-3307
[14] Ayesha Mariyam, SK Althaf Hussain Basha,, S Viswanadha Raju, "A Literature Survey On Recurrent Attention Learning For Text Classification", 2nd International Conference on Machine Learning, Security and Cloud Computing (ICMLSC2020), Springer Conference 18th & 19th December2020.
[15] SK Althaf Hussain Basha, Ayesha Mariyam, and S Vishwanadha Raju "Applications of Multi- Label Classification", International Journal of Innovative Technology and Exploring Engineering,pp.86-89,ISSN:2278-3075,Volume-9, Issue-4S2, March 2020.
[16] Ayesha Mariyam, SK Althaf Hussain Basha Sk, and Viswanadha Raju S , "A Brief Literature Survey on Text Classification Applications", International Conference on Devices, Intelligent Systems & Communications (DISC)2020.
[17] Venkata Pavan Kumar Savala, Sk Althaf Hussain Basha, Ranganath P, P V Ravi Kumar, "Information Inclusion: The Modern Rank and The Approach Forward", International Journal of Computer Engineering and Applications, Volume 13, Issue 6 , December2020, ISSN2321-3469.
[18] Sreedhar Jinka, Sk. Althaf Hussain Basha, Suresh Dara, Baijnath Kaushik, "Sequence Labelling for Three Word Disambiguation in Telugu Language Sentences", International Journal of Scientific Research in Computer Science, Engineering and Information Technology Volume 2 , Issue 7 , pp.311-315, 2017, ISSN :2456-3307.
[19] Baijnath Kaushik, Sk. Althaf Hussain Basha, Sreedhar Jinka, D Praveen Kumar, " Sequence Labelling for Two Word Disambiguation in Telugu Language Sentences", International Journal of Scientific Research in Computer Science, Engineering and Information Technology Volume 2 , Issue 7 , pp.321-327, 2017, ISSN

:2456-3307.
[20] Sreedhar Jinka, Sk. Althaf Hussain Basha, Baijnath Kaushik, D. Praveen Kumar, A. Jagan, " Empirical Analysis of Context Sensitive Grammars and Parse Trees for Disambiguiting Telugu Language Sentences", International Journal of Scientific Research in Computer Science, Engineering and Information Technology Volume 2 , Issue 7 , pp.328-331, 2017, ISSN :2456-3307.
[21] GNR Prasad , SK Althaf Hussain Basha, Mallikharjuna Rao K M GnanaVardhan "A Review of Predictive And Descriptive Data Mining Techniques in Higher Education Domain, International Journal of Computer Engineering and Applications(IJCEA),Volume 13, Issue 6, January. 21, ISSN2321-3469.
[22] G.V.S.Raju, Sk Altaf Hussain Basha, K.Venkata Subbaih, A.V.N. Sharma, "Facilities Layout Optimization using Genetic Algorithm", International Conference on Advanced Computing Technologies (ICACT 2008), GRIET,Hyderabad,32-36,2008, ISSN: 9788178003
[23] Dr. G. N. R. PRASAD, "Evaluating student performance based on bloom's taxonomy Levels", Journal of Physics: Conference Series IOCER2020, Vol 1797 12/ 2021, Doi:10.1088/1742-6596/1797/1/012063