Human Activity Recognition using Resnet-34 Model

Akansha Abrol, Anisha Sharma, Kritika Karnic, Raju Ranjan

Abstract*: Activity recognition has been an emerging field of research since the past few decades. Humans have the ability to recognize activities from a number of observations in their surroundings. These observations are used in several areas like video surveillance, health sectors, gesture detection, energy conservation, fall detection systems and many more. Sensor based approaches like accelerometer, gyroscope, etc., have been discussed with its advantages and disadvantages. There are different ways of using sensors in a smartly controlled environment. A step-by-step procedure is followed in this paper to build a human activity recognizer. A general architecture of the Resnet model is explained first along with a description of its workflow. Convolutional neural network which is capable of classifying different activities is trained using the kinetic dataset which includes more than 400 classes of activities. The videos last around tenth of a second. The Resnet-34 model is used for image classification of convolutional neural networks and it provides shortcut connections which resolves the problem of vanishing gradient. The model is trained and tested successfully giving a satisfactory result by recognizing over 400 human actions. Finally, some open problems are presented which should be addressed in future research.

***Keywords*: Video Surveillance, Resnet, Convolutional Neural Network, Kinetic Dataset.**

I. INTRODUCTION

Human activity recognition has gained a wide range of attention in the past few decades. The data collected through activity monitoring can be used in several places like safe driving, controlling the crime rates, taking suitable actions during medical treatment and many more. Activity recognition also assists elderly people to make their life easier and simpler. Human beings have the potential to identify other human's activities through observation and communication. However, machines need to go through a learning phase to be able to recognize activities. Activity recognizer is capable of recognizing a wide range of activities like walking, applauding, reading, washing hands, and many more. The activities like handshaking, hugging and more comes under the category of human-to-human interaction in which two humans are involved whereas reading a book or newspaper comes under the category of human-to-object interaction in which one human and one object is involved. Some of these activities can either be simple or very complex. Complex activities may be broken down into simpler activities which will make them easier to understand.

The data can be collected through various sources like accelerometer, sensors, images or video frames. While collecting the data through sensors, people need to wear more than one sensor in different parts of their body. The data collected needs to be processed in various ways. The raw data is segmented and different features are extracted. This process might be a challenging task in case of sensors. Deep neural network which is used in this paper, helps in the extraction of more significant features. Classification algorithms are then applied on these features. The classification model is based on a training dataset and is used to recognize activities. Hidden Markov Models (HMMs), Support Vector Machine Classifier and Feed-forward neural network are few classification algorithms. Recent modern-day approaches include Recurrent Neural Networks (RNN), Convolutional neural networks and more. These approaches are far more suitable and convenient as the processing time of the image is reduced.

In this paper, an intelligent human activity recognition system is developed through the Resnet-34 algorithm. Resnet is one of the most successful architectures in image classification. It provides shortcut connections that allow a signal to bypass one layer and move to the next layer in sequence. It consists of two convolutional layers, and each layer is followed by batch normalization and a rectified linear unit (ReLu). In the case of plain neural networks which simply stacks layers, accuracy starts deteriorating as deeper it starts to converge. Therefore, the main advantage of using the Resnet model is resolving the problem of vanishing gradient. This model is easy to optimize and shows less training error as compared to other models.

II. LITERATURE SURVEY AND COMPARATIVE ANALYSIS

There are several researches on the human activity recognition system which shows several aspects of this model. A simple yet robust activity recognition system uses smartphone sensors like accelerometer and gyroscope to combine the output of an activity classification algorithm [1].

Manuscript received on May 17, 2021.
Revised Manuscript received on May 21, 2021.
Manuscript published on May 30, 2021.
* Correspondence Author
Akansha Abrol*, Department of Computing Science and Engineering, Galgotias University, Greater Noida (U.P), India. Email: abrolakansha@gmail.com
Anisha Sharma, Department of Computing Science and Engineering, Galgotias University, Greater Noida (U.P), India. Email: sanisha0601@gmail.com
Kritika Karnic, Department of Computing Science and Engineering, Galgotias University, Greater Noida (U.P), India. Email: kritkakarnic73@gmail.com
Raju Ranjan Department of Computing Science and Engineering, Galgotias University, Greater Noida (U.P), India. Email: drraju.ranjan@galgotiasuniversity.edu.in

However, sensor-based approaches face a lot of complexities in distinguishing different features. The model proposed in this paper is successfully able to differentiate between activities like walking and running due to the efficiency of the Resnet-34 model which provided high optimization and reduced overfitting during training. A CNN based patient activity recognition model obtained an improved performance with minimum latency [2]. However, wearing sensors might not be comfortable for the patients for a long amount of time. A survey on CNN based approach states the cold start problem in CNN [3]. The results showed that a pretrained CNN is efficiently able to address this problem.

In real world scenarios, the use of sensors requires some skills and sophisticated equipment that makes them accessible to only experienced users. There always needs to be a user's willingness to perform continuous monitoring tasks. Vision based systems do not require ordinary users to wear several uncomfortable devices on different parts of their body which is a drawback of technique [4]. Hence vision-based systems gain more acceptability of use from society. A framework which mainly consists of a deep attentive 3D residual network and a multistage fusion strategy showed higher recognition accuracy and satisfactory training efficiency [5]. Kumar, Sagar, Awasthi proposed a real time action system where classification is done using the K-nearest neighbor (KNN) algorithm [6]. Videos are more accurate as relationships between two successive frames can be established. A two-phase recognition method is used for recognition of complex activities [7]. The first phase described automatic learning while second phase threw light on modelling temporal dependencies between their activation. The result shows that concurrent activities are recognized better but more complex activities are unable to be recognized appropriately on a larger dataset. An optical flow descriptor considered only the features derived from motion [8].

One simple accelerometer may be utilized to recognize daily life activity much more efficiently based on self-collected and public dataset [9]. The results might have adverse effects on users even if there is a slight inaccuracy. Wi-Fi based activity recognition systems are gaining a huge popularity [10]. In Wi-Fi based activity recognition different activities are mapped to the received signal records. Sensors embedded in smart phones is a very hassle-free technique [11]. Exercise activity recognition (EAR) is considered an important part of smart healthcare and ambient living [12]. These technologies are far better than burdensome wearables or other wireless devices. It is no doubt that these are some of the best systems proposed so far. Human activity recognition is becoming a necessity today due to its beneficial uses in different fields. It plays a crucial role in the health care sector to monitor the patients. There are still a lot of things that need to be explored more so that this system works more efficiently and effectively in recognizing more complex activities. Recognition of group activities is a challenging task which still needs to be addressed. Deep learning-based approaches are getting huge attention nowadays due to its promising results and progress they have made in terms of detection and recognition. It has made the complex process of recognition of activity much simpler and easier.

III. METHODOLOGY

The two crucial steps for the implementation are training and recognition. Stochastic gradient descent is used. Training samples are generated from the training data. To proceed next, a temporal position in the video is selected for the generation of training samples. Then around the selected position, a sixteen-frame clip is generated. Start looping around the video as much as required if it is less than sixteen-frames. Next, a spatial position and spatial scale is selected as per requirement. Then follow the corner cropping strategy and crop four corners to 112 X 112 pixels. A weight decay of 0.001 is included in the training parameter. Learning rate is started from 0.1 and after validation loss saturates, reduce it to 0.01.

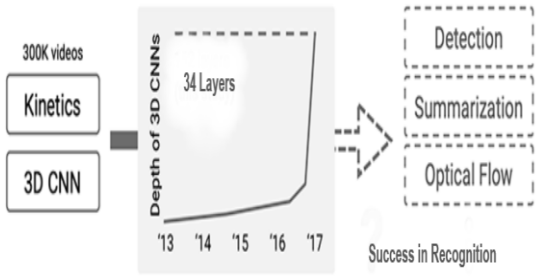

Fig. 1 Flow process of resnet model.

For the recognition, looping over each frame from 0 to sample duration is done. A frame is grabbed by reading the video and if the image is grabbed successfully then return true. Then the frame is resized to 400 pixels and added to the frame list. Next, the frame is resized again so that the images are of the same size. Once the frame array is completely filled then creation of the blob will begin. A blob of input frames is constructed which is passed through the network. The image is preprocessed first to obtain the correct prediction. Subtraction of mean value and scaling by using a factor of 0.1 is done to preprocess the image. This helps to make the image less sensitive to background and lighting conditions. All this is done with the help of OpenCV's deep neural network module. This blob is created to have images with the same spatial dimensions. A forward pass is applied after this and the system grabs a label with the correct prediction value.

Kinetic dataset is used to train the model. This dataset consists of 400 classes of human activity including more than 650,000 trimmed videos which lasts around 10 seconds. The videos are resized without any change in the aspect ratio. The dataset contains a huge range of activities including applauding, archery, bartending, crying, handshaking, and many more. This dataset includes frames that relate to target action. The absence of noise and unrelated frames makes this dataset the most suitable for training. The number of training, validation and test sets are approximately 580000, 30000, 40000 respectively. Training the Resnet-34 model on kinetic dataset does not result in overfitting. The result of this experiment can be very important for future progress in the field of computer vision.

Retrieval Number: 100.1/ijrte.A58960510121
DOI: 10.35940/ijrte.A5896.0510121

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

IV. RESULT AND DISCUSSION

The validation loss on kinetic dataset is found to be fairly higher than training loss which concludes that training resnet-34 model on kinetic dataset
does not result in overfitting.

The analysis shows that the kinetic dataset is efficient for the training of resnet-34 without causing any overfitting. An average accuracy of 71.2 is achieved on the kinetic validation set over Top-1 and Top-5. Hence kinetics can train deep 3D CNN from scratch.

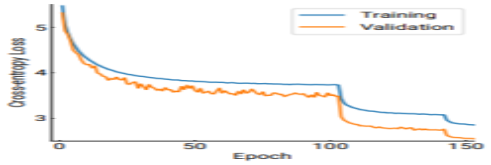

Fig. 2 Training and validation loss on kinetic dataset.

Next, it is attempted to adjust kinetics pre-trained CNN model on HMDB-51 and UCF-101 datasets. HMDB-51 includes videos extracted from movies. It contains almost 6 thousand video clips divided into 51 categories. UCF-101 includes a total number of 101 action classes which are further divided. It has 13320 videos from 101 classes. This dataset has realistic videos where there is a large amount of clutter in background and variations in camera motion. All these things make this dataset very challenging.

Table- I: The Top-1 accuracy for Resnet-34 (Scratch) model on HMDB-51 and UCF-101

Dataset	Accuracy
HMDB-51	19.2%
UCF-101	43.4%

Table- II: The Top-1 accuracy for the proposed model on HMDB-51 and UCF-101

Dataset	Accuracy
HMDB-51	58.8%
UCF-101	86.7%

The results showed that the model which originally was pre-trained on kinetic dataset showed efficacious results on HMDB-51 and UCF-101 dataset. The model pre-trained on the kinetic dataset is better than the one trained from scratch. This shows that a model pre-trained on a large dataset shows a good performance on a small dataset.

It is observed that personal actions as well as person to person actions like drinking, writing, shaking hands are labelled very accurately. Fine grained actions like swimming, yoga, cooking requires temporal reasoning to distinguish. Such actions showed a reduced amount of accuracy. This is because there is parent-child grouping for example music can be labelled into playing drums, violin, harp, etc. Similarly, actions like cooking, dancing can also be divided into smaller classes. This decomposition will allow having better prediction as well as label. The main focus of this dataset is on human actions rather than events. A more detailed dataset can be taken for this purpose to obtain an improved result.

Fig. 3 Demonstration of activity (Stretching arms)

Fig. 4 Demonstration of activity (washing hands)

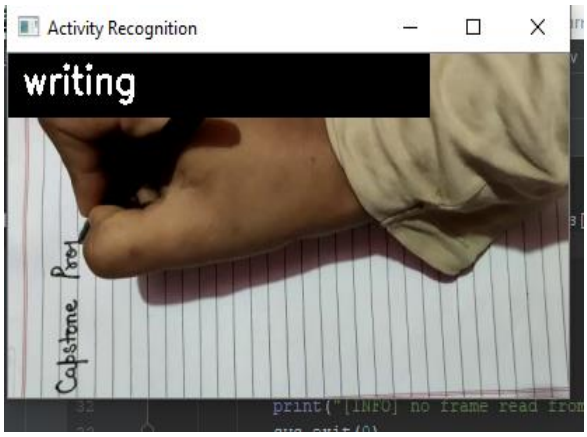

Fig. 5 Demonstration of activity (writing)

V. CONCLUSION AND FUTURE SCOPE

In this work the process of activity recognition is discussed and different methods of activity recognition are compared. Image recognition has become an important area of research for the advancement of computer vision. Actions could be anything like playing football, eating, dancing, etc. It is seen how deep learning models allow learning from simple to complex features due its layer-by-layer structure. All this has been possible due the capabilities of modern computers and datasets. Deep learning is continuously providing solutions to a variety of problems. Activity recognition benefits many applications such as smart homes and ambient living. Recognition of activities like fighting in public places and vandalism becomes very necessary to enforce a state of law and peace in the world.

There has been a lot of progress in recognition of activities. However, there are a lot of challenges like recognition of complex as well as simultaneous activities. Activities like walking while listening to music, singing while dancing are known as simultaneous activities. These activities become confusing and difficult to recognize.

Many studies are still being done in order to fully overcome such problems. Sensor based technologies also face some challenges like installation of devices on different parts of human bodies to directly measure activities. It becomes a burden for users to wear sensors installed in their watches, clothes, bracelets, etc. External sensors are installed in the environment at different locations. GPS receivers are limited to the outdoor environment which makes the usage of sensors limited to particular regions. In a smart home, sensors need to be installed in every door and equipment of use. Installation and maintenance of such a huge network is quite cumbersome. These sensors can be replaced with the help of cameras. In this paper Resnet-34 model is used in which the process of training and recognition is followed. An in-depth implementation of the work is explained. The model is successfully tested to achieve the desired result. Resnet-34 provides an optimized result due to the fact that it follows multiple levels of skip connections. A weight matrix can be used in order to call these skip weights. During recognition blob is used because a batch of multiple images is built that needs to be passed through the action recognition network. This gave it an advantage of taking spatiotemporal information. The most important feature of such a method is automatic learning features from huge amounts of data. Kinetic dataset is used which provides a satisfactory accuracy level in order to derive almost 400 classes of human actions. The video clips are from YouTube with a high-quality resolution. There could be more than one action in a particular clip. If simultaneous activities are taking place like “texting” while “walking” or “eating” while “chatting”, then it will be labelled only under one of the classes and not both. Some activities require more emphasis on the object in order to differentiate, like playing different kinds of musical instruments. The proposed system could be used to monitor new staff to ensure they are working properly, keep a check in restaurants if the customers are served properly and automatically categorize a dataset of video on disk. Therefore, activity recognition systems are becoming a basic tool in many aspects of life. For further work, usage of a dataset with more than 400 activities could be made in order to increase the level of accuracy and make the system more flexible. It is observed that if there is a deep hierarchy of activities like yoga which has different positions, dance which has different forms, cooking in which there are different kinds of food and many other such activities, then it could significantly help to achieve better performance.

ACKNOWLEDGMENT

The team members of the research project want to sincerely thank our guide Assistant professor Mr. Raju Ranjan and the Department of Computing Science and Engineering, Galgotias University, India for their encouragement and support for completion of this work.

REFERENCES

1. A. Wang, S. Zhao, C. Zheng, H. Chen, L. Liu and G. Chen, "HierHAR: Sensor-Based Data-Driven Hierarchical Human Activity Recognition," in *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3353-3365, 1 Feb.1, 2021, doi: 10.1109/JSEN.2020.3023860.
2. M.A. Gul, M.H. Yousaf, S. Nawaz, Z.U. Rehman, and H.W. Kim, “Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture”. Electronics 9:1993 (2020). doi: 10.3390/electronics9121993
3. F. Cruciani, A. Vafeiadis, C. Nugent, et. al., “Feature learning for Human Activity Recognition using Convolutional Neural Networks”. CCF Trans. Pervasive Comp. Interact. 2, 18–32 (2020). https://doi.org/10.1007/s42486-020-00026-2
4. A. V. Vesa *et al*., "Human Activity Recognition using Smartphone Sensors and Beacon-based Indoor Localization for Ambient Assisted Living Systems," *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2020, pp. 205-212, doi: 10.1109/ICCP51029.2020.9266158
5. Z. Zhang, Y. Yang, Z. Lv, C. Gan and Q. Zhu, "LMFNet: Human Activity Recognition Using Attentive 3-D Residual Network and Multistage Fusion Strategy," in *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 6012-6023, 1 April1, 2021, doi: 10.1109/JIOT.2020.3033449
6. R. Kumar, L.K. Sagar, S. Awasthi, “Human Activity Recognition from Video Clip”. Solanki V., Hoang M., Lu Z., Pattnaik P. (eds) Intelligent Computing in Engineering. Advances in Intelligent Systems and Computing, Vol 1125. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2780-7_31
7. K. Thapa, Z.M.A. Al, B. Lamichhane, S-H. Yang, “A Deep Machine Learning Method for Concurrent and Interleaved Human Activity Recognition”. Sensors, 20, 5770 (2020). https://doi.org/10.3390/s20205770
8. A. Ladjailia, I. Bouchrika, H.F Merouani, N. Harrati and Z. Mahfouf, “Human activity recognition via optical flow: decomposing activities into basic actions”. Neural Comput & Applic 32, 16387–16400 (2020). https://doi.org/10.1007/s00521-018-3951-x
9. J. Lu, X. Zheng, M. Sheng, J. Jin and S. Yu, "Efficient Human Activity Recognition Using a Single Wearable Sensor," in *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11137-11146, Nov. 2020, doi: 10.1109/JIOT.2020.2995940
10. Q. Zhou, J. Xing, Q. Yang, “Device-free occupant activity recognition in smart offices using intrinsic Wi-Fi components”. Building and Environment, 172, 106737(2020). https://doi.org/10.1016/j.buildenv.2020.106737
11. A. Ferrari, D. Micucci, M. Mobilio and P. Napoletano, "On the Personalization of Classification Models for Human Activity Recognition," in *IEEE Access*, vol. 8, pp. 32066-32079, 2020, doi: 10.1109/ACCESS.2020.2973425
12. X. Yang, R. Cao, M. Zhou and L. Xie, "Temporal-Frequency Attention-Based Human Activity Recognition Using Commercial WiFi Devices," in *IEEE Access*, vol. 8, pp. 137758-137769, 2020, doi: 10.1109/ACCESS.2020.3012021

AUTHORS PROFILE

Akansha Abrol, will be receiving B-tech degree in 2021 from Galgotias University, Greater Noida. She has participated in various workshops and successfully completed them. She has received digital credential from the Cisco Networking Academy on completion of ITN course. She has completed online internships as a junior software developer. She is the winner of many competitive programming contests and has received various certifications in the field of computer science. She has done several projects at college level. Her research focuses in the area of computer vision, design and implementation of human activity recognition systems, data science and their implementation in the real world.

Anisha Sharma, will be receiving B-tech degree in 2021 from Galgotias University, Greater Noida. She has participated in various hackathons and competitive programming contests held at different levels. She has received digital credential from the Cisco Networking Academy on completion of ITN course.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

She has completed online internships as a web developer and participated in various training programs. She has participated in the Advanced certification program in Cybersecurity and networking by E-learning Center IIT Roorkee. She has done several projects at college level. Her research focuses in the area of human activity recognition systems, classification models, motion detection and their implementation in the real world.

Kritika Karnic, will be receiving B-tech degree in 2021 from Galgotias University, Greater Noida. She has participated in various competitive programming contests and was able to reach a certain level. She has received digital credential from the Cisco Networking Academy on completion of ITN course. She has completed online internships as a junior software developer and participated in various events held online by Microsoft. She took part in the Certification program in AI and ML by Eckovation. She has done several projects at college level. Her research focuses in the area of human activity recognition systems, motion detection and their implantation in the real word.

Raju Ranjan, profile Assistant professor CSE department, Galgotias University, Greater Noida. He holds a Ph.D. in C.S.E. (Uttarakhand Technical University) and has done M-Tech. in C.S. (JRN University), Masters in Physics (Magadh University). He has participated in a faculty development program named Ubiquitous Computing and staff development program. He participated in the National seminar on computing and intelligence systems sponsored by AICTE. He has successfully completed a workshop on high impact technical skills from Dale Carnegie Training and chaired sessions at different conferences. He was a reviewer of several conference papers and journals. He is a lifetime member of Computer Society India and Indian society of Technical Education. His research area focuses on Data Mining, Network Security and Cryptography.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication