Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7 July 2021: 12898 – 12908

Research Article

Human Activity Recognition using 3D CNN

[1]Monalika Padma Reddy, , [2]Sheba Selvam [3]Meghana AC, [4]Ashwitha NA
Visvesvaraya Technological University, Karnataka, India
Email: [1] monalika.6671@gmail.com, [2]shebaselvam@bnmit.in, [3] meghanachandar87@gmail.com, [4]ashwitha.na.19@gmail.com

ABSTRACT

One of the most common biometric strategies is Human Activity Recognition, which is popularly known as HAR, which has gained immense popularity and significance over the last few years because of its applications. Human Activity Recognition deals with predicting the activity of a person based on the usual movements of humans such as sitting, standing, and so on. This paper presents an implementation of a Vision-Based Human Activity Recognition. The model is designed to predict the user's physical activity with sufficient confidence and overcome the disadvantages of the sensor-based methods in terms of cost, need for fixed infrastructure, inaccurate results, and discomfort due to the physical contact with the person's body. The proposed model uses convolution neural network architectures to meet the requirements. The 3D CNN model is used to implement the Vision-Based Human Activity Recognition. The UCF-50 dataset is used to train the model. The model will classify the images based on the activity that is being performed. Videos will be given as an input to the model and then the pre-processing will be done, which will be given to the model, and then the model will classify the activity. The vision-based human activity recognition system has numerous advantages in terms of accuracy, cost-effectiveness, and user-friendliness. It has a wide range of applications in elderly care, surveillance system, and anomalous behavior detection.

Keywords - *Human Activity Recognition, convolution neural network, Vision-Based Human Activity Recognition, 3D CNN architecture*

1 Introduction

Smart homes play a significant role in providing intelligent dwellings in the future is an inevitable fact. Not only does smart home technology control the incorporated lighting, heating, electrical, and all domestic components, but it also can recognize the activity of all home residents. Moreover, in conjunction with recognizing the activity of occupants, by utilizing machine learning techniques it can make a decision and prepare sufficient devices and services based on the user's need. Therefore, due to the increasing demand for human activity recognition in terms of security and health care especially elderly and child care, it has become a noteworthy issue in recent years.

Human Action Recognition (HAR) from a set of video sequences is a challenging problem in computer vision technology and is fundamental to a variety of applications in many different research areas, such as academia, security, industry, and consumer electronics. It is the problem of predicting the activity of a person by tracing or tracking the movements.

The basic idea of human activity recognition is to recognize the subject's activity, which can be used for monitoring. Deep learning models such as convolution neural networks have demonstrated the state

of the art results by learning the features from the raw data. The methods of activity recognition include sensor and vision-based categories.

Sensor-based methods which contain wearable and ambient sensors seem to be traditional methods of human activity recognition in smart homes. In both sensors-based approaches to analyze human motion, data have been collected and conveyed by sensors. Since wearable sensors which have been attached to the body and ambient sensors have been installed all around the home, they can be annoying for residents. In addition, sensors can produce noise and wrong alarms which can lead to inaccurate results. In recent years to overcome the mentioned drawbacks of sensors in collecting accurate and sufficient data, vision-based methods (as shown in Fig.1) have gained popularity in human activity recognition researches.

Fig.1. Vision-based human activity recognition

The primary technique used in vision-based human activity recognition in computer vision. It is used for tracking and understanding the behavior of people through videos taken by cameras. Vision-based methods take the advantage of using diverse camera types to provide more accurate and adequate data than sensor-based methods. This is helpful in daily life monitoring, personal biometric signature, elderly and youth care, localization, industry manufacturing, and assisting surveillance systems. Fig.2. shows the applications of vision-based Human activity Recognition

Fig.2. Applications of vision-based Human activity Recognition

1.1 Problem Statement

Human activity recognition is a broad field of study concerned with identifying the specific movement or action of a person. The sensor-based approach to human activity recognition has disadvantages in terms of inaccurate results and discomfort to the users.

The purpose of the Vision-Based Human activity recognition is to design a generalized classifier is designed to predict the user's physical activity with sufficient confidence and overcome the disadvantages of the sensor-based methods.

2 Literature Review

In the work presented by A.Murugeshwari, S. Tamil Selvi [1], a deep learning convolutional neural network is presented. The model can classify the activities being performed directly on the raw inputs. Inception v3 is used to extract the features from the images. A multi-layered CNN and LSTMs are used. LSTMs are used for the prediction of visual time series problems. The CNN layers used for feature extractions are combined with LSTMs. The softmax function is used for activity recognition. A 2048 layered LSTM was followed by a 512 thick layer and utilized a 0.5 dropout.

In this work presented by R. Mutegeki and D. S. Han [2], the generic HAR framework based on Long Short-Term Memory (LSTM) networks for time-series domains for smartphone sensor data is proposed. A 4-layer CNN-LSTM, a hybrid LSTM network is proposed to improve recognition performance. The model is trained on the UCI-HAR dataset and 10-fold validation protocols and LOSO cross-validation were used for model evaluation. The model accuracy was around 95%. Bayesian technique optimization techniques are also used for hyperparameter tuning.

The paper [3] by I.Lillo, et al. presents an approach to recognize human activities using body poses estimated from RGB-D data. The focus is laid on recognizing complex activities composed of sequential or simultaneous atomic actions characterized by the body motions of a single actor. The results show the benefits of using a hierarchical model that exploits the sharing and composition of body poses into atomic actions and atomic actions into activities. Multi-class discrimination providing useful mid-level annotations is achieved by the model.

In this work presented by Z. Wharton, et al. [4], a vision-based approach that unifies transfer learning and deep convolutional neural network (CNN) for the effective recognition of behavioral symptoms of dementia was developed. The state-of-the-art CNN features with the hand-crafted HOG-feature performance, as well as the combination using a basic linear SVM, are compared here. Automatic recognition of these behaviors was done and an alert was given to family members and caregivers, which was helpful in planning and managing the daily activities of people with dementia.

In the work presented by P. Y. Han, et al. [5], a localized Spatio-temporal representation, alongside Motion History Image (MHI), Motion Energy Image (MEI), and Binarized Statistical Image Features (BSIF), is proposed for human action recognition. In this work, the information of timestamp and ratio of colors are extracted from the silhouette of the MHI template. This information is utilized for encoding movement dynamics to derive a temporal representation. This temporal representation preserves transient information of actions. Subsequently, local descriptors are computed from MHI and MEI temporal templates via BSIF. The computed localized temporal representation is classified by using a linear SVM. The proposed system offers promising performance in human action recognition with about 90% accuracy.

In the work presented by J. Cai, et al. [6] a framework for human action recognition based on Procrustes analysis and Fisher vector encoding is proposed. A pose-based feature extracted from silhouette image by employing Procrustes analysis and local preserving projection was applied. The discriminative shape information and local manifold structure of human pose are preserved and are invariant to translation, rotation, and scaling. After the pose feature is extracted, a recognition framework based on Fisher vector

encoding and a multi-class supporting vector machine is employed for classifying the human action. Experimental results on benchmarks demonstrate the effectiveness of the proposed method.

3 Implementation

3.1 Dataset

UCF 50 is an activity or action recognition dataset. It has 50 action categories that are taken from YouTube consisting of realistic videos. The videos are grouped into 25 groups, for all the 50 categories.

50 action categories in the UCF-50 dataset are as follows - Baseball Pitch, Basketball Shooting, Bench Press, Biking, Biking, Billiards Shot, Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Lunges, Military Parade, Mixing Batter, Nunchucks, Playing Piano, Punch, Push-Ups, Pizza Tossing, Pommel Horse, Pole Vault, Pull-Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skate Boarding, Skiing, Skijet, Soccer Juggling, Swing, Tennis Swing, Trampoline Jumping, Playing Tabla, TaiChi, Playing Violin, Volleyball Spiking, Walking with a dog, and Yo-Yo.

Fig.3. UCF-50 dataset

3.2 Convolution Neural Networks

Convolution Neural Network [7] is a type of neural network, widely used for image processing. It was inspired by the working of the visual cortex region in the brain. A CNN has three layers - the input layer, hidden layer, and output layer. The two major parts of a convolution neural network are feature extraction and classification. Convolutional layers, pooling layers, and fully connected layers are present in the hidden layer.

Fig.6. shows the working of convolution neural network The input of the CNN model will be an image. A feature map is obtained by applying several different filters to the input image. Pooling will be applied to the obtained feature map which will then be flattened into a single long continuous linear vector. The single long continuous linear vector obtained at the end of flattening will be given as an input to a fully connected artificial neural network. The features are processed through the network. The model is trained through forward and backward propagation for many epochs. The fully connected layer produces the output.

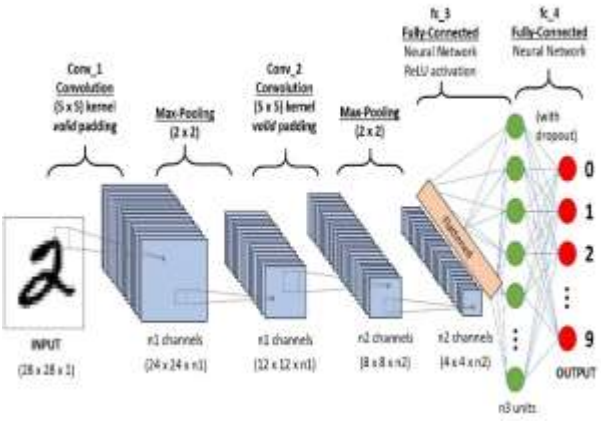

Fig.4. Working of Convolution neural network

3.3 3D Convolution Neural Network

In a 3D Convolution Neural Network, the 3-dimensional filters are applied during the 3D convolutions. The filer moves along 3 directions i.e the x,y, and z directions, and the low-level feature representations are being calculated. The shape of the output obtained from a 3D Convolution Neural Network is a three-dimensional volume space. To capture the motion-related information in human activities, the feature maps present in the convolution layer are linked to multiple frames arranged contiguously in the previous layer.

A 3D filter is convolved to obtain the 3D convolution. The stacking of multiple contiguous frames will be performed to produce a 3D cube. 3D CNN models can learn the local spatial filters which are quite useful in the classification tasks. The 3D CNN consists of two convolution layers. These are interspersed with two max-pooling layers, followed by two fully connected layers. It consists of two kernels, having the size 3x3x16. Here, 3x3 refers to the spatial dimension and 16 refers to the spectral dimension. The two kernels are used to convolve the input of the first convolution layer and four kernels having the size 3x3x16 which were used in the second convolution layer. The activation function ReLu is used for the output of the convolution. 2x2x2 max pooling is applied to the output of the convolution layers. To prevent the model from overfitting, a dropout mechanism is used. The output of the second fully connected layer is given to the softmax function. Dropout having a probability of 0.25 is performed after the initial max-pooling operation and a probability of 0.5 is performed after the first fully connected layer. The Fig.5. shows the 3D CNN architecture.

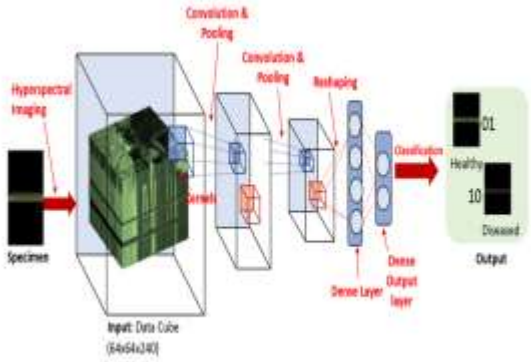

Fig.5. The architecture of 3D Convolution neural network

3.6. Proposed Model

The proposed vision-based system for human activity recognition consists of the following steps – input, data pre-processing, activity recognition, output as shown in Fig.6. The input data is received in the form input video frame which is preprocessed by resizing and normalizing the frames from the input video. A list of feature vectors and associated labels are obtained. The 3D CNN model is used for activity recognition. The model recognizes the activity in the input video. The output obtained from this algorithm can be used as input parameters for wider applications such as elderly care, anomalous behaviors detection, and surveillance system.

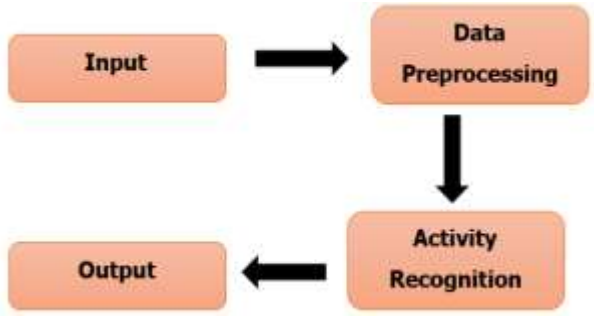

Fig.6. Scheme of the proposed model

3.7 Methodology

The input to the module is given in the form of video frames. OpenCV library is used for obtaining the input and extracting the frames. The frames will be extracted from each video while performing other pre-processing operations like resizing and normalizing images.

The module reads the video file frame by frame, resizes each frame, normalizes the resized frame, appends the normalized frame into a list, and then finally returns that list. The steps being followed in Data Pre-processing is as follows –

1. In the list of classes defined, iterate through all the classes defined in the list.
2. For each class obtained, iterate through all the video files present
3. Exact each frame in the video files
4. After all videos of the classes defined in the list are obtained and processed, select video frames randomly and extract the features and labels.
5. Store the features and labels as NumPy arrays in the pickle file.

The working of the proposed model is as shown in Fig.6.

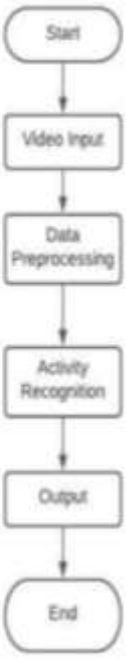

Fig.6. Working of the proposed model

The activity recognition is performed by the 3D CNN model. The model consists of two convolution layers that are three-dimensional. The activation function used is ReLu and the kernel initializer used is 'he_uniform'. The 3D max-pooling is applied. 2x2x2 pool size is used. The obtained feature maps are then flattened and given to the dense layer. The model is trained for six classes of the UCF-50 dataset – Walking With Dog, Horse Race, Jumping Jack HulaHoop, YoYo, and Pizza Tossing. Hence, the number of classes in the dense layer will be six, and the softmax activation function is used, which gives the multiclass probability distribution over all the possible target classes. The Fig.7. shows the working of 3D CNN.

Adam optimizer is used to train the 3D CNN model weights. The batch size used is 32. The 3D CNN model is trained for 150 epochs. The trained 3D CNN model is saved in the h5 format. Batch Normalization is used to stabilizing the learning process and reduce the number of training epochs.

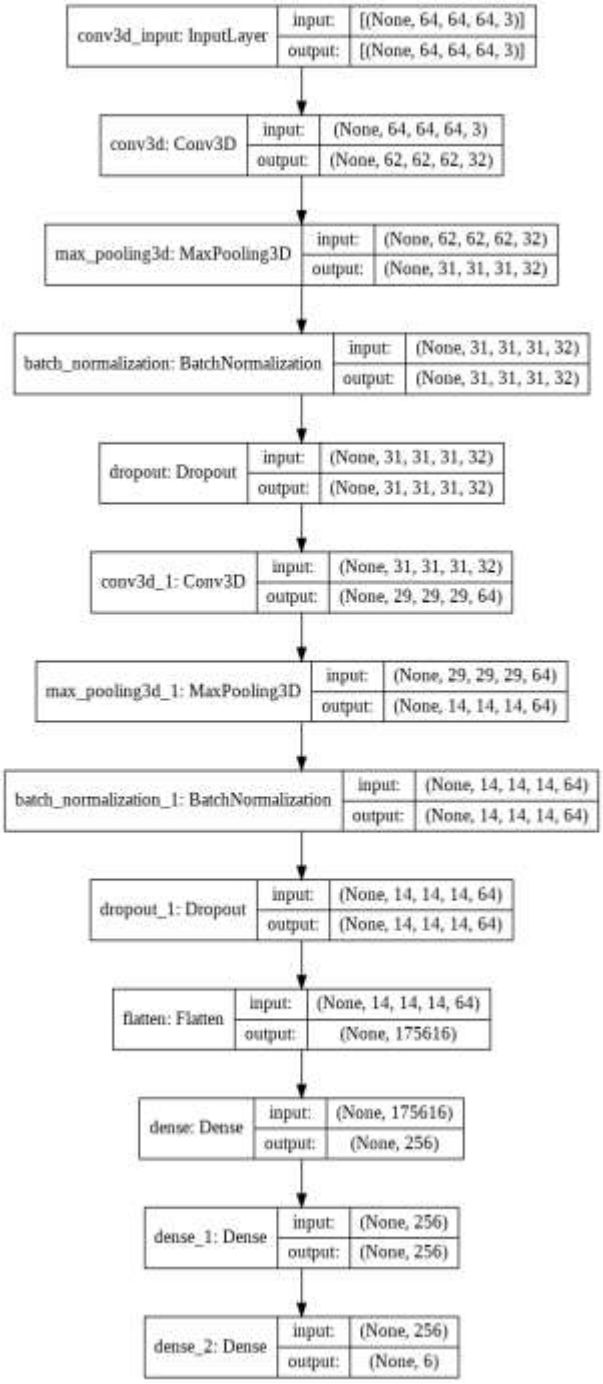

Fig.6. Working of 3D CNN

The steps that are followed for recognizing the human activity using 3D Convolution Neural Network are as follows :

1) Obtain the input
2) Extracting the frames from the video
3) Image pre-processing and normalization
    a. Converting BGR to RGB
    b. Resizing and Normalization
    c. Extraction of features and class Labels
    d. Converting and saving into Pickle files

4) Building the 3D CNN model
5) Compiling and training the 3D CNN model
6) Testing the trained model
7) Evaluating the accuracy
8) Saving the inference model
9) Obtain the output

4. Experimental Results

The model is trained for 150 epochs having 2000 images per action class.

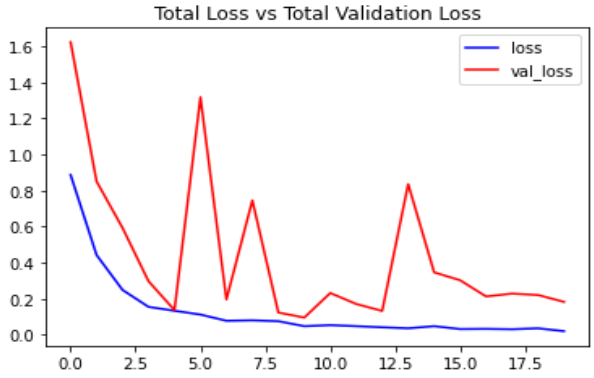

Fig.7. Total Loss vs Total Validation Loss

The Fig.7. and Fig..8. shows the graph obtained during the model training. The graph in Fig.7. represents the Total Loss vs Total Validation Loss and the graph in Fig.8. show the Total Accuracy vs Total Validation Accuracy

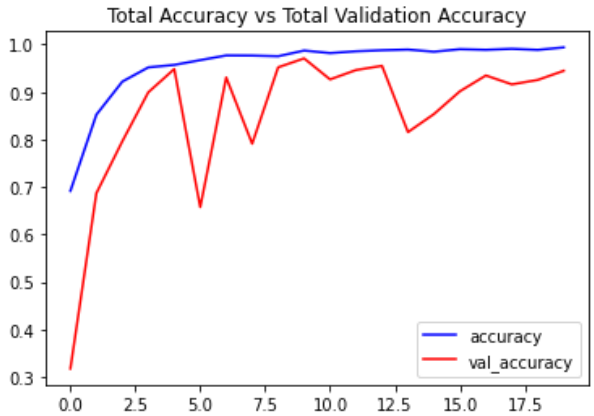

Fig.8. Total Accuracy vs Total Validation Accuracy

The accuracy score of the model is 95.33% as shown in Fig.9.

```
from sklearn.metrics import accuracy_score
score = accuracy_score(labels_test, labels_pred)
print(score*100)

95.33333333333334
```

Fig.9. Accuracy score of the model

The loss function used is categorical_crossentropy. The Fig.10. shows the classification report for the human activity recognition.

Classification Report

	precision	recall	f1-score	support
WalkingWithDog	0.96	0.98	0.97	421
HorseRace	0.99	1.00	0.99	401
JumpingJack	1.00	0.77	0.87	372
HulaHoop	0.86	0.99	0.92	406
YoYo	0.98	1.00	0.99	385
PizzaTossing	0.97	0.97	0.97	415
accuracy			0.95	2400
macro avg	0.96	0.95	0.95	2400
weighted avg	0.96	0.95	0.95	2400

Fig.10. Classification Report

The confusion matrix obtained is as shown in Fig.11. The values along the diagonal of the 6x6 confusion matrix depict correct predictions made. The values on either side of the confusion matrix depict the incorrect predictions.

```
confusion

array([[414,   4,   0,   1,   1,   1],
       [  1, 400,   0,   0,   0,   0],
       [ 12,   1, 285,  63,   1,  10],
       [  3,   0,   0, 401,   0,   2],
       [  0,   0,   0,   0, 385,   0],
       [  1,   1,   0,   3,   7, 403]])
```

Fig.11. Confusion matrix

The Fig.12. shows the model prediction for the video of the Horse race and Fig.13. shows the model prediction for the video of a person walking a dog

Fig.12. Model Prediction for Horse Race

Fig.13. Model Prediction for Walking with a Dog

5. Conclusion

Vision-based methods take the advantage of using a camera to provide more accurate and adequate data than sensor-based methods. It mainly aims to design a real classifier for human activity recognition that can be used for recognition of human activity and overcome the disadvantages of the sensor methods. The model has an accuracy of 95.33%. The accuracy of classification is relatively high here.

It overbalances the previously developed projects in this field it proves that it is safer. The objective is to develop a system that is cost-effective and affordable. This is helpful in daily life monitoring, personal biometric signature, elderly and youth care, localization, and industry manufacturing and assisting.

6. References

[1] A.Murugeshwari, S. Tamil Selvi "Human Activity Recognition Using CNN", Journal of Xi'an University of Architecture & Technology, Volume XII, Issue V, 2020, ISSN No : 1006-7930

[2] R. Mutegeki and D. S. Han, "A CNN-LSTM Approach to Human Activity Recognition," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 362-366

[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, 2019

[4] Z. Wharton, E. Thomas, B. Debnath, and A. Behera, "A vision-based transfer learning approach for recognizing behavioral symptoms in people with dementia," in 2018, 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018

[5] P. Y. Han, K. E. Yee, and O. S. Yin, "Localized temporal representation in human action recognition," in Proceedings of International Conference on Network, Communication and Computing. ACM, 2018

[6] J. Cai, X. Tang, and R. Zhong, "Silhouettes based human action recognition by procrustes analysis and fisher vector encoding," in International Conference on Image and Video Processing, and Artificial Intelligence, vol. 10836. International Society for Optics and Photonics, 2018, p. 1083612

[7] Kunihiko Fukushima (2007) Neocognitron. Scholarpedia, 2(1):1717.

[8] Daniel Maturana and Sebastian Scherer, "VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition", 2019