International Journal of Research in Advent Technology, Vol.7, No.2, February 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Human Action Recognition based on Spatio-temporal Feature Extraction using Pre-trained CNN Model and LSTM approach

J. Jerusha[1], S. Aravind Kumar[2]
Department of Computer Science and Engineering[1, 2], Jeppiaar SRR Engineering College[1, 2]
Email: jerushajoseph13@gmail.com[1], saravind123@gmail.com[2]

Abstract- Human Action Recognition (HAR) is one of the major research areas in Image Processing in association with Machine Learning. HAR is mainly used to recognize the various actions performed by a human, based on the given input frame. Some of the techniques for performing HAR are effective, but have restrictions such as low accuracy levels and high computational complexities. In the existing system, Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) units are used to perform deep learning for classifying videos based on the activity performed. The proposal contains pre-trained Convolutional Neural Network (CNN) models recognizing actions using Transfer Learning method. The top classification layer of the pre-trained CNN model is removed and additional layers are created and trained to process a specific dataset. Various pre-trained CNN models like VGG16, InceptionV3, Resnet50, Resnet150 and Resnet152 are used to extract the visual features, based on those features the videos of the UCF101 dataset are classified.

Index Terms- Convolutional Neural Network; Human Action Recognition; LSTM approach; Transfer Learning; Video Classification.

1. INTRODUCTION

In this modern era of technology, most of our day-to-day activities are digitized and machines have become an indispensable part of life. By performing HAR, machines can recognize human actions. Image processing is an active field that has been researched upon for decades and new techniques are emerging every moment. Acquiring images from the source, analyzing the image and extracting useful information from that image is called Image Processing. Image processing is used to perform Face recognition, Iris recognition, Fingerprint authentication, Human action recognition, etc.

Machine Learning (ML) is a kind of Artificial Intelligence (AI) that is used to build models that can form its own logic by using various algorithms and solve problems. ML is classified into Supervised and Unsupervised Learning. In supervised learning, the training data is properly labeled; consisting of both input values and the corresponding output values. During the training phase, the model will create the logic to solve the problem. The model is tested is by giving new input values for which it will predict the output values. By comparing the predicted output values with the actual output values, we can find the accuracy of the model. In unsupervised learning, the model is not trained with any labeled data. The model simply classifies the input data based on the similarities and differences present among them. Generally, the accuracy of unsupervised learning models is less compared to supervised learning models.

In the proposed ML model, supervised learning is performed using pre-trained CNNs. A CNN is a type of deep neural network used in image processing domain because of its capability to process visual data accurately. As the name suggests, a deep neural network contains numerous deep layers for processing the features of the images and predicting accurate classification results. Fig.1. depicts the different layers present in a CNN model.

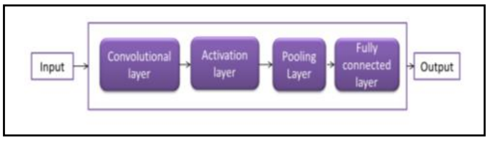

Fig.1. Different types of layers in a CNN model

A CNN consists of four types of layers: convolutional layer, activation layer, pooling layer and fully connected layer. A combination of these different kinds of layers is used to build a CNN model. Training an entire CNN from the beginning with a particular dataset is very difficult and time-consuming, so we perform Transfer Learning from various pre-trained CNN Models. Using a pre-trained CNN model to create a new CNN by removing the top classification layer and adding new layers is called transfer learning. Using this technique, CNNs process and classify the action recognition videos of UCF101 dataset with more accuracy and less computational complexities. Tensorflow is the library used for implementing the deep neural networks. Python programming language is used because of its huge popularity among ML tools and various advantages like simplicity, code readability, compatibility etc.

2. RELATED WORK

The HAR systems based on ML techniques have been implemented in numerous ways in the recent years. In this section, some of the related approaches are discussed. In [1], RNNs are used to perform action recognition by remembering the past outputs to process the forthcoming inputs. LSTM units act as the internal memory of the RNN. The Trust Gates ensure the reliability of the inputs by performing forget, remember or update actions in the internal memory. This system has been tested with seven bench-mark datasets: NTU RGB+D dataset (73.2%), UT-Kinect Dataset (95.0%), SBU Interaction Dataset (93.3%), SYSU-3D Dataset (76.5), ChaLearn Gesture Dataset (92.0%), MSR Action3D Dataset (94.8%) and Berkeley MHAD Dataset (100%). But RNNs process the inputs recurrently to provide accurate results which will consume lots of time and resources. The computational resources required for training an RNN is very high and not scalable. In [2], Deep ConvNets are utilized for extracting the spatio-temporal features of the videos. The VGGNet extracts the spatial features and Two-stream ConvNet extracts the temporal features of the frames followed by pooling strategies. An accuracy of 92.08% is achieved by this technique using the UCF101 dataset. In [3], the performance of CNNs in large-scale video classification is studied by evaluating Sports-1M dataset and UCF101 dataset. Sports-1M dataset is evaluated with Multi-resolution CNN architecture, achieving 41.4% accuracy. The transfer learning experiment of CNNs were found to achieve the highest performance of 65.4% in UCF101 dataset. In [4], CNNs are used for tracking the movement of humans in the input video sequence by extracting the spatio-temporal features using CNNs. The CNN architecture combines the local and global features together. The CNN learns the object tracked and its surrounding context by processing more number of training samples to achieve robust performance. The performance was measured by calculating the Position Error and the proposed system achieved significantly low error values. In [5], ImageNet is a large-scale database consisting 3.2 million in-the-wild images with annotations based on the WordNet database structure. Forming such a scalable and diverse dataset is very difficult, so it will be a very challenging dataset for evaluation by scientists. The pre-trained CNN models VGG16, Inceptionv3, ResNet50, ResNet150, ResNet152, etc., are trained with this standard ImageNet database and are capable of detecting objects belonging to 1000 classes from a very wide variety of images. In [6], this study attempts to shed light on the various aspects associated with forming a real-time action recognition machine. Different approaches for performing action recognition and complex activity recognition are discussed. It is suggested that collaborating different methods will further increase the performance and achieve the goal of state-of-the-art performance in real-world conditions. In [7], the InceptionV3 pre-trained CNN model, introduced by Google is the upgraded version of the previous Inception versions. It won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014. Unlike the CNNs that contained convolutional layers that go deeper and deeper, this model was more complex including many activation layers. In [8], the Visual Geometry Group, University of Oxford introduced the VGG-16 model containing 16 weighted layers and was the Runner-up in ILSVRC 2014. In [9], Kaiming introduced the Residual Neural Network (ResNet) that won the ILSVRC 2015. In [10], the popular in-the-wild action recognition video dataset UCF101, introduced by the University of Central Florida consists of 101 classes of action videos collected from YouTube and is a popular and challenging dataset for evaluating action recognition systems. In [11], using a combination of CNN and LSTM network achieves an accuracy of 68.56% in CIFAR-100 dataset. In [12], a LSTM based classification model is implemented and is shown to achieve 68.26% accuracy on Stanford Background Dataset and 22.59% accuracy on SIFT Flow Dataset. Apart from all the methods discussed above, the proposed system compares the performance of various pre-trained CNN models including LSTM approach and determines the most accurate method.

3. PROPOSED SYSTEM

Human Action Recognition System classifies the action recognition videos of UCF101 dataset by analyzing the features of the extracted frames spatio-temporally. The various pre-trained CNN models like VGG16, InceptionV3, ResNet50, ResNet150, ResNet152, etc., are pre-trained with the ImageNet dataset and are capable of detecting objects belonging to 1000 classes accurately. In Fig.2, the over-all architecture of the proposed models and techniques is shown. The input video from the UCF101 dataset is converted into frames. Various pre-trained CNN models extract the visual features of the frames. These features are processed spatio-temporally by the proposed modified CNN models.

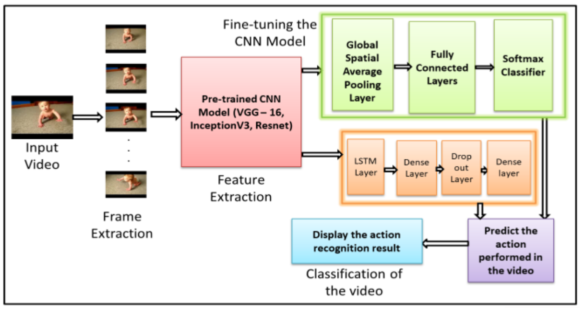
Fine-tuning the CNN Model
Global Spatial Average Pooling Layer
Fully Connected Layers
Softmax Classifier
Input Video
Frame Extraction
Pre-trained CNN Model (VGG – 16, InceptionV3, Resnet)
Feature Extraction
LSTM Layer
Dense Layer
Drop out Layer
Dense layer
Predict the action performed in the video
Display the action recognition result
Classification of the video

Fig.2. Architecture diagram of the proposed methods including the pre-trained CNN model and LSTM approach evaluated using action recognition video dataset.

To reduce the computational complexity in processing the video sequence and increase the accuracy, pre-trained CNN models are modified in two methods:

3.1. *Spatial CNN Model*

After removing the classification layer of the pre-trained CNN models, the Global Spatial Average Pooling Layer (GAP), Fully Connected Layers (FC) and Softmax classifier are added to the top of the pre-trained CNN models. These additional layers are trained with the UCF101 action recognition video dataset. After the pre-trained CNN model performs feature extraction in the frames, the GAP layer reduces the dimensions and down-samples the input features; followed by the FC layer that connects all the pooling outputs to the classifier layer. The Softmax classifier maps the pooled values to the output classes. This modified CNN model will predict the action performed in the input video by considering each frame independently. The most accurate prediction is displayed as the action recognition result by the CNN.

3.2. *Temporal LSTM Model*

In this approach, the pre-trained CNN models are modified by removing the classification layer and adding the LSTM layer and Dense Layers with a Dropout layer. Many LSTM units together constitute a LSTM layer, which becomes the storage layer for the CNN for processing the forthcoming inputs more accurately with the past memory. In Fig.3, the inputs given to the LSTM unit are processed by three gates, namely the Forget gate, Remember gate and Update gate.

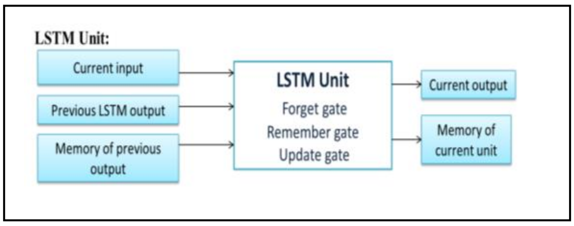

Fig.3. Structure of an LSTM unit that remembers the long-term information about the inputs and corresponding outputs of the CNN model.

These additional layers are trained with the UCF101 action recognition video dataset. The extracted features are given as inputs to the LSTM layer in sequences. The LSTM layer processes and remembers the long-term temporal information by acting as the internal memory of the CNN model. The Dense layer predicts the output by normalizing the LSTM output and mapping it to the action recognition video classes. After measuring the accuracy of the predictions, the most accurate prediction is provided as the action recognition result by the CNN model.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed models are validated by calculating the accuracy of the recognition output with regard to the actual output. Firstly, the Spatial CNN Model in III-A is evaluated using the UCF101 Dataset.

The performance metric used for calculating the accuracy of the recognition results is as follows:

$$\boxed{\text{Accuracy} = \Sigma^{Q}_{i=1} \text{TP}_i \,/\, \text{Total}}$$

Where,
Q represents the number of action classes considered.
True Positive (TP) represents the correct output results.
Total represents the number of all predictions.

Table 1. The accuracy values of Spatial CNN Model approach

Pre-trained CNN model	*Accuracy*
ResNet-50	24.2%
VGG-16	40.2%
InceptionV3	97.6%

In Table 1, the recognition accuracy values of the various pre-trained CNN models implemented using Spatial CNN Model approach are presented and the InceptionV3 model achieves the highest accuracy of 97.6%. So, it is evident that removing the classification layer and adding three new layers to the InceptionV$_3$ CNN model achieves state-of-the-art performance and accuracy higher than the VGG-16 and ResNet-50 models.

Secondly, the Temporal LSTM approach proposed in Section III-B is also evaluated using the UCF101 Dataset.

Table 2. The accuracy values of Temporal LSTM Model approach

Pre-trained CNN model	***Accuracy***
ResNet-50	68%
VGG-16	70%
InceptionV3	75%

In Table 2, the recognition accuracy values of the various pre-trained CNN models implemented using temporal LSTM model approach are presented and the InceptionV3 model achieves the highest accuracy of 75%. So, including the LSTM approach to InceptionV3 model achieves high performance in terms of accuracy compared to VGG-16 and ResNet-50 models.

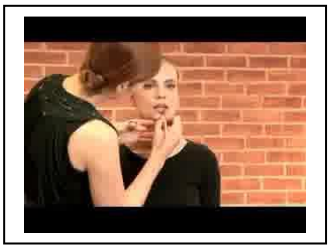

Fig.4. v_ApplyLipstick_g13_c03-0161 - Input frame from the UCF101 dataset processed using the spatial CNN model approach on pre-trained CNN models.

Considering Fig.4 as the input frame processed using the spatial CNN model approach on InceptionV3 CNN model, the model predicts the correct action recognition output for this frame with an accuracy of 97%.

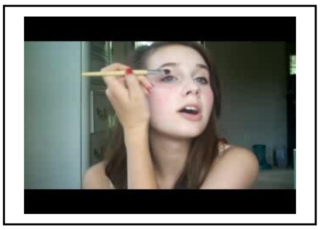

Fig.5. v_ApplyEyeMakeup_g01_c01 - Input frame from the UCF101 dataset processed using the temporal LSTM model approach on pre-trained CNN model.

Considering Fig.5 as one of the input frames from the sequence of frames processed by the temporal LSTM model approach on InceptionV3 CNN model, the model predicts the correct action recognition output for this sequence of frames with an accuracy of 75%.

5. CONCLUSION

The HAR system was implemented using various pre-trained CNN models like InceptionV3, VGG-16, ResNet50, etc., based on transfer learning method. The pre-trained CNN models were trained and tested using the UCF101 action recognition video dataset. Based on the test results, the most accurate CNN model was determined.

The future enhancements of this project include: Improvising the system to process live videos given dynamically and perform HAR. The 101 categories by which the videos are classified can be extended to accommodate more number of actions. This system can be modified to categorize the recognized actions based on behavioral characteristics. The HAR system can be enhanced to recognize the identity of the human performing any particular action in the video.

REFERENCES

[1] Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot and Gang Wang, “Skeleton-Based Action Recognition using Spatio-Temporal LSTM Network with Trust Gates”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume:40, Issue:12, pp.3007-3020, Dec. 2018.

[2] Shichao Zhao, Yanbin Liu, Yahong Han, Richang Hong, Qinghua Hu, Qi Tian, “Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition”, IEEE Transactions on Circuits and Systems for Video Technology, Volume: 28, Issue: 8, pp. 1839 - 1849, Aug. 2018.

[3] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fie-Fei, “Large-Scale Video Classification with Convolutional Neural Networks”, IEEE

Conference on Computer Vision and Pattern Recognition, June 2014.

[4] Jialue Fan, Wei Xu, Ying Wu, Yihong Gong, “Human Tracking using Convolutional Neural Networks”, IEEE Transactions on Neural Networks, Volume: 21 , Issue: 10 , pp. 1610 - 1623, Oct. 2010.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, “ImageNet: A large-scale hierarchical image database”, IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[6] Pavan Turaga, Rama Chellapa, V. S. Subrahmanian, Octavian Udrea, “Machine Recognition of Human Activities: A Survey”, IEEE Transactions on Circuits and Systems for Video Technology, Volume: 18, Issue: 11, pp. 1473-1488, Nov. 2008.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, “Rethinking the Inception Architecture for Computer Vision”, arXiv:1512.00567[cs], Dec. 2015.

[8] Karen Simonyan, Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv:1409.1556[cs], Apr. 2015.

[9] [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, arXiv:1512.03385[cs], Dec. 2015.

[10] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, “UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild”, CRCV-TR-12-01, Nov. 2012.

[11] JingLin Chen, YiLei Wang, YingJie Wu, ChaoQuan Cai, “An Ensemble of Convolutional Neural Networks for Image Classification Based on LSTM”, International Conference on Green Informatics (ICGI), Aug. 2017.

[12] Wonmin Byeon, Thomas M. Breuel, Federico Raue, Marcus Liwicki, “Scene labeling with LSTM recurrent neural networks”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.