International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Open Education Resource for School Children with Down Syndrome

Kotur Guna Pragna, Dindi Dhanunjai

Abstract*: Action recognition in video sequences is a challenging problem of computer vision due to the similarity of visual contents, changes in the viewpoint for the same actions, camera motion with action performer, scale and pose of an actor, and different illumination conditions. Also, there is no designated action recognition model for hazy videos. This model proposes a novel unified and unique model for action recognition in haze built with Convolutional Neural Network (CNN) and deep bidirectional LSTM (DB-LSTM) network. First, every frame of the hazy video is feed into the AOD-Net (All-in-One Dehazing Network). Next, deep features are extracted from every sampled dehazed frame by using VGG-16, which helps reduce the redundancy and complexity. Later, the sequential and temporal information among frame features is learnt using DB-LSTM network, where multiple layers are stacked together in both the forward and backward passes of DB-LSTM to increase its depth. The proposed unified method is capable of learning long term sequences and can process lengthy videos (even hazy videos) in real time by analyzing features for a certain time interval. Experimental results on both synthesized and natural video datasets show decent results on par with other state of the art methods in action recognition using the proposed method on the benchmark data set UCF-101. This helps the Down Syndrome Students to recognize an action faster.

***Keywords* : CNN, Haze, Bidirectional LSTM, Deep Learning**

I. INTRODUCTION

The all in one dehazing network (AOD-Net) is formed by using Convolution Neutral Network (CNN). This network helps in the removal of haze from the images. AOD-Net consists of two modules such as a K-estimation module which contains of five convolutional layers to estimate K (x) from I (x), followed by a clean image generation module to estimate J (x) from both K (x) and I (x) [8]. All the above-mentioned methods are designed for single-image dehazing, without taking into account the temporal dynamics in video. When it comes to video dehazing, a majority of existing approaches count on post processing to correct temporal inconsistencies, after applying single image dehazing algorithms frame-wise. Action recognition using deep networks is developed through 3D convolutional kernels which are applied on video frames in a time axis to capture both temporal information and spatial information. In this paper [3] the dehazing is done is a quick way by reducing the noise, no distortion of the color and maintaining the high quality.

The single image hazy removal [4] is done in this paper by using the multi-channel multi scale convolutional neural network (McMs-CNN). The dehazing of the images is performed using the TSNET for the application of the HAR for experimentally verifying by using the two benchmark datasets like UCF-101 and hazy-HMDB51 [1]. The dataset UCF-101 has the various kinds of actions taken from the YouTube videos. This method is experimentally evaluated on a new ‘‘YouTube 1 million videos dataset’’ of 487 classes. Their recognition rate on UCF101 is 63.3%, which is still too low for such important task of action recognition. Recently, for wide range of tasks like style transfer super-resolution deblurring and classification, there is a growing interest in modelling video using CNNs [2].

The usage of the RFID Technology to improve the learning of the down syndrome by using 3D objects and GUI cardboards [7]. For the teaching the down syndrome students with help of videos we the technology of VGG-16. The feature extraction from the video or the image takes place in this Visual geometry group which has the trained dataset in it. The VGG-16 helps in the fast training of the data and small in size [9]. The Recurring networks can’t remember the short-term memory so we use the Long Short-Term Memory (LSTM) network [10] to increase the long-term learning capabilities. The bidirectional LSTM performs both the backward and forward processing to the high-quality data. The LSTM is used since it has the Long-term time series and memory which is built in to store the input information. The LSTM learns the action or a sentence in one direction and reverse direction also for more understanding. The SoftMax layer takes the input from the extracted data which has the various features. The probability of each feature is given in this layer [11]. The feature with the highest probability is taken as the particular action being performed.

II. METHODLOGY

1. AOD-NET

The recognition of the image and object detection becomes very difficult in the presence of the haze. We have used AOD-NET to dehaze the video. All in one dehazing network using the lightweight CNN produces the clean image. This network performs high level tasks when embedded with deep models on the hazy image with one pipeline. AOD-NET consists of a K estimation module which has 5 convolution layers to estimate K(x) and many element wise multiplication layers and addition layers to recover the clean image.

Manuscript published on November 30, 2019.
* Correspondence Author
Kotur Guna Pragna*, Computer Science Engineering, Vellore Institute of Technology, Vellore, India.
Dindi Dhanunjai, Computer Science Engineering, Vellore Institute of Technology, Vellore, India.

The K-estimation module consists of five convolution layers, which has "concat1" layer which concatenates features from the layers "conv1" and "conv2". Similarly, "concat2" concatenates those from "conv2" and "conv3"; "concat3" concatenates those from "conv1", "conv2", "conv3", and "conv4". The need for using K-estimate module is for complete end-to-end modelling for restoring clean image. One of the important reasons to use AOD-NET is because it can be seamlessly embedded with other deep models, to constitute one pipeline that performs high-level tasks on hazy images, with an implicit dehazing process. Fig 1 shows the different types of convolution layers present in AOD-NET. To this model we give 5 sampled frames of video as input and these frames are then dehazed by the network.

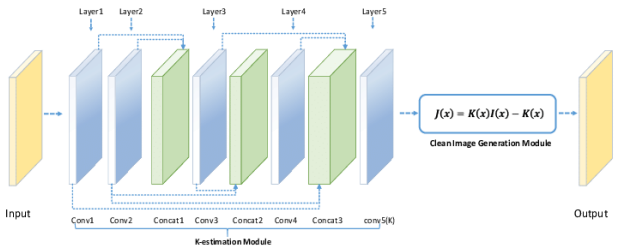

Figure 1. AOD-NET

2. VGG-16

In traditional architectures like VGG, each successive layer detects features at some more abstractly semantic level than the layer below. In VGG-16 the image is sent as the input of the convolutional layer which is of fixed size. This image is passed through the stack of convolutional layers. In this layer the Spatial and Max pooling is performed. There are various channels in the fully connected layers in the stack. Convolution happens with various filters in the VGG which are a total of 16 layers which performs the following functions. VGG-16 is better than any other neural network for feature extraction because the kernel size is less and many convolutional layers are used which gives better results compared to networks with large kernel size and less layers. These dehazed frames are then fed into the VGG-16 network for feature extraction.

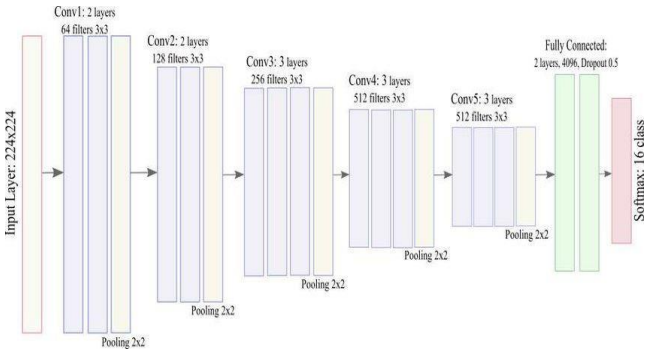

Figure 2. VGG-16

3. BI-DIRECTIONAL LSTM

The analysis of hidden patterns in the sequential way happens in the recurrent neural network using both the temporal and spatial sequential data. The RNN's major disadvantage is that it fails to derive the time steps since they are far away, as there is increase in time steps. Hence, short term memory sequences are accepted by the RNN. The solution to this problem is the Long Short-Term Memory (LSTM) networks. Using the networks of LSTM we are able to learn long term dependencies. The sequence patterns which are long term are controlled by various gates namely input, output and forget gates. The two RNNs in the Bidirectional LSTM which are in the stack format one after the another first RNN moves in the forward direction and second in the backward direction. The hidden state of the two RNNs gives the total output. In our proposed model, we have many LSTM layers for the two passes which are forward and backward. The features which are extracted such as feature vectors are sent into the network of bidirectional LSTM. The outputs are the temporal and spatial interpretations from the network. These outputs are sent into the softmax layer as the inputs. This softmax layer produces the output as the probabilities for every class. The probability which is high is taken as the predicted action in the video. In the training process, the video is taken as frames and then features are extracted. The common features are taken each frame in the video, these features are stacked up and saved into a npy file. Therefore, the features from the videos are present in the npy files.

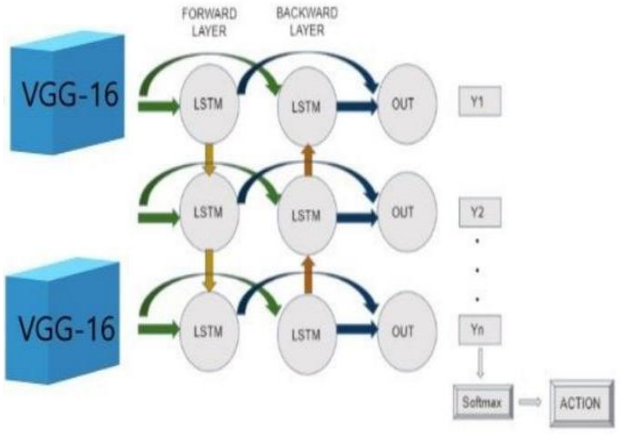

Figure 3. Bi-Directional LSTM

4. DATA SET

The huge dataset which consists of the human actions is the UCF101. This dataset contains of over 13k clips which has the 101 action classes and video data in 24 hours. The video uploaded by user are present in the database having the camera motion and proper background. This dataset also has the additional feature which is results of action recognition from the new dataset using the method of standard bag of words with best overall performance of about 44%. UCF101 is the most challenging dataset which consists of different actions which has the huge of classes and clips. The nature of the clips in this dataset are limited. Due to the lack of hazy video dataset available, we create our own synthesized hazy video dataset by (1), using the ground-truth images with depth meta-data from the UCF101 dataset. For every second frame in the clear video, we calculate the depth map by taking it as a stereoimage based on the assumption that there won't be any significant difference between two continuous frames as we are running them on code snippets.

This is anyway done only during the training phase and not during prediction phase and hence there is no problem with the above assumption. Now that we have calculated the depth maps, the transmission map can be obtained by using (2). Next, I(x) is calculated by using (1). The I(x) calculated is our hazy image. The above process is repeated for each frame of the clear video and finally, a hazy video is obtained. Therefore, we finally generate a synthetic hazy video dataset consisting of 98 videos each for the 101 classes in the UCF101 dataset.

Figure 4: UCF-101 Dataset

III. MATH

All the work carried in the image dehazing focused on the classical atmospheric scattering model:

$I(x) = J(x)t(x) + A(1 - t(x))$ (1)

where I(x) is observed hazy image, J (x) is the scene radiance ("clean image") to be recovered. A denotes the global atmospheric light, and t(x) is the transmission matrix.

$t(x) = e^{-\beta d(x)}$ (2)

Later many CNN based methods employ CNN as a tool to

$$J(x) = \frac{1}{t(x)} I(x) - A\frac{1}{t(x)} + A. \qquad (3)$$

regress t(x) from I (x). With *A* estimated using some other empirical methods, they are then able to estimate J(x) by (3).

Lately AOD-NET(All-in-One Dehazing Network) has a complete end-to-end CNN dehazing model based on re-formulating (1), which directly generates J(x) from I(x) without any other intermediate step:

$$J(x) = K(x)\,I(x) - K(x)$$

$$K(x) = \frac{\frac{1}{t(x)}(I(x) - A) + A}{I(x) - 1}.$$

IV. RESULTS AND DISCUSSION

1. Features Extraction:

All features are extracted similarly shown in Figure 5.

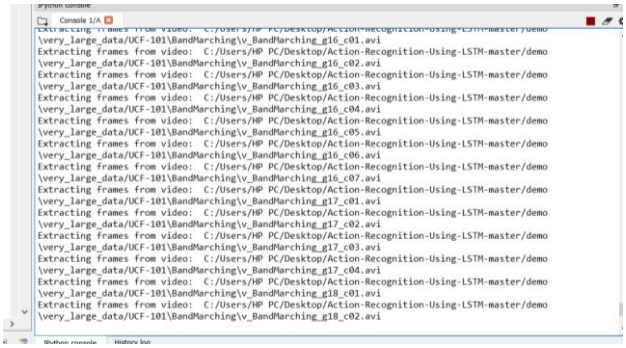

Figure 5. Frames extraction

2. Training:

Extracted frames are been trained to the model as shown

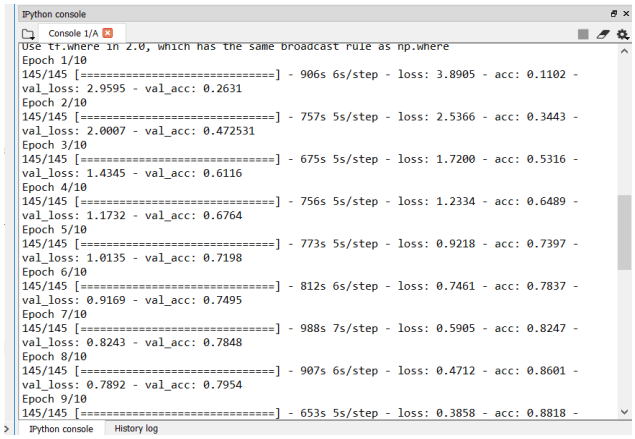

Figure 6. Training of Frames

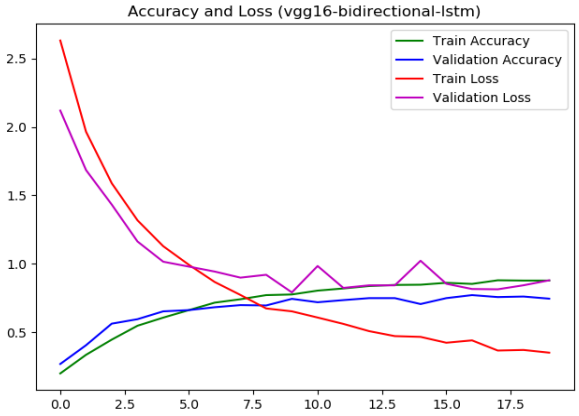

Figure 7. Accuracy and Loss Graph

3. Video Prediction:

Action in video is been predicted is been Predicted by calculating probability and conclude as shown in Figure 8.

Figure 8. Action Prdiction

V. JUSTIFICATION OF ALGORITHM

we use AOD-NET because it contains five Convolution layers in the K estimation module to estimate K(x). It Dehazes (clean) the image. VGG-16 is used because every preceding layer extract feature at a more abstractly Semantic level. It is suggestible than other neural networks for feature detection since the size of the kernel is small and many convolutional layers which leads to a good result. RNNs can remember only short-term memory sequences whereas LSTM (Long Short-Term Memory) is useful to learn long term sequences. Bidirectional Long Short-Term Memory contains 2 RNNs stacked one on one, where one goes in the forward direction and other in the backward direction. we used the UCF-101 data set for training because it provides different and efficient data to the model.

VI. CONCLUSION

In the paper, we proposed an action recognition in haze framework which first dehazes the sampled frames, then learns the features and is then fed into the DB-LSTM for classification of action. The proposed model is the first of its kind with unified end-to-end modelling. After CNN features are extracted from the dehazed video frames, they are fed into DB-LSTM, where two layers are stacked on both forward and backward pass of the LSTM. This helped in recognizing complex frame to frame hidden sequential patterns in the features. The experimental results indicate that the recognition score of the proposed method gives exceptional results on UCF-101 datasets. These characteristics make our proposed method more suitable for processing of visual data and can be an integral component of smart systems. The proposed method extracts feature from the whole frame of the video. In future, we aim to analyses only the salient regions of the frames for action recognition. Furthermore, we have intention to extend this work for activity recognition in Large Duration videos. This prototype helps the children who are suffering with Down Syndrome.

REFERENCES

1. S. Chaudhary and S. Murala, "TSNet: Deep Network for Human Action Recognition in Hazy Videos," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 2018, pp. 3981-3986.
2. S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 91-99, 2017.
3. J. Tang, Ketan, Yang Jianchao, "Investigating haze-relevant features in a learning framework for image dehazing", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2995-3000, 2014.
4. A Dudhane, S. Murala, "C2Msnet: A Novel Approach for Single Image Haze Removal", Winter Conference on Applications of Computer Vision (WACV), pp. 1397-1404, 2018.
5. . K. Soomro, A. R. Zamir, M. Shah, "UCFI0l: A Dataset of 101 Human Actions Classes From Videos in The Wild", arXiv Prepr. arXiv, 2012.
6. M.-H. Y. W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, "Single image dehazing via multi-scale convolutional neural networks", European Conference on Computer Vision (ECCV), pp. 154-169, 2016.
7. J. Jadán-Guerreroa, L. A. Guerrerob and T. Sharmac, "Improving the interaction of Down syndrome students through the use of RFID technology," 2016 IEEE MTT-S Latin America Microwave Conference (LAMC), Puerto Vallarta, 2016, pp. 1-4.
8. B. Li, X. Peng, Z. Wang, J. Xu and D. Feng, "AOD-Net: All-in-One Dehazing Network," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 4780-4788.
9. H. Qassim, A. Verma and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2018, pp. 169-175.
10. J. Li and Y. Shen, "Image describing based on bidirectional LSTM and improved sequence sampling," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(, Beijing, 2017, pp. 735-739.
11. B. Yuan, "Efficient hardware architecture of softmax layer in deep neural network," 2016 29th IEEE International System-on-Chip Conference (SOCC), Seattle, WA, 2016, pp. 323-326.

AUTHORS PROFILE

Kotur Guna Pragna graduated with B.Eng. degree in Computer Science Engineering, majoring in Data Science from Vellore Institute of Technology, Vellore in year 2016-2020.

Dindi Dhanunjai graduated with B.Eng. degree in Computer Science Engineering, majoring in Data Science from Vellore Institute of Technology, Vellore in year 2016-2020.