Bi-LSTM Model to Recognize Human Activities in UAV Videos using Inflated I3D-ConvNet

Sireesha Gundu[1], Dr.Hussain Syed[2]
Research Scholar[1], Associate Professor[2]
School of Computer, Science and Engineering, VIT-AP University
Andhra Pradesh[1,2]

***Abstract*—Human activity recognition in aerial videos is an emerging research area. In this paper, an Inflated I3D-ConvNet (Inflated I3D) and Bidirectional Long Short-Term Memory (Bi-LSTM) based human action recognition model in UAV videos have been proposed. The initial module was pre-trained using the Kinetics-400 video dataset, which consisted of 400 classes of human activities and around 400 video clips for each class culled from real-world and arduous YouTube videos. The proposed inflated I3D-ConvNet which was built on 2D-ConvNet inflation learns and extracts spatio-temporal features from aerial video while leveraging the architectural design of Inception-V1. The proposed model employs Bi-LSTM architecture for human action classification on the Drone-Action dataset which is a smaller benchmark UAV-captured video dataset. This model considerably improves the state-of-the-art results in activity classification using the SoftMax classifier and retains an accuracy of about 98.4%.**

Keywords*—*2D-ConvNet; Bi-LSTM; drone-action; inception-V1; inflated I3D-ConvNet; Kinetics-400

I. Introduction

Drone technology has advanced rapidly in recent decades. Unmanned aerial vehicles (UAVs), or drones, are proving very useful in areas that humans cannot reach or cannot reach quickly and effectively. Drone technology has much greater potential, scope, and size when used in a variety of settings, including businesses, governmental institutions, military installations, etc. They take little effort, energy, or time and can travel to even the most remote locations where fewer people are needed.

In the next generation of drones, the key characteristics of size, autonomy play, and propulsion are crucial. The type of drone is determined by the technologies used to operate it. Multirotor is the most prevalent type of drone extensively used by most of the professionals. Video surveillance, aerial photography, etc. are some of the multirotor drone applications. Multirotor is easy to manufacture and most economic to fly. With a multirotor, you can position and frame the camera more precisely for good aerial shots. The most popular and commonly used multirotor aircraft are quad-copters. Multirotors have many drawbacks such as limited speed, reduced endurance, and less flying time. With light cargo, the maximum flight time for a multirotor is 20 to 30 minutes. Due to this, multirotors are not appropriate for applications involving long-distance travel, inspection, or mapping.

Understanding how a human body is articulated in an image and identifying human movement in videos taken by these types of drones [1] is a challenging research problem. Sometimes in many situations such as those including visual blur, perspective distortion, low-resolution scenes, occlusions, etc., it is highly challenging to discern human motions.

With the swift advancement of deep learning algorithms and neural network architectures, numerous great accomplishments have been made in image recognition tasks on various datasets [2], [3], [4]. The CNN model pre-trained on ImageNet [5] has already achieved significant performance in image recognition. The same methods can be adopted in the existing action recognition mainstream models for extracting temporal and spatial features from frames of RGB video and optical flow subsequently. Although great video-based action recognition performance can be attained, our research work discovered that pre-train on still image dataset ImageNet is not an optimal preference. Since still images differ from video sequences to a significant extent, the novel dataset known as Kinetics for human action recognition [6] was released. This dataset is appropriate for pre-training the architectures for video recognition. The outcome of the experiment [7] shows that every convolutional neural network architecture pre-trained on this dataset outperforms the architectures pre-trained on ImageNet. In action recognition, an important factor along with image recognition is sequence modeling. In sequence modeling problems, compared with RNN (Recurrent Neural Network) and unidirectional LSTM [8], bidirectional LSTM [9], [10], [11] is broadly used architecture because it can address the most resilient problems such as vanishing gradient and exploding gradient up to some extent.

The majority of existing techniques suffer from a high false alarm rate. Furthermore, while these techniques perform well on simple datasets, their performance is restricted when dealing with real-world scenarios. To address these issues, we present in this paper a robust and efficient model that learns visual features from a sequence of frames by integrating them as spatiotemporal information from raw video. The following are the key contributions of current work:

- For drone-based outdoor surveillance networks, we propose a reliable ConvNet-based sophisticated paradigm. For deep spatiotemporal feature extraction, we use a pre-trained inflated inception-V1 architecture, followed by a sequential learning method that outperforms state-of-the-art approaches in terms of accuracy.
- The proposed framework employs a multi-layer Bi-LSTM architecture. Because of the forward and backward passes applied at each layer of the LSTM model, this improves the capabilities of effective learning. As

a result, the final trained model is not limited to the data used for training, but it can also be used in video surveillance networks.

- In this paper, we proposed a novel network that combines the advantage of a sophisticated ConvNet feature extractor pre-trained on Kinetics and robust sequence modeling tools such as Bidirectional LSTM. Initially, like I3D Inception 3D-CNN [12], [13], [14] pre-trained on Kinetics is exercised as the feature extractor. Next, to capture high-level temporal features, the feature vectors created by I3D are input into a bidirectional LSTM network. Finally, to create useful predictions of these high-level features, a SoftMax classifier is used. We can train the network on an end-to-end basis.

The rest of the paper is structured into sections as follows: Section II presents the overview of the existing models for action recognition, Section III gives the problem definition, Section IV describes the proposed methodology, Section V reports the datasets adopted for the proposed work, experimental results, performance evaluation, comparison with the state-of-the-art models and discussion. Section VI concludes the proposed work. Section VII provides the future work. Finally, Section VII gives the acknowledgments.

II. Literature Survey

A. 3D-ConvNet for spatio-temporal feature extraction

Since 3D-CNN was first introduced for action recognition tasks, it has always been a prevalent research method. Using this method, spatial and temporal data are directly extracted from raw surveillance video. C3D [15] is one more predominant research effort. Using a simple linear classifier on four main-stream benchmarks, C3D can outperform other algorithms by learning spatial-temporal properties from 3D-CNN. Recently, large-size datasets such as Kinetics and ActivityNet [16] have been introduced. After pre-training on a large video dataset, a few other works incorporating a 3D-CNN model outperform leading-edge models. Before 3D-CNN came into existence, due to its ability to directly extract spatial characteristics using 2D convolution, which is more effective and exact than creating features by hand, 2D-CNN models totally rule the image recognition domain.

The previous research study says that for the action recognition task, very deep 2D-CNN models can be outperformed by very shallow and Kinetics pre-trained 3D-CNN models to a great extent. A novel model called inflated I3D-ConvNet has been proposed that enhances kernels of inception module [17] such as convolution and pooling in GoogleNet [18] into 3D. In the UCF-101 dataset [19], this model has achieved maximum accuracy.

B. Bi-LSTM for Action Recognition

In sequential modeling tasks like human action recognition in videos, RNN is a very powerful and widely used network architecture. LSTM (Long Short-Term Memory) is an RNN-based network that is extensively used in video-based action recognition for learning motion features. This also can be used to prevent gradient explosion problems and gradient vanishing problems up to some extent during the training process. Another research work [20] has suggested a video as an ordered sequence based on LSTM architecture. In this, the underlying CNN network generates the output features which are fed to the LSTM network yielding an exceptional performance in the UCF-101 dataset. Another model LRCN [21] is very broad in both spatial & temporal dimensions. In this, we utilized CNN for spatial feature learning and LSTM for temporal feature learning. Also, the LRCN model has different input and output lengths.

Bidirectional RNNs are initially addressed in [22], later for speech recognition tasks, the work [23] has been proposed by using the bidirectional RNN to outperform the unidirectional RNN. Next, bidirectional LSTMs were introduced to predict frame-wise phoneme classification [24], network-wide traffic speed [25], etc. In terms of prediction bidirectional LSTMs give good results than unidirectional LSTMs. In various tasks involving videos like object segmentation in video [26], video-super resolution [27], fine-grained action detection [28] and spatio-temporal feature learning for gesture recognition [29] bidirectional LSTM architecture has been leveraged. Another research work [30] extends the idea of coupling the convolutional module with the RNN module. This architecture leverages the element-wise max pooling and BiConvLSTM and includes the temporal encoding in forward temporal directions and backward temporal directions as well because in most heterogeneous datasets accessing the forthcoming details from the present state is very much advantageous for accurate predictions.

CNNs are capable of extracting local features whereas Bi-LSTMs are very good at handling long-term dependencies. Unlike LSTMs (which utilize only past information), when complete time-series sequence data is at hand Bi-LSTMs make use of both past and future information which permits the network to make more accurate predictions. In sequential data, to capture the different temporal local dependencies, this model employs different sizes of convolutional kernels. Datasets such as WISDM, UCI-HAR, and PAMAP2 are adopted to calculate the performance of the proposed multiple human activities identification model. Merits and demerits of the existing research works are described in the below Table I.

It is possible to reach a conclusion by utilising some of the constraints and potential improvements learned from previous research. Existing approaches were insufficiently effective. This is primarily due to the algorithm's complexity and potential. Our proposed approach is used to fill gaps in the existing scientific literature. For this, we introduced a novel model named I3D-BiLSTM that employs inflated I3D-CNN pre-trained on the Kinetics dataset and bidirectional LSTM architectures for learning spatio-temporal features of adjacent video frames. This approach achieved considerably highest performance in terms of accuracy compared to existing works.

III. Problem Definition

Human activity recognition in a fixed camera-based surveillance system is relatively one of the well-studied areas of interest whereas human activity recognition in UAV-based surveillance systems is comparatively less studied [31], [32], [33]. This area of research has drawn a great deal of attention

TABLE I. MERITS AND DEMERITS OF RELATED WORKS

Method	Dataset	Merits	Demerits
C3D [15]	UCF-101	Effective approach for spatiotemporal feature learning using deep 3D ConvNets	Achieved less accuracy
ActivityNet [16]	ActivityNet	Provides largescale video benchmark for human activity understanding	Needs more categories and samples per category to unveil new challenges in understanding and recognizing human activities in present scenarios
Inception-v3 [17]	ILSVR 2012	Enhances the kernels of inception module such as convolutional and pooling in GoogleNet into 3D	Lack of verification of ensemble result on the test set as the test and validation set error tends to correlate very well
GoogleNet [18]	ILSVRC14	Improved utilization of the computing resources inside the network	Improvement needed in classification of groudtruth predictions
UCF-101 [19]	UCF-101	Large benchmark dataset consists of unconstrained videos	Less number of classes and categories, below baseline action recognition results
CNN+LSTM [20]	Sports-1M+ UCF-101	Yeilds good performance using CNN and LSTM for video classification	Deeper integration of the temporal sequence information into the CNNs is required
LRCN [21]	Flickr30k+ COCO2014	Achieved good results in CNN special learning and LSTM temporal feature learning with different input and output lengths	Needs more investigation on perceptual problems with time-varying visual input or sequential outputs
BRNN [22]	TIMIT	Efficient estimation of the conditional posterior probability of complete symbol sequences without making any explicit assumption about the shape of the distribution	Need more investigation on classification
DBLSTM-HMM hybrid [23]	TIMIT speech	Outperforms both RNNs and unidirectional LSTM	Experiments in LSTM learning algorithms, output error functions, and improved generalisation is required
Frame-wise Phoneme classification [24]	TIMIT speech	Outperforms both RNNs and unidirectional LSTM	Experiments in LSTM learning algorithms, output error functions, and improved generalisation is required
SBU-LSTM [25]	loop detector data+INRIX data	Best prediction performance for the traffic speed on different types of traffic network in both accuracy and robustness	Requires more attention in learning and interpreting spatial features towards graph-based structure
ConvGRU+ConvLSTM [26]	DAVIS+ FBMS+ SegTrack-v2+ FT3D	Improved encoding techniques used in spatio-temporal evolution of objects in a video for motion segmentation	More research is required on promising algorithms like Instance-level video object segmentation
BRCN for multi-frame SR [27]	25 YUV format videos+Set1+Set2	Powerful temporal dependency modeling for SR videos containing complex motions, achieved better performance and faster speed	1) The model has a very shallow architecture (i.e., only 3 layers), so its performance is poor 2) Due to the low-resolution images the computational complexity grows with the upsampled spatial size 3) Due to the lack of large-scale video SR dataset, model cannot directly learned from scratch (raw videos)
Multi-stream CNN +Bidirectional LSTM [28]	MPII Cooking 2+new MERL Shopping	Effective use of pixel trajectories rather than stacked optical flow as input to the motion streams, leading to a significant improvement in results	Extraction of pertinent data is required to increase more accuracy
MoIWLD [29]	Hockey Fight +BEHAVE +Crowd Violence	Modified sparse-representation-based classification model is used to control the reconstruction error of coding coefficients and minimize the classification error for violence detection in videos	Requires more training to improve the accuracy
BiConvLSTM [30]	Hockey Fights+Movies+Violent Flows	Leverages the element-wise max pooling and BiConvLSTM, includes the temporal encoding in forward temporal directions and backward temporal directions for the detection of violence in videos	Further investigation is needed for the best results on more dynamic and heterogeneous datasets

in recent years as drones are becoming more popular and extremely beneficial in different applications such as military, agriculture, search and rescue, security monitoring systems, etc [34]. Human activity recognition in aerial videos is a challenging task because of various factors such as environmental background, camera motion, object distances and variety of cameras, longitude, the latitude of the UAV flight, etc. [35], [36], [37]. Another challenging task is human activity recognition in a video (sequence of frames) where each frame has different variations such as spatio-temporal features loss, background clutter, occlusion, low inter-class and high intra-class variances for some classes, background lighting change, image distortion, and a person pose variation [38], [39], [40], [41]. We need to develop a sophisticated model that addresses these issues and accurately classifies the human activities in UAV captured video datasets.

IV. PROPOSED METHODOLOGY

A. 3D-ConvNet

In the realm of computer vision, Convolutional Neural Network (CNN/ConvNet) is one of the well-known deep learning neural networks that can handle massive volumes of data in the early 1980s. It is a feed-forward neural network with multiple layers that agglomerates many convolutional layers called grouping layers, hidden layers, and activation layers on top of each other in a particular order. This sequential design allows ConvNet to learn hierarchical attributes using the multi-layers of artificial neurons. An artificial neuron or perceptron is the function that evaluates the weighted sum of one or more inputs and passes this value to the non-linear function called the activation function which produces the output as an activation value.

Initially, 2D-ConvNet was pre-trained on ImageNet for image recognition. When an image is passed as an input to the 2D-ConvNet, it generates different activation functions that are directed to the next layers. During the convolution, a kernel/filter of fixed size should be applied to the input image to obtain convolved features. The initial layer extracts primitive information such as diagonal or horizontal edges. The next layer takes hold of the output from the previous layer and extracts more intricated features such as amalgamated edges or corners. Hence, the deeper layers of the network can extract even more complex features such as faces, objects, etc. The terminal convolutional layer generates the activation map. Based on this, the confidence score (between 0 and 1) will be produced by the classification layer to classify the image. 2D-ConvNet architectures pre-trained on ImageNet and other datasets failed to recognize the objects in the images or videos which are having different angles and different illumination conditions. Later on, Carreira and Zisserman introduced a new 3D-ConvNet architecture pre-trained on Kinetics for human action recognition to extract spatio-temporal features in videos.

In this work, we proposed a novel architecture shown in Fig. 1 where each short video clip classification is treated as an action recognition problem. For this subject, 3D-CNN has traditionally been a popular strategy for facilitating spatio-temporal learning. A substantial amount of training data is often required to train a 3D-CNN from scratch. Carreira and Zisserman introduced a new model that inflates a pre-trained 2D-CNN on ImageNet architecture along the temporal dimension to produce an inflated I3D-ConvNet. They used the Kinetics human action dataset to train the inflated I3D-ConvNet to address the action recognition challenge. Based on their work, we exercised Inception-v1 I3D for our challenge and fine-tune it on our Drone-Action dataset [42]. In this paper, we have chosen the pre-trained weights of the RGB stream as the initial weights are publicly accessible for Inception-v1 I3D.

In this model architecture, we manifested the evolution of 3D-ConvNet from ImageNet-based 2D-ConvNet architecture. Like standard conventional networks, 3D-ConvNet is a natural video modeling approach with spatio-temporal filters explored previously in [43], [44], [45]. Inflating 2D-ConvNets into 3D implies converting successful image classification models such as deep 2D-ConvNets into 3D-ConvNets by inflating all the filters and pooling kernels of size ‘n × n’ into ‘n × n × n’ with a supplementary temporal dimension. One important characteristic of these models is that they precisely generate hierarchical illustrations of the spatio-temporal data. Compared to 2D-ConvNet, it has many parameters due to the added kernel dimension, which makes training very difficult but results have proven that they have produced promising results for evaluation on the larger benchmark dataset (e.g. Kinetics). Fig. 2 depicts the feature extraction process for an inflated I3D-ConvNet architecture.

We used a modest modification of C3D for this research, which contains 14 convolutional layers, 27 BatchNormalization layers, 27 ReLU layers, 13 DepthwiseConv layers, 4 ZeroPadding layers, one GlobalAveragePooling layer, and A top layer with 4 fully connected layers. As in the original version, short clips of 244 × 244 pixels each with a crop of 112 × 112 pixels make up the model’s inputs. After all convolutions and completely connected layers, we used batch normalization. Another variation of the original model is, that instead of using a temporal stride of 1 stride 2 is used. This moderates the memory usage and permits greater batch sizes. When fully connected layers without weight bindings have been constructed, batch normalization is very crucial at this stage.

B. LSTM

Long Short-Term Memory (LSTM) deep learning architecture was first developed by Hochreiter and Schmidhuber in 1997. It is a gated Recurrent Neural Network (RNN) architecture. RNN network works on the present input by considering the previous output as feedback and stores the information in its memory for future prediction. Unlike RNN, it is capable of handling long-term dependencies. LSTM networks address the following issues that RNN fails to solve:

- Preserving the information for a long period or handling long-term dependencies to predict the current output.
- Fine control over carrying forward the important information and forgetting the unnecessary past information.
- Removal of exploding and vanishing gradient problem which occurs during the network’s training phase, backtracking is used to reduce loss and achieve the best outcomes.

C. LSTM Architecture

LSTM works very similar to RNN but it has a key feature that differs it from RNN that is it preserves information for a long period for future cell processing. Three gates make up an LSTM cell: Forget gate, Input gate, and Output gate. The internal process of an LSTM cell is depicted in the following Fig. 3:

It has a memory pool that has two key vectors of states.

1) Short-term state: It is also known as a hidden state. This state keeps the output at the current time step. The short-term state of the previous timestamp is represented with S_{t-1} and the current timestamp is S_t.
2) Long-term state (L_{t-1}): It is also known as the cell state. This state reads and rejects the information while passing through the network which is meant for long-term storage. The long-term state of the previous timestamp is represented with L_{t-1} and the current timestamp is L_t.

Here, the cell state contains all timestamps and information as well. As depicted in the figure, the decision of reading, writing, and storing depends on the activation functions whose outputs are in between (0, 1).

1) Forget gate: This is the first state in the cell of the LSTM network. This gate decides whether to store the information of the previous timestamp or forget it. The Forget gate Eq. (1) is as follows:

$$F_t = \sigma(C_t * U_f + S_{t-1} * W_f) \tag{1}$$

Where,

C_t = Current timestamp t input,
U_f = Weight connected with the input,
S_{t-1} = Previous timestamp’s short-term state or hidden state,
W_f = The weight matrix for the short-term state.

Next, the activation function i.e. the $sigmoid$ function is applied to it which gives the value of f_t in between (0, 1). It is then multiplied by the previous timestamp’s long-term state, as shown in the calculations below by using Eq. (2) and (3).

$$L_{t-1} * f_t = 0, \qquad if \quad f_t = 0 \tag{2}$$

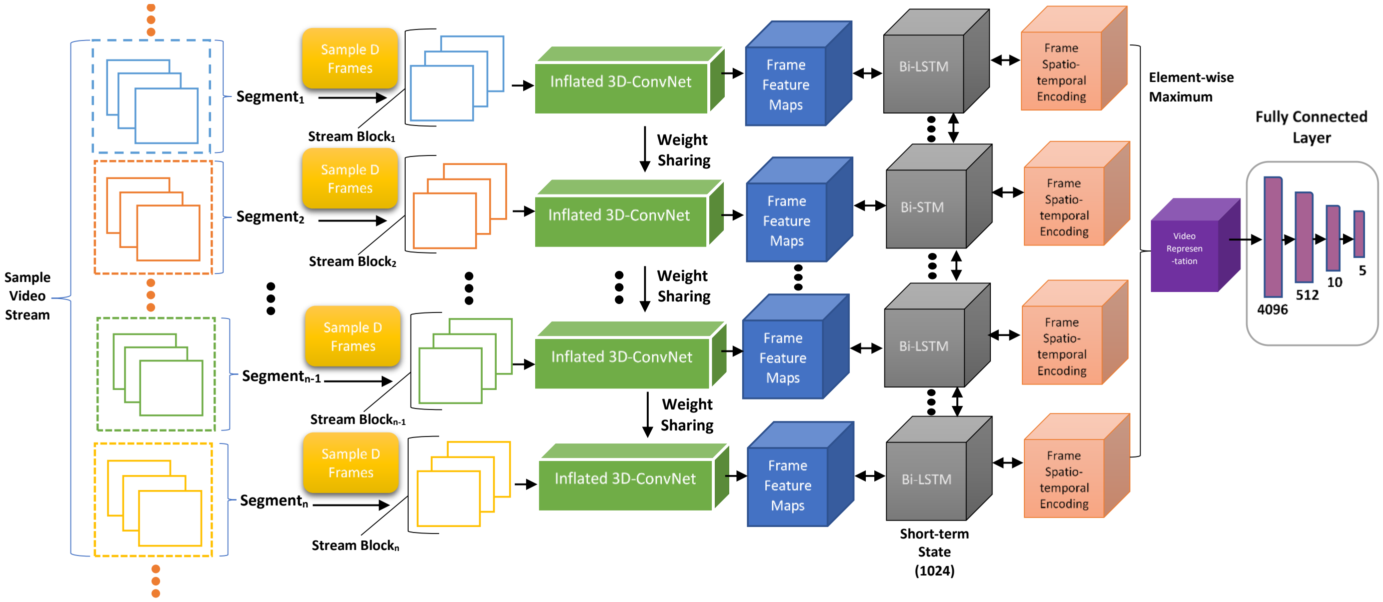


Fig. 1. Framework of the proposed Inflated I3D-ConvNet-BiLSTM for action recognition

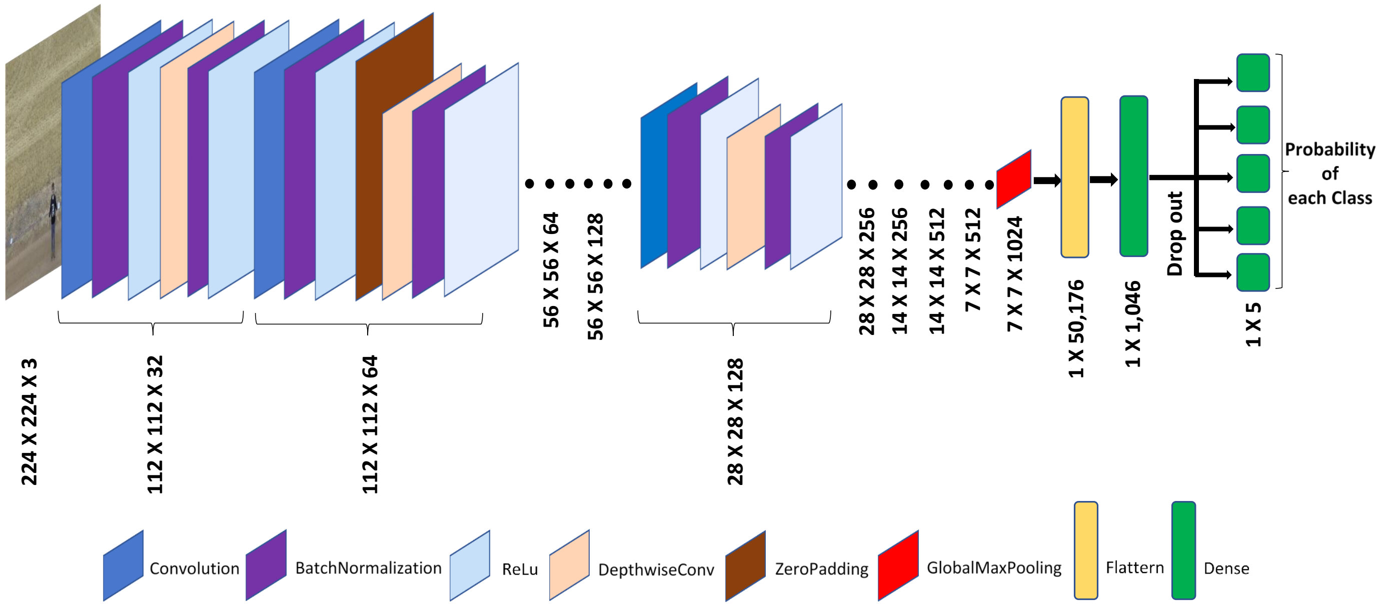


Fig. 2. The feature extraction workflow of Inflated I3D-ConvNet architecture

$$L_{t-1} * f_t = L_{t-1}, \qquad if \quad f_t = 1 \tag{3}$$

If the f_t value is 0 it forgets everything, otherwise, it forgets nothing.

2) Input gate: This is employed to manage the input value flow and quantifies the importance of the latest information in the cell. The Input gate Eq. (4) is as follows:

$$I_t = \sigma(C_t * U_i + S_{t-1} * W_i) \tag{4}$$

Where,

C_t = Current timestamp t input,
U_i = Matrix of weights associated with the input,
S_{t-1} = Previous timestamp's short-term state or hidden state,
W_i = The weight matrix for the short-term state.

Next, the activation function is then subjected to the $sigmoid$ function, resulting in the value of 'I' at timestamp 't'. The value lies between (0, 1).

In LSTM, the latest information also known as new infor-

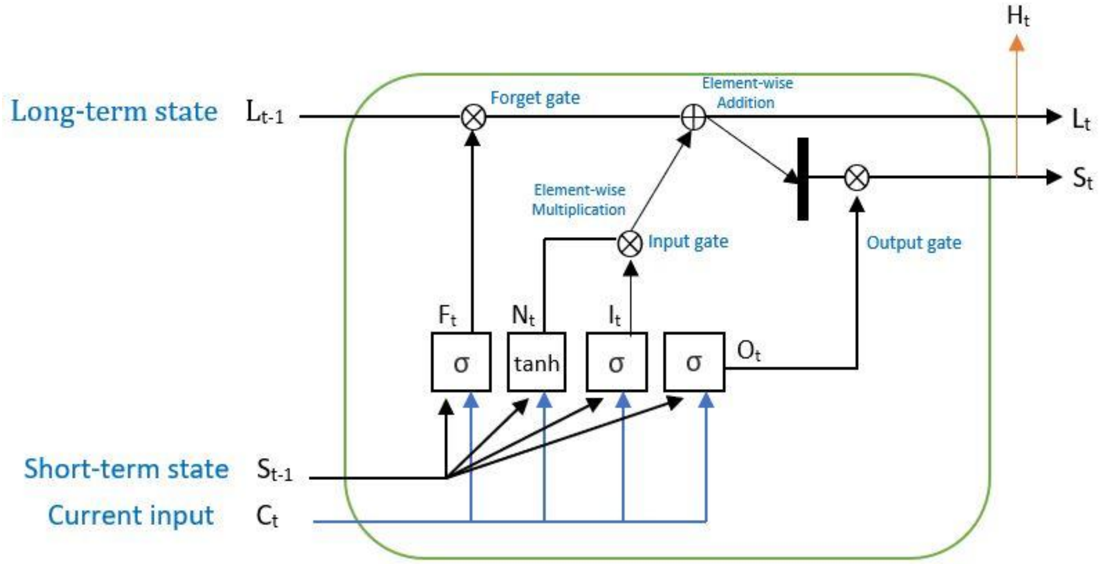


Fig. 3. Long short-term memory architecture

mation in Input Gate is used to update the long-term state.

Latest Information (or New Information):

$$N_t = tanh(C_t * U_c + S_{t-1} * W_c) \qquad (5)$$

This latest information shown in Eq. (5) is a function of the short-term state at the previous timestamp ‘t-1’ & input ‘C’ at the timestamp ‘t’. This information is necessary to pass it through the long-term state. After the activation function, $tanh$ is applied over it, therefore the value of the latest information lies between (-1, 1).

If N_t is negative, this information is subtracted from the long-term state; if N_t is negative, this information is added to the long-term state at the current timestamp. The following Eq. (6) is an updated equation to add N_t to the long-term state.

$$L_t = F_t * L_{t-1} + I_t * N_t \qquad (6)$$

Where L_{t-1} represents the long-term state at the current timestamp and others represent previously determined values.

3) Output gate: This is used to control the cell utilization for calculating the output activation of the LSTM unit and is used to determine the generation of the output from the current internal long-term state to the next short-term state. The Output gate Eq. (7) is as follows:

$$O_t = \sigma(C_t * U_o + S_{t-1} * W_o) \qquad (7)$$

This equation is similar to Forget gate and Input gate. The output value of this equation lies between 0 and 1 when the $sigmoid$ activation function is applied to it.

Next, the following Eq. (8) will be used to compute the current short-term state O_t and $tanh$ of the updated long-term state:

$$H_t = O_t * tanh(L_t) \qquad (8)$$

That is, the short-term state is the function of present output and $tanh$ of the long-term state. Then apply the SoftMax activation function to the short-term state L_t to get the output of the present timestamp using the following Eq. (9):

$$Output = SoftMax(S_t) \qquad (9)$$

In this, the token with the highest score in the output is the prediction.

D. Bidirectional LSTM

Bidirectional LSTM or Bi-LSTM is an extension of the LSTM model. Bi-LSTMs, unlike baseline LSTMs (trains a model in unidirectional i.e., forward direction and use only past information), train the model in two directions called forward direction and backward direction as depicted in the following Fig. 4.

In the forward direction, the model learns a sequence of inputs from past to future whereas in the backward direction model learns from future to past when the complete sequence of time-series data is available. This denotes that the calculation of the output frame at timestamp ‘t’ depends on the preceding frame at time ‘t-1’ and the subsequent frame at time ‘t+1’ because it performs processing in both directions.

This approach uses two hidden states, one for the forward pass and another for the backward pass, to preserve the past and future information. To enable the network to make more accurate predictions, these states should be combined and this mechanism is called merging. This can be achieved through the functions called sum, averaging, multiplication, and concatenation. Among these functions concatenation is the default mechanism.

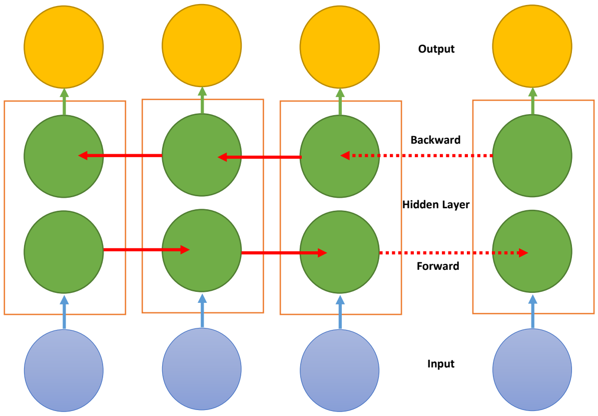


Fig. 4. Bidirectional LSTM architecture

V. Experimental Results and Discussion

First, we briefly described the datasets that are utilized to evaluate the models in this segment. Secondly, we provided the classification outcomes after implementing the Inflated I3D-ConvNet model with Bi-LSTM for recognizing human actions. In the performance evaluation section, charts of model accuracy and loss are provided. We finally reported the comparison results of existing models with the proposed model in the following section. All these experiments are performed on the anaconda platform.

A. Overview of the Datasets

1) Kinetics-400: The below Table II presents an overview of the Kinetics-400 dataset.

TABLE II. Summary of Kinetics-400 Dataset

Feature	Value
Actions	400
Clips per class	min 400
Total clips	306, 245
Repetitions per clips	3-5
Each clip length	10 Sec

In this paper, we used a human action video dataset called Kinetics-400. It contains 400 human action classes, with at least 400 video clips per action, each clip lasting 10 seconds, for a total of 306,245 videos. Each clip in this dataset is retrieved by first searching YouTube for suggestions and then using AMT (Amazon Mechanical Turkers) to determine if the clip contains an action. At least three or more out of five confirmations were needed to accept the clip. Next, we ran de-duplication on the dataset and verified that the clips were retrieved from each video and that the clips did not share common video footage. Finally, the noise reduction and overlap of the class were reviewed.

Rather than activities or events, the dataset is focused on a broad range of classes and human actions. Some of the human actions are drinking, drawing, pumping fist, laughing, etc., human-human actions such as shaking hands, hugging, kissing, etc., as well as human-object actions such as washing dishes, opening a present, mowing the lawn, etc.

Some actions in the Kinetics-400 dataset are fine-grained and require emphasis on an object to categorize, e.g., playing various kinds of wind instruments. Some other actions require temporal reasoning to categorize e.g., various kinds of swimming. There are also different Parent-Child groupings like Personal Hygiene (cutting nails, washing hands, Brushing teeth, . . .), Music (trombone, playing drums, violin, . . .), Cooking (peeling, frying, cooking, . . .), Dancing (macarena, ballet, tap, . . .), etc.

2) Drone-Action: The dataset consists of 13 actions, each action having 10-20 clips, a total of 240 clips, and 66919 frames. All videos in the dataset were delivered in HD format (1920 × 1080 resolution) and recorded at 25 fps, with an average duration of 11.15 seconds for each action. Midway through a wheat field, actions were recorded on an unsettled road while the hover and follow modes were active. The actions were divided into three categories called following, side-view, and front-view based on the camera movement and the viewpoint. The video recording camera was 10 MP, A 3DR SOLO rotor-craft served as the unmanned aerial vehicle, and the camera was a GoPro HERO 4 Black with 5.4 mm, IR CUT anti-fish eye replacement, and a 3-axis Solo gimbal. The flight ended between 8 and 12 metres low altitude and was slow. The following Table III provides an overview of the Drone-Action dataset.

TABLE III. Summary of Drone-Action dataset

Feature	Value
Actions	13
Actors	min 10
Clips	240
Clips per class	10-20
Repetitions per class	5-10
Mean clip length	11.15 Sec
Total duration	44.6 min
Frames	66919
Frame rate	25 fps
Resolution	1920 × 1080
Camera motion	Yes (hover and follow)
Annotation	Bounding box

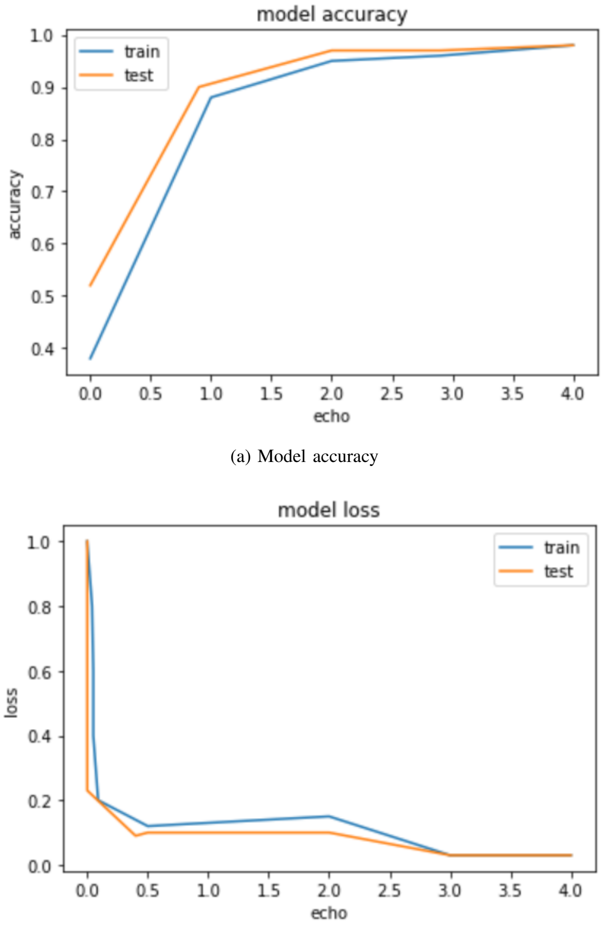
In this research work, out of 13 action classes, only 5 classes [clapping, waving hands, walking (front, back, left, right), jogging (front, back, left, right), running (front, back, left, right)] with a minimum of 20 video clips for each action class and an average duration of 11.15 seconds for each action were utilized for classification of human actions.

B. Results

Initially, to make your I3D-ConvNet network a strong feature extractor, pre-train it with the Kinetics-400 video dataset. Referring to the RGB I3D model [12] and using the open-source I3D model, we attempt to emulate the results to achieve

97.6% accuracy on the Drone-Action dataset with RGB modality. With reference to the previous work, to represent high-level temporal features, we integrated the Bi-LSTM network with I3D and used the Drone-Action dataset to validate the proposed model and attempt to obtain an accuracy of 98.4% which outperforms the other mainstream models and also the original I3D model.

Below Table IV provides instances of sequences that correlate to various types of actions and scenarios in the Drone-Action dataset and Table V shows some of the sample results with a percentage of accuracy after validating the model using the Drone-Action dataset.



(a) Model accuracy

(b) Model loss

Fig. 5. Performance of the proposed model

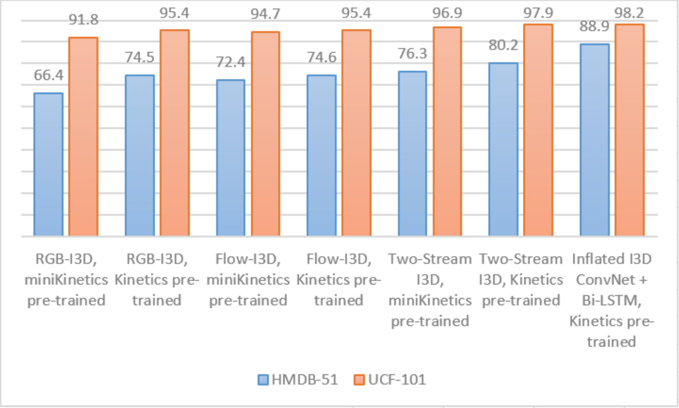


Fig. 6. Bar chart of the average accuracies of the various models

C. Performance Evaluation

The evaluation measure we have chosen for this experiment was accuracy. The action classifier returns scores based on this the accuracy was evaluated. The Drone-Action dataset was divided into train-to-test split sets. Each train-to-test split set was randomly generated at a ratio of 70:30.

The two measures which we considered to assess the proposed model's accuracy trained on Kinetics-400 are as follows:

- Firstly, freeze the network weights and utilize it to generate features for the Drone-Action video dataset, and then train the SoftMax classifier for the Drone-Action dataset classes that use train data to evaluate a test set.
- Secondly, fine-tune the network for classes of the Drone-Action dataset by utilizing its training data and re-evaluating its test sets.

This clearly says that the proposed inflated I3D-ConvNet architecture with Bi-LSTM perks from pre-training on the Kinetics-400 dataset and also after pre-training on the Kinetics dataset training only the model's last layers yield much better performance. The proposed model's model accuracy and model loss per epoch is depicted in the Fig. 5.

1) Comparison with classical models: We compare the performance of I3D models to earlier classical approaches in Table VI, on HMDB-51 & UCF-101 datasets and the bar chart of the various model's average accuracies is shown in the following Fig. 6. We provide the findings while pre-training on miniKinetics and the complete Kinetics dataset.

D. Discussion

All activities were correctly recognised with a high classification performance using the proposed approach. The bidirectional LSTM ability to use past and future inferences in the signal allows for accurate differentiation of the dynamic activities like clapping, jogging, wavinghands, running, walking.

The results of this work are comparable to the best related works using the same datasets. In contrast to earlier 3D-ConvNets (C3D) models, Kinetics' pre-trained I3D models demonstrate a far higher difference. This could be due to Kinetics higher quality, but it could also be due to I3D's superior architecture. The I3D RGB stream on HMDB-51 improves after switching from miniKinetics to Kinetics pre-training, indicating that 3D-ConvNets may permit a massive volumes of information to understand robust motion characteristics. The two streams achieve equal performance after Kinetics pre-

TABLE IV. DRONE-ACTION DATABASE: INSTANCES OF SEQUENCES CORRESPONDING TO VARIOUS TYPES OF ACTIONS AND SCENARIOS

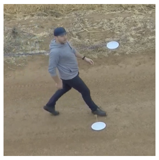
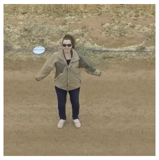
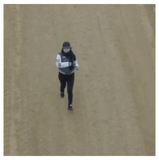
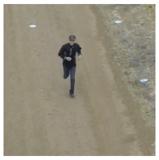
Actor	walking	jogging	running	clapping	waving hands
Actor 1					
Actor 2					
Actor 3					
Actor 4					
Actor 5					
Actor 6					
Actor 7					
Actor 8					
Actor 9					
Actor 10					

TABLE V. THE ACTION CLASSES OF THE DRONE-ACTION DATASET

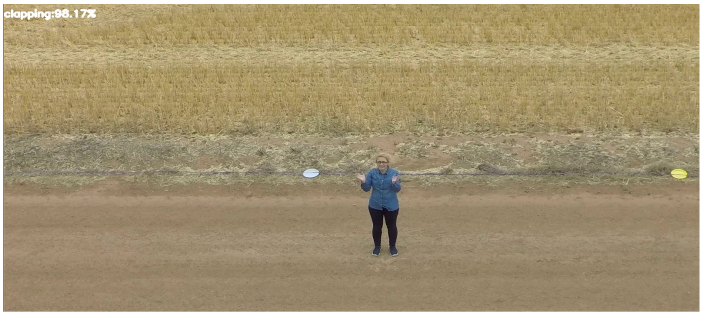
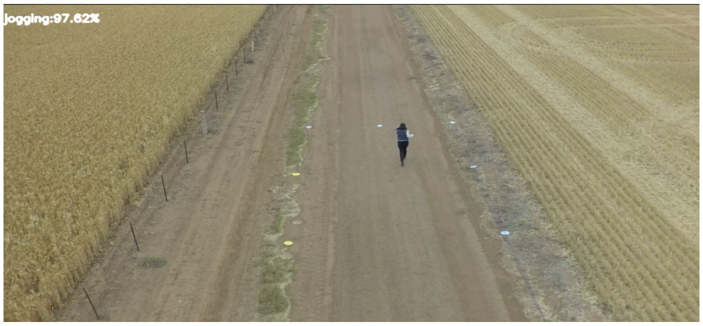
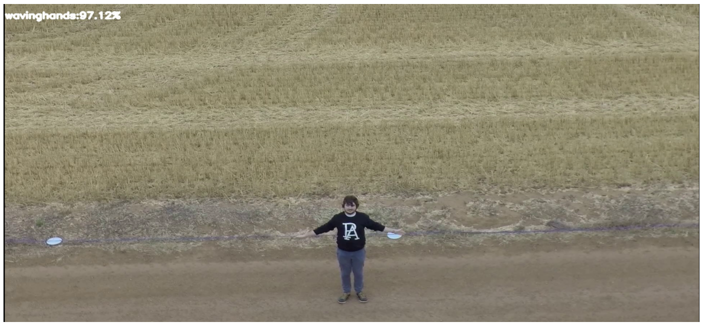
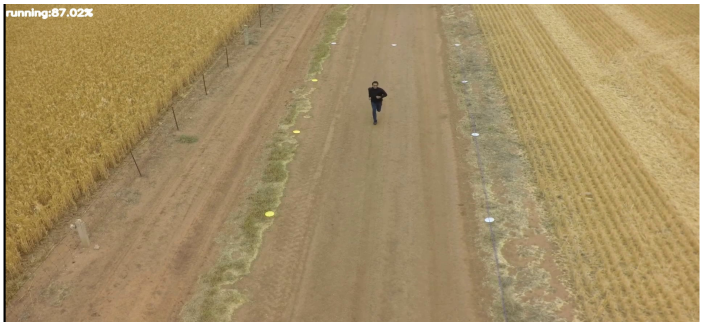
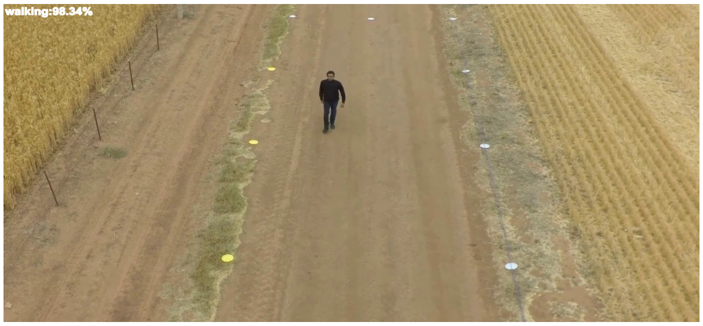
Action	Result
clapping	
jogging	
waving hands	
running	
walking	

TABLE VI. COMPARISON WITH CLASSICAL MODELS AVERAGED OVER THREE TRAIN/TEST SPLITS ON HMDB-51 & UCF-101 DATASETS

Model	**HMDB-51**	**UCF-101**
RGB-I3D, miniKinetics pre-trained	66.4%	91.8%
RGB-I3D, Kinetics pre-trained	74.5%	95.4%
Flow-I3D, miniKinetics pre-trained	72.4%	94.7%
Flow-I3D, Kinetics pre-trained	74.6%	95.4%
Two-Stream I3D, miniKinetics pre-trained	76.3%	96.9%
Two-Stream I3D, Kinetics pre-trained	80.2%	97.9%
Inflated I3D-ConvNet + Bi-LSTM, Kinetics pre-trained	88.9%	98.2%

training, but they're still good together. When their predictions are averaged, the results range from 74.6% to 80.2%.

The average accuracy over the three common train/test splits was used to compare our techniques. When pre-trained on Kinetics, a model from our RGB-I3D or RGB-Flow lineup outperforms any model combination or any previously published performance by any model. Our proposed model significantly improves on previous models, bringing overall performance to 88.9% on HMDB-51 and 98.2% on UCF-101 respectively.

VI. CONCLUSION

In this paper, we introduced a new architecture known as inflated I3D-ConvNet with Bi-LSTM pre-trained on the Kinetics-400 video dataset. First, on the Kinetics-400 dataset, we initially model the low-level features of successive frames using a pre-trained I3D model, and then we learn high-level special features using a Bi-LSTM network. The leading performance of inflated I3D-ConvNet pre-trained on the Kinetics-400 video dataset also corroborates that this novel model is more advanced and efficient than other mainstream models. The proposed model can serve as an alternative solution to the data loss and long-term dependency problems that arise as the training data size increases. This architecture uses various convolutional filter sizes by capturing different local dependencies to enhance feature extraction and also can able to classify simple activities like waving hands, walking, clapping, jogging, and running with outstanding accuracy. The suggested model uses the inflated process as a domain adaptation job for 3D-ConvNets, and the input structure of split video segments outperforms the standard sequential input structure.

VII. FUTURE WORK

In the proposed model we utilized Kinetics-400 large-scale and high-quality video dataset which includes 400 human action classes for pre-training the architecture. As a future plan, we plan to rerun all the experiments using other benchmark datasets such as Kinetics-600 and Kinetics-700 that cover 600 & 700 human action classes consequently with and without pre-training, explore inflating operation on 3D-ConvNets and also examine the efficacy of implementing our inflated I3D-ConvNets into two and three-stream architectures.

ACKNOWLEDGMENTS

We would like to extend our profound gratitude and sincere appreciation to everyone on the Kinetics dataset project, especially Tim Green and Brain Zhang for creating data that can be used for experiments and also we would like to express our profound gratitude to Asanka G.Perera, Yee Wei Law and Javaan Chahl for sharing `Drone-Action` dataset for our experiments.

REFERENCES

[1] V. Krassanakis, M. P. Da Silva, and V. Ricordel, "Monitoring human visual behavior during the observation of unmanned aerial vehicles (Uavs) videos," *Drones*, vol. 2, no. 4, pp. 1–19, dec 2018.

[2] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[3] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016. [Online]. Available: http://arxiv.org/abs/1609.08675

[4] S. Gundu, H. Syed, and J. Harikiran, "Human detection in aerial images using deep learning techniques," in *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2022, pp. 1–10.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[7] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?" *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

[8] R. L. Abduljabbar, H. Dia, and P. W. Tsai, "Unidirectional and bidirectional LSTM models for short-term traffic prediction," *Journal of Advanced Transportation*, vol. 2021, 2021.

[9] A. Mihanpour, M. J. Rashti, and S. E. Alavi, "Human action recognition in video using db-lstm and resnet," in *2020 6th International Conference on Web Research (ICWR)*. IEEE, 2020, pp. 133–138.

[10] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features," *IEEE Access*, vol. 6, pp. 1155–1166, nov 2017.

[11] S. S. Saha, S. S. Sandha, and M. Srivastava, "Deep convolutional bidirectional lstm for complex activity recognition with missing data," in *Human Activity Recognition Challenge*, 2021, vol. 199, pp. 39–53.

[12] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4724–4733, 2017.

[13] Y. Huang, Y. Guo, and C. Gao, "Efficient Parallel Inflated 3D Convolution Architecture for Action Recognition," *IEEE Access*, vol. 8, pp. 45 753–45 765, 2020.

[14] X. Wang, Z. Miao, R. Zhang, and S. Hao, "I3d-lstm: A new model for human action recognition," in *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 3. IOP Publishing, 2019, p. 032035.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[16] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 1–9, 2015.

[19] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," Tech. Rep., 2012. [Online]. Available: http://crcv.ucf.edu/data/UCF101.php

[20] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 4694–4702, 2015.

[21] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[23] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 273–278.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, pp. 2047–2052, 2005.

[25] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction," pp. 1–11, 2018. [Online]. Available: http://arxiv.org/abs/1801.02143

[26] P. Tokmakov, K. Alahari, and C. Schmid, "Learning Video Object Segmentation with Visual Memory," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 4491–4500, 2017.

[27] Y. Huang, W. Wang, and L. Wang, "Video Super-Resolution via Bidirectional Recurrent Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 1015–1028, 2018.

[28] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 1961–1970, 2016.

[29] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative Dictionary Learning with Motion Weber Local Descriptor for Violence Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 696–709, 2017.

[30] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional convolutional lstm for the detection of violence in videos," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[31] S. Zaghbani and M. S. Bouhlel, "Mask rcnn for human motion and actions recognition," in *International Conference on Soft Computing and Pattern Recognition*. Springer, 2020, pp. 1–9.

[32] M. Ding, N. Li, Z. Song, R. Zhang, X. Zhang, and H. Zhou, "A lightweight action recognition method for unmanned-aerial-vehicle video," in *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*. IEEE, 2020, pp. 181–185.

[33] W. Sultani and M. Shah, "Human action recognition in drone videos using a few aerial training examples," *Computer Vision and Image Understanding*, vol. 206, p. 103186, 2021.

[34] B. Mishra, D. Garg, P. Narang, and V. Mishra, "Drone-surveillance for search and rescue in natural disaster," *Computer Communications*, vol. 156, pp. 1–10, apr 2020.

[35] M. Teutsch and W. Krüger, "Detection, segmentation, and tracking of moving objects in UAV videos," *Proceedings - 2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012*, pp. 313–318, 2012.

[36] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 28–35.

[37] A. G. Perera, Y. W. Law, T. T. Ogunwa, and J. Chahl, "A Multiviewpoint Outdoor Dataset for Human Action Recognition," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 405–413, oct 2020.

[38] K. V. V. Subash, M. V. Srinu, M. Siddhartha, N. S. Harsha, and P. Akkala, "Object detection using ryze tello drone with help of mask-rcnn," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020, pp. 484–490.

[39] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A Social Signal Processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2011.12.038

[40] Z. Cheng, L. Qin, Q. Huang, S. Yan, and Q. Tian, "Recognizing human group action by layered model with multiple cues," *Neurocomputing*, vol. 136, pp. 124–135, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2014.01.019

[41] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 438–445, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2011.05.015

[42] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, pp. 1–16, dec 2019.

[43] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[44] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[45] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*. Springer, 2010, pp. 140–153.