IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 March 2023, accepted 27 March 2023, date of publication 30 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263155

RESEARCH ARTICLE

Human Activity Recognition Based on Deep-Temporal Learning Using Convolution Neural Networks Features and Bidirectional Gated Recurrent Unit With Features Selection

TARIQ AHMAD[1], JINSONG WU[2,3], (Senior Member, IEEE), HATHAL SALAMAH ALWAGEED[4], FAHEEM KHAN[5], JAWAD KHAN[6], AND YOUNGMOON LEE[6], (Member, IEEE)

[1]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin 541004, China
[2]School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 510004, China
[3]Department of Electrical Engineering, University of Chile, Santiago 8370451, Chile
[4]College of Computer and Information Sciences, Jouf University, Sakakah 72314, Saudi Arabia
[5]Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea
[6]Department of Robotics, Hanyang University, Ansan 15588, South Korea

Corresponding authors: Jinsong Wu (wujs@ieee.org), Youngmoon Lee (youngmoonlee@hanyang.ac.kr), and Jawad Khan (jkhanbk1@hanyang.ac.kr)

This work was supported by the Research Fund of Hanyang University under Grant HY-2022-0010.

ABSTRACT Recurrent Neural Networks (RNNs) and their variants have been demonstrated tremendous successes in modeling sequential data such as audio processing, video processing, time series analysis, and text mining. Inspired by these facts, we propose human activity recognition technique to proceed visual data via utilizing convolution neural network (CNN) and Bidirectional-gated recurrent unit (Bi-GRU). Firstly, we extract deep features from frames sequence of human activities videos using CNN and then select most important features from the deep appearances to improve performance and decrease computational complexity of the model. Secondly, to learn temporal motions of frames sequence, we design Bi-GRU and feed those deep-important features extracted from frames sequence of human activities to Bi-GRU which learn temporal dynamics in forward and backward direction at each time step. We conduct extensive experiments on realistic videos of human activity recognition datasets YouTube11, HMDB51 and UCF101. Lastly, we compare the obtained results with existing methods to show the competence of our proposed technique.

INDEX TERMS Human activity recognition, recurrent neural networks (RNNs), convolution neural networks (CNNs), bidirectional-gated recurrent unit (Bi-GRU), deep learning.

I. INTRODUCTION

The recent era of artificial intelligence has witnessed the fame of human activity recognition because of its wide range of real-world applications such as healthcare, videos surveillance, smart-homes and human-computer interaction. Human activity recognition plays vital roles in these domains, but more specifically in surveillance applications, it becomes the key factor due to sensitivity of surveillance applications such as employee safety, public security, public transportation and

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

analysis of shopping behavior. Human activities refer to the movement of one or more parts of human body, while human activity recognition is the process of allocating different descriptions to human activities in videos such as (walking, playing, laughing, eating, and so on) and train a system based on these descriptions which enable the trained system to intelligently and spontaneously recognize different activities performed by human in unseen videos.

It is necessary to recognize improper activities of humans in surveillance applications because these may lead to the attempt of stealing, harm of human and many others destructions. Optimal evaluation of human activity is a crucial task which yields challenges to computer vision researchers because of camera motions, human to human interaction, visual similarity, human to object interaction, facial action with object interaction and the same viewpoint of different human activities. To alleviate the problem of sub-optimal recognition performance of human activities many early approaches have been introduced in literature [1], [2], [3]. Although these techniques achieved state-of-the-art performances, but these techniques mostly employed visual data of pre-planned actions in control and unrealistic conditions which may cause many challenges in real-life surveillance applications, because the situation of real-world visual data are more crucial and unstoppable due to clutter background, illumination conditions and variations of camera motions [4], [5], [6].

Several other earlier studies have been presented for human activity recognition based on human silhouettes [7], [8], in which the outline of a person in frames sequence of human activities was extracted via analyzing, examining, decomposing and subtracting noisy background in order to achieve discrimination among different activities, where temporal information was obtained via tracking the movement of human body parts from the combination of extracted silhouettes. However, these approaches extract only local contents of human activities which are suitable only for simple activity recognition and could not be effective in situations where multiple persons perform some activities.

Besides these traditional approaches many deep learning based approaches have been constructed for human activity recognition [9], [10], [11], [12], [13]. The key motivation for computer vision researchers to work in deep learning based human activity recognition was the remarkable improvement in performance via using deep learning in many other domains such as image recognition, face recognition, object detection and person re-identification [14], [15], [16], [17], [18]. Some researchers designed 3-dimensional (3D) Convolution Neural Network (CNN) for human activity recognition [11], and used 2 dimensions for learning the spatial appearances and the third dimension of CNN was allocated to learn temporal motions of human activity frames sequence. Their method outperformed for realistic videos of human activity recognition but 3D CNN faced challenges in long stream realistic videos because the third dimension of CNN can only learn temporal motions of few frames.

The problem of temporal modeling for human activity recognition can be reduced via using RNNs and their variants, due to their gated structures they memorize the earlier information very efficiently of the sequence. RNNs and their variants have showed significant contributions and achieved great results to process sequential data such as time series analysis, audio processing and sequential text data. The sequential characteristics of visual data attracts researcher to work with RNNs and their variants based human activity recognition [9], [12]. Nevertheless, these methods performed well for realistic visual data but RNNs based methods mostly produced vanishing gradients problem for long stream videos due to extensive calculations and sharing the same weights at every time step t, while the RNN variant i.e., long-short term memory (LSTM), requires extensive computations for long stream and high dimensional realistic visual data to process because of its complicated gated structure.

Inspired by the above mention facts, we propose in this paper to tackle these challenges such as sub-optimal evaluation and computational complexity faced by human activity recognition. The main contributions of this paper are summarized as follows:

1) The basic purpose of video surveillance systems is to correctly identify different human activities of realistic visual data. To achieve this we use realistic benchmark videos datasets, YouTube11, HMDB51 and UCF101 [4], [5], [6], while, to improve recognition rate of human activities, we extract deep features from deep network VGG16 [14], which has been trained on million of images. Thus, we argue that extracting deep features for human activities from pre-trained model can enhance recognition performance.
2) We employee random forest algorithm to select most important features from deep features and reduce the dimension of features map. After that, the reduced features vector which consists of only important features fed to train our model with the aim to decrease computational complexity.
3) Bi-GRU consists of simple gated structure i.e., reset and update gates which memorize a long sequence of data. Moreover, it propagates the input sequence in forward and backward directions. Therefore, we propose Bi-GRU which effectively learns the frame to frame changes of human activities at each time step t to alleviate the problem of temporal modeling.

II. RELATED WORKS

In this section we review the literature which are related to our proposed method. For human activity recognition, some traditional methods have been studied which based on hand-crafted features extraction [19], [20], [21], [22] for non-realistic videos, which defined visual data of human activity as local descriptors. One popular method [23], uses two kinds of features for human activity recognition such as histogram of oriented gradients (HOG) and motion history image (MHI). For instance, HOG based features are

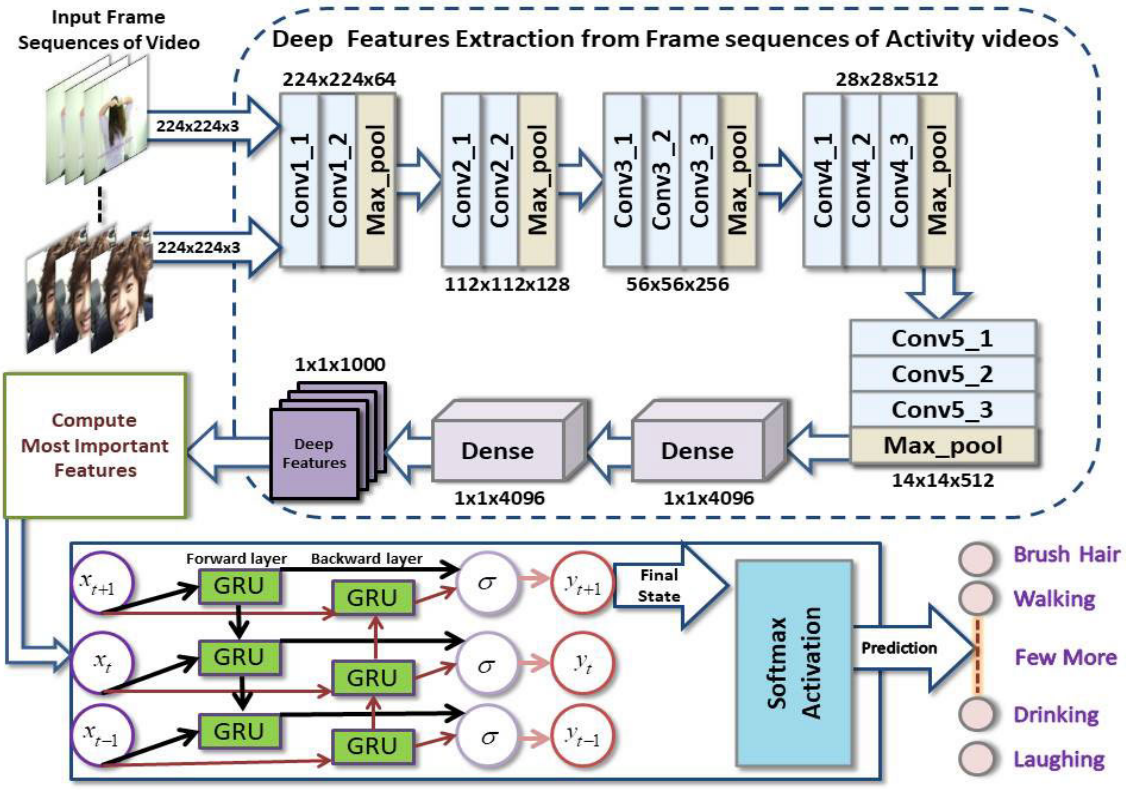


FIGURE 1. Structure of the proposed deep-temporal learning with features selection for human activity recognition.

extracted from boundaries of human activities videos via analyzing magnitude and viewpoint, while in MHI the noisy background is subtracted. Afterward, these features are fused and input for recognition through generic classifier which is implemented on top of these features. Subsequently, Liu et al. [6] presented classification of human activities for realistic videos via extracting spatial and motion features. They applied motion statistics to transform noisy motion features into well-structure shapes. Furthermore, they also used data mining technique of PageRank to select different spatial features which are informative for particular human activity class. However, human activity recognition based on hand-crafted features only captures limited contents of particular activity which may cause confusions between two different activities. In addition, these methods may also not feasible for multiple person activity recognition in visual data.

In contrast to hand-crafted based methods, many deep learning approaches have been introduced for human activity recognition [10], [11], [24], [25], [26], [27], [28]. Karpathy et al. [10] presented a framework for features connectivity in time axis via capturing local spatial-temporal contents of human activity. In addition, they evaluated their experiments via introducing one million videos dataset. To improve recognition performance, [28] applied features fusion strategy for two types of features acquired from different fully connected (FC) layers of pre-trained CNN. Furthermore, they also reduced dimensions of features map via exploiting features selection techniques and then at later stages a generic classifier trained over those selected features. Others famous approaches relied on 3D CNN [11], [24], [25], which were the extensions of 2D CNN, and used two dimensions of CNN to learn spatial information and the third dimension devoted to capture temporal motions. However, 3D CNN requires high training complexity and only processes few frames efficiently in the third dimension. By these constructions, 3D CNN faces challenges to capture fine details of temporal motions in processing long stream videos.

To improve the performance of temporal motions several works have been proposed based on RNNs [29], [30]. The former approach used framework of features extraction from different FC layers of pre-trained CNN for human activity frames sequence and then input those features to multilayer gated recurrent unit (GRU) for prediction whereas the later approach applied features fusion technique of two-stream LSTM for human activity recognition task.

Recently, several extensions have been presented to improve recognition of human activity by self-supervised video transformer [31], VideoMoCo [32], and confidence distillation [33]. Furthermore, graph convolution networks (GCN), and attention based models [34], [35], [36] have been used extensively in recent years due to optimal performance

for skeleton based human activity recognition. One popular approach mRi multi-model [37], proposed 3d human pose estimation skeleton dataset. In addition, this method provide the most comprehensive sensing modalities and action detection via using pyramid attention architecture.

Deep learning based methods have the capability to capture fine details of visual data [38], due to huge weighted layers of features representation [39], [40]. However, deep learning based approaches need extensive computations and huge amounts of data for training the model from the scratch. Thus, in this paper we propose a framework to stabilize computation complexity via using pre-trained CNN for features extraction and then select the most important features with the aim to reduce dimension of feature maps. Furthermore, we design Bi-GRU for modeling temporal dynamics of human activity videos.

III. PROPOSED FRAMEWORK

In this section, we describe the key components of our proposed method for human activity recognition including deep features Ft_{dp} extraction from frames sequence F_s of human activity videos V_h, after that we select most important features Ft_{imp} from deep extracted features to reduce dimension of features map by using random forest algorithm which helps in decreasing computational complexity in training, lastly we feed important features of every individual frame Ft^i_{imp} along with descriptions (one-hot labels) T_{lbs} of every class to Bi-GRU for training. Bi-GRU learns deep and temporal information of each frame at each time step t, and then the trained Bi-GRU evaluates the unseen frames of test data for prediction of different classes C_A of human activity recognition. The workflow structure of our proposed method is shown in Fig. 1.

A. PREPARATION AND PREPROCESSING OF FRAMES

Videos are the combination of frames sequence at 30 frames per second. To understand the story of running video, people need to analyze several frames in a sequence. In addition, features extraction from pretrained model for video data also require RGB frames sequence. Therefore, we extract frames sequence of size $224 \times 224 \times 3$ from human activity videos which is the desired and fixed input size of VGG16 model [14]. Preprocessing plays vital role for any machine learning model [41], without which may lead to extensive computations and sub-optimal recognition performance. Thus, we use mean subtraction preproceesing strategy for frames sequence of human activity videos. The given frames sequence $F_s = \{F_1, F_2, F_3, \ldots . F_i \ldots, F_N\}$, where each frame $F_i \in \mathbb{R}^{h \times w \times c}$, h represents height, w width and c color channel of RGB frame. The mean of frame F_i can be computed as follows

$$\mu F_i = \frac{F_i^{h \times w \times c}}{h \times w \times c}, \tag{1}$$

where the superscript $h \times w \times c$ represents pixel values and divisor $h \times w \times c$ are the total count of height, width and color channel of frame F_i. Now, to achieve mean subtraction preprocessing for RGB frame F_i, we require to subtract μF_i from ImageNet mean $\mu ImgN_i$ such as

$$F_{pre} = \mu ImgN_i - \mu F_i, \tag{2}$$

where F_{pre} is our required preprocessing of frame F_i for features extraction from VGG16 model and the values of $\mu ImgN_i$ equivelent to [0.485, 0.456, 0.406] which is provided by ImageNet dataset [42].

B. DEEP FEATURES EXTRACTION

Standard Convolution neural networks (CNNs) contains stack of convolution layers which perform dot product of input image and filters (kernels) with respect to defined strides [43], and then followed by pooling layers to reduce size of features map acquired from convolution layers. Moreover, pooling layers also help speed up the required calculations. At the end, the convoluted features extracted from previous layers are injected to FC layers with the aim to obtain intrinsic information from convoluted features [44] and then those features are used for prediction of image classes. Training CNNs from scratch required huge dataset, extensive computation and powerful hardware resources. On the contrary, the same results can be obtained for classification and prediction task via using pretrained CNNs. Therefore, we exploit the pretrained model for feature extraction from human activity frames sequence because pretrained CNNs learn deep hidden patterns from million of images in ImageNet dataset. In addition, pretrained CNNs also have the ability in knowledge inferences and features representation. However, there are many famous CNNs such as VGG19, DenseNet and Xception [14], [15], [16] but we select VGG16 model for features extraction because of the following factors.

1) VGG16 model consists of only 16 weighted layers for features representation and uses the small number of weighted layers which can help reduce the computation cost during features extraction from frames sequence.
2) Optimization of machine learning model requires manual setting of hyper-parameter but VGG16 model uses 3×3, 2×2 filters and 1, 2 strides for every convolution and max-pooling layers respectively instead of using large number of hyper-parameters.
3) Human activity recognition requires deeper network for discriminative features extraction, and VGG16 is generally deeper than other models because it has 138 million learning parameters.

According to the above facts, we use VGG16 as base model for features extraction. The architecture of VGG16 model with features map representation of frames during features extraction is given in Table. 1. The base model divided into five convolution blocks, whereas every block consists of convolution layers followed by max-pooling layers and at the end three FC layers were injected. The ReLU non-linear activation function is used regularly in training the model.

TABLE 1. Features map representation of human activity video frames at each layer of VGG16 model.

Layers	Conv1	Conv2	Conv3	Max-Pooling	FC6	FC7	FC8
Features map of Block-1	224×224	224×224	-	112×112	-	-	-
Features map of Block-2	112×112	112×112	-	56×56	-	-	-
Features map of Block-3	56×56	56×56	56×56	28×28	-	-	-
Features map of Block-4	28×28	28×28	28×28	14×14	-	-	-
Features map of Block-5	14×14	14×14	14×14	7×7	-	-	-
Filters size	3×3	3×3	3×3	2×2	-	-	-
Stride	1	1	1	2	-	-	-
Dimension of Block-1	64	64	-	64	-	-	-
Dimension of Block-2	128	128	-	128	-	-	-
Dimension of Block-3	256	256	256	256	-	-	-
Dimension of Block-4	512	512	512	512	-	-	-
Dimension of Block-5	512	512	512	512	-	-	-
Dimension of FC Layers	-	-	-	-	4096	4096	1000

We preprocess frame F_{pre} of size 224 × 224× 3, then by feeding it to the first convolution layer of base model we get

$$Ft_{conv1} = F_{pre} * K, \tag{3}$$

where Ft_{conv1} are the features map of frame F_{pre} obtained from the first convolution layer of base model and K represents filters (kernels) with size of 3 × 3 which is pre-defined in base model. Now, if we further elaborate (3), we get:

$$Ft_{conv1} = \sum_{p}\sum_{q} K[p, q]F_{pre}[m, n], \tag{4}$$

where m, p represents rows and n, q are columns of frame F_{pre} and kernel K respectively. In (4), the dot product will be performed between part of input frame F_{pre} and kernel K with respect to defined sliding window e.g., (stride = 1) and the resultant vector Ft_{conv1} considers features map of frame F_{pre} acquired from first convolution layer. After the first convolution layer, VGG16 model uses the second convolution layer to further shrink the spatial dimension of input image whereas the same intuition were applied in conv2 layer, we can write (4) for frame F_{pre} as

$$Ft_{conv2} = \sum_{p}\sum_{q} K[p, q]Ft_{conv1}[m, n], \tag{5}$$

where Ft_{conv2} are features map of frame F_{pre} extracted from conv2 layer of base model. After the second convolution, one max-pool layer is added to reduce size of features vector.

The output of Max-pool layer can be computed via applying it on (5), and then we get

$$Pool^{1}_{max} = max^{j}_{i}(Ft^{n}_{m_{conv2}}), \tag{6}$$

where i, j are filters of max-pool operation whereas n, m represents rows and columns of Ft_{conv2}. After that, the same intuition is applied for features extraction in subsequent layers up-to block five. We can write extracted features map of frame F_{pre} for max-pool of block five in equation such as

$$Pool^{5}_{max} = max^{j}_{i}(Ft^{n}_{m_{conv3}}) \tag{7}$$

Lastly, (7) followed by three FC layers with aim to extract intrinsic information from frame F_{pre}. The operation in FC layer change the dimension of convoluted features extracted in (7) to be flattened. The stack of three FC layers are of dimension 4096, 4096 and 1000 were applied on (7), respectively. The mathematical representation of first FC layer for frame F_{pre} can be written as

$$\begin{aligned} Ft_{C_A} &= (Pool^{5}_{max}, D_{flatn}), \\ Ft_{C_A} &= Ft^{pool5}_{pre}, \end{aligned} \tag{8}$$

where Ft_{C_A} represents features of human activities classes while D_{flatn} operation convert (7) features i.e., $(Pool^{5}_{max})$ into intrinsic information and change the dimension to 4096. However, the second FC layer is applied in the same manner but the last FC layer which is the deep representation of frame F_{pre} consists of 1000 dimension and its our desired features extraction layer for human activity recognition. The extracted features for frame F_{pre} from last layer of VGG16 model can be written as

$$Ft_{dp} = (Ft_{C_A}, D_{flatn}) \tag{9}$$

C. IMPORTANT FEATURES SELECTION

Important features selection from the dataset can play a significant role in reducing the model training complexity [45]. Additionally, feeding all features to the model may decrease the model performance [46] due to the presence of some noisy and redundant features. However, there are numerous features selection techniques such as filter based and wrapper methods. The former select features based on uni-variate statistics instead of cross validation which face difficulties in optimal set of selected features whereas the later select features based on greedy search algorithm which is computationally expensive. Therefore, we use random forest algorithm which carefully select important features from deep representation Ft_{dp} of frame F_{pre} on basis of their contribution towards human activity. We need to implement random forest algorithm between deep features Ft_{dp} and ground truth T_{lbs} (one-hot-labels). First, we need to construct decision tress based on information gain between deep features of human activity frames Ft_{dp} and ground truth target labels T_{lbs}.

$$\begin{aligned} Inf_{gain}(Ft_{dp}, T_{lbs}) &= Entropy(T_{lbs}) \\ &\quad - Entropy(Ft_{dp}, T_{lbs}) \end{aligned} \tag{10}$$

Information gain $Inf_{gain}(Ft_{dp}, T_{lbs})$ helps us find the purest child nodes of decision tree because its reduce uncertainty recursively and continues until all child nodes are not pure. In (10), *Entropy* equivalent to

$$Entropy = \sum_{i} -p_i(log_2 p_i),$$

where p_i is the probability of i^{th} feature. We assume the constructed decision tree D_{tree}, which we rewrite (10) in mathematical form as

$$\begin{aligned} D_{tree} &= L_1, L_2, \ldots L_n \\ L_1 &= Ch_1, Ch_2 \ldots . Ch_n \\ &\vdots \\ &\vdots \\ &\vdots \\ L_n &= Ch_n 1, Ch_n 2 \ldots . Ch_n m \end{aligned}$$

where $L_1, L_2, \ldots L_n$ are leaves nodes of features and $Ch_1, Ch_2 \ldots . Ch_n$ represent child nodes of leaves nodes respectively. Nodes importance can be calculated by using gini impurity index. We assume gini index applied on leave node L_i, then we get:

$$IL_i = wV - w_{lft}V_{lft} - w_{rt}V_{rt} \ldots \ldots w_n V_n, \tag{11}$$

where IL_i represent importance of L_i, w is weighted number of features reached to L_i where V is impurity value and w_{lft}, w_{rt} are left and right childs node of leave node. Next, the importance of each feature computed in decision tree via calculating their weighted average as

$$Ft_i = \frac{\sum IL_i}{\sum L_n}, \tag{12}$$

where Ft_i is the i^{th} deep feature of human activity frame. Moreover, feature Ft_i normalized between 0 and 1 via dividing it on the summation of all important features.

$$normFt_i = \frac{Ft_i}{\sum Ft_{all}} \tag{13}$$

Lastly, the $normFt_i$ feature is divided by total decision tree, and then we get the most important features Ft_{imp} from random forest algorithm of human activity frame such as

$$Ft^i_{imp} = \frac{\sum normFt_i}{D_{tree}}. \tag{14}$$

D. BIDIRECTIONAL GRU

RNNs are type of artificial neural network design for processing sequential data e.g., time series, text mining, audio data. In sequential data, every data point in a sequence depends upon the previous data point such as sentence, audio and speech data, but traditional artificial neural networks cannot process such data efficiently because they construct only independent data points. RNNs have the principle of memory cells which helps store information of previous inputs to generate next outputs of the sequence. Simple RNNs have made great successes in modeling sequential data, but, for long stream and high dimensionality of sequential visual data, they may suffer vanishing gradient problems. Therefore, we use Bi-GRU a variant of RNN to learn visual data of human activity to overcome vanishing gradient problem. Bi-GRU consists of two GRU stack one after another where the input information of sequence propagate by one GRU in forward direction to previous time step and the other propagate in backward direction to later time step to make prediction of current state. The general structure of Bi-GRU for processing sequential data is presented in Fig. 2. We have deep important features Ft^i_{imp} obtained from (14), and then input those feature to Bi-GRU for training. The training process of Bi-GRU for human activity recognition can be summarized as follows

$$I_t = \sigma(Ft^i_{imp}W_r + H_{t-1}W_r + b_r), \tag{15}$$

$$Z_t = \sigma(Ft^i_{imp}W_z + H_{t-1}W_z + b_r), \tag{16}$$

$$\tilde{H}_t = tanh(Ft^i_{imp}W_h + (I_t + \odot H_{t-1})W_h + b_h), \tag{17}$$

$$\vec{H}_t = Z_t \odot \vec{H}_{t-1} + (1 - Z_t) \odot \tilde{H}_t), \tag{18}$$

$$\overleftarrow{H}_t = Z_t \odot \overleftarrow{H}_{t-1} + (1 - Z_t) \odot \tilde{H}_t). \tag{19}$$

For the given time step t the human activity deep important feature Ft^i_{imp} input for processing then reset gate I_t and update gate Z_t of GRU computed, where H_{t-1} represent hidden state of previous time step t. Next, reset and update gates integrated to form candidate hidden state $\tilde{H}_t$. We use Bi-GRU, and then the intuition of new hidden state H_t is applied separately for each GRU cell in both forward and backward direction. Forward hidden state $\vec{H}_t$ of one GRU cell is compared with the earlier hidden state H_{t-1} in order to resembles the new candidate hidden state $\tilde{H}_t$, and the second GRU cell compare its hidden state $\overleftarrow{H}_t$ with new candidate state $\tilde{H}_t$ to know the upcoming information of next frame for human activity in the sequence. Finally, the trained Bi-GRU forward the information to softmax activation function for prediction. We can represent softmax function as

$$final_{state} = softmax(H_t). \tag{20}$$

IV. EXPERIMENTAL EVALUATION

A. ENVIRONMENT

The experiment has been conducted in python 3.8 with cuda and cudnn versions of 11.2 and 8.8.1. CoreTM i7-4790 processor with 16 GB RAM and GeForce GTX 1080Ti GPU of 11GB used for processing our experiment. We use deep learning framework 'keras' for features extraction, Sklearn library for features selection and tensorflow for implementation of Bi-GRU. For training, our proposed Bi-GRU is with the total number of 100 epochs, mini-batch size of 256, weight decay of $1e^{-6}$. and the initial learning rate of 0.001 are applied which later decrease to 0.0001. The dataset are split in ratio of 60:20:20, where (60%) are allocated for training, (20%) for validation and for performance assessment on unseen data by our proposed Bi-GRU, (20%) of data assigned to testing. As for performance evaluation, we have used three

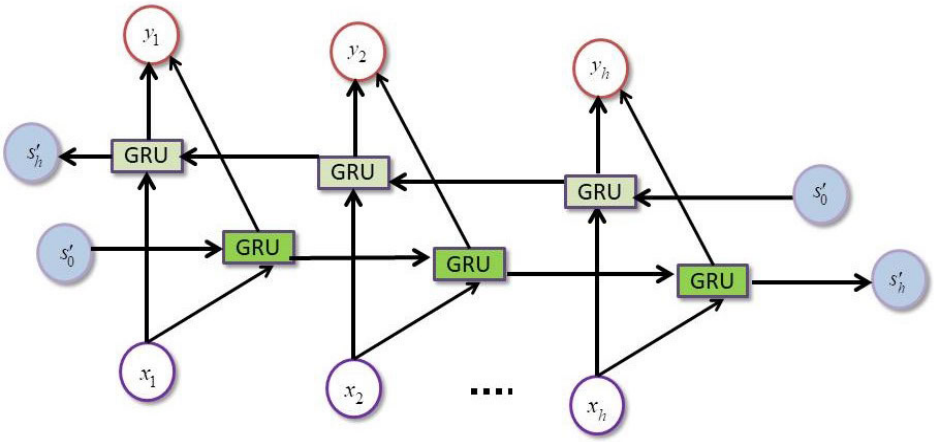


FIGURE 2. General structure of bidirectional gated recurrent unit for processing sequential data.

metrics such as score of average accuracy, confusion matrix and class-wise accuracy for the three benchmark dataset i.e., (YouTube11, HMDB51 and UCF101) to show effectiveness of our method.

B. DATASETS

1) *YouTube11 Dataset:* Its a realistic video dataset of human activity recognition which contain 1160 sports videos and divided into 11 classes i.e., walking, volleyball, jumping, tennis swing, swing, soccer, horse riding, golf swing, diving, biking and basektball. Each class of YouTube11 dataset further divided into 25 groups where every group consists of more than 4 videos of human activities which share some similarity such as similar actor, viewpoint and background conditions. This dataset of human activity recognition is small but very challenging due to clutter background, illumination condition and camera motion of videos.
2) *HMDB51 Dataset:* its a large dataset of human activity videos based on realistic conditions. The dataset include 6849 videos which are divided in 51 classes where each class holds at least 101 videos. The videos of dataset were captured from movie clips and youtube in short duration which make it more challenging than any other dataset. Additionally, classes of the dataset based on similiarity among different human activities such as facial action of human:(talk, laugh), facial action by using objects:(eat, drink), body movement:(climb, walk), human to human body movement:(shake hands, hugging), body movement by interacting some objects:(shooting a gun, hiting something), which make recognition of human activity more difficult for any system.
3) *UCF101 Dataset:* The largest realistic dataset of human activity recognition include approximately 13320 videos which split up in 101 classes. Videos of UCF101 dataset capture from youtube in five different types such as sports, human body movement, paly-ing music, human to human interaction and human to object interaction. The dataset is very challenging due to similarity between different classes, different illumination condition and viewpoint. UCF101 dataset consists of real-life videos where many others dataset of human activity recognition based on unrealistic and pre-planned action perform by actor.

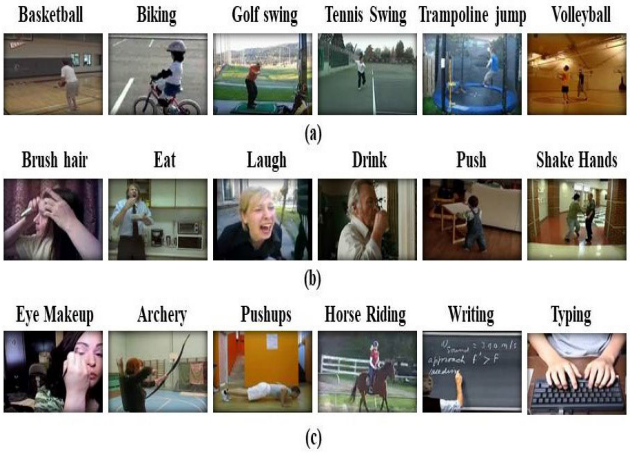


FIGURE 3. Sample frames of human activities datasets in (a) YouTube11, (b) HMDB51 and (c) UCF101.

The sample frames of different human activities classes from the three benchmark dataset given in Fig. IV-C (a) YouTube11 (b) HMDB51 and (c) UCF101.

C. COMPARATIVE ANALYSIS

The proposed method is compared with existing methods with respect to average accuracy captured from test set of data given in Table. 2. The dash "-" sign indicates accuracy not reported by the reference paper for the corresponding dataset and accuracy given in bold represent high accuracy.

TABLE 2. Comparison of proposed method with existing methods.

CNNs based and Others State of the Art methods	YouTube11 dataset %	HMDB51 %	UCF101 %
Large-scale video classification CNNs [10]	-	-	65.4
Factorized Spatio-Temporal CNNs [24]	-	59.1	88.1
Two-stream CNNs [26]	-	59.4	88.0
Self-supervised Video Transformer [31]	-	67.2	90.8
VideoMoCo [32]	-	49.2	78.7
Confidence Distillation [33]	-	-	91.2
Single stream CNNs [47]	93.1	-	-
Improved trajectories [22]	-	57.2	-
Hierarchical multi-task learning [48]	89.7	51.4	76.3
Dense trajectories [20]	-	52.1	83.8
Soft kernel learning [49]	91.6	28.2	79.7
Key volume mining [52]	-	43.7	65.2
RNNs based Methods	**YouTube11 dataset %**	**HMDB51 %**	**UCF101 %**
Unsupervised learning using LSTMs [9]	-	44.0	84.3
P-RRNNs [50]	-	68.2	91.4
Spatio-temporal bidirectional LSTM [51]	-	70.4	-
Encoding RNNs [53]	-	54.9	81.9
Deep Bi-Directional LSTM [54]	92.84	-	91.21
Proposed method	**93.38**	**71.89**	**91.79**

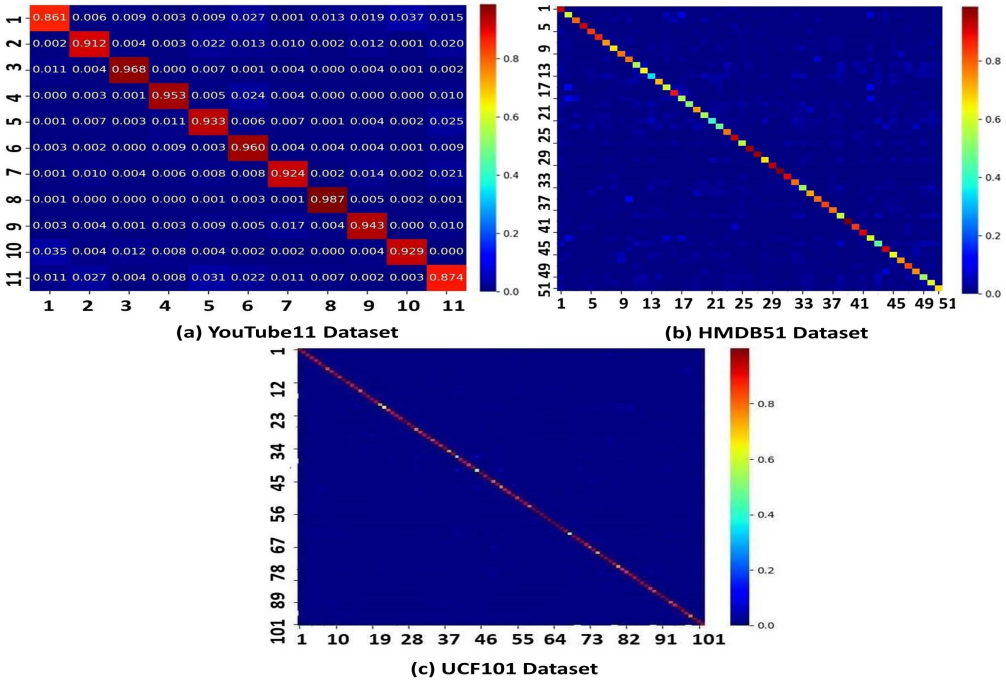


FIGURE 4. Confusion matrices of proposed method for human activity recognition obtain from test set of data, in (a) YouTube11 dataset, (b) HMDB51 dataset and (c) UCF101 dataset.

The existing methods such as single stream CNNs [47], hierarchical multi task learning [48] and soft kernel learning [49] reported 93.1%, 89.7% and 91.6% average accuracy for human activity recognition using YouTube11 dataset while our proposed method obtain 93.38% for the same dataset and improve 0.28% recognition accuracy.

In the given Table 2, we can observe average accuracy for HMDB51 dataset by two stream CNNs [26], hierarchical multi-task learning [48], improved trajectories [22], unsupervised LSTM [9], P-RRNs [50]. They claimed 59.4%, 51.4%, 57.2%, 44.0% and 68.2% accuracies for HMDB51 dataset respectively. The second highest accuracy of 70.4% reported for this dataset by bidirectional LSTM [51]. However, our method holds the highest accuracy of 71.8% which enhance recognition performance up to 1.4% for HMDB51 dataset. As for UCF101 dataset the comparison based on average accuracy are demonstrated in Table 2. Few existing methods claimed more than 75% accuracy for UCF101 dataset such as hierarchical multi-task learning [48] and soft kernel learning [49]. On the other hand, several existing state-of-the-art

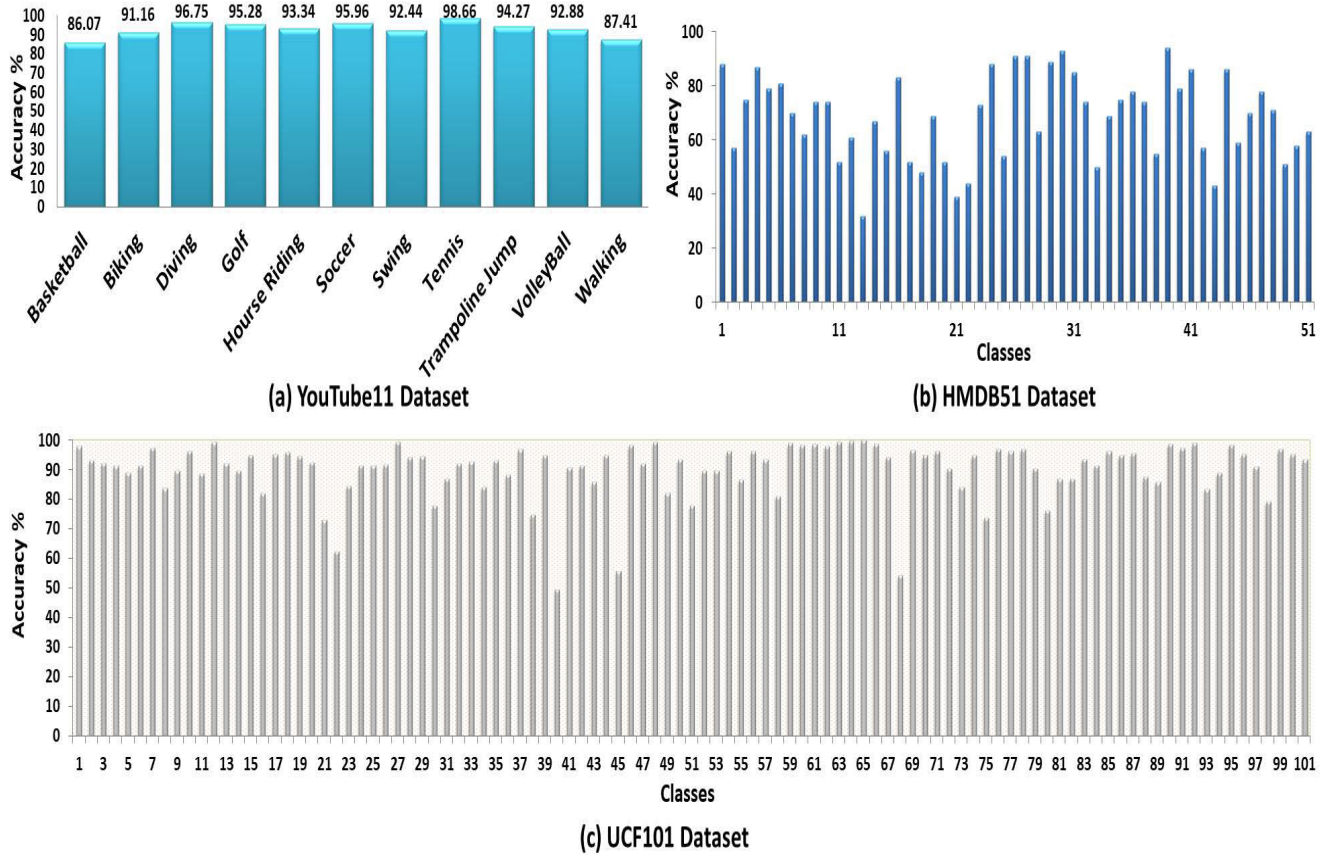


FIGURE 5. Class-wise accuracy of proposed method for human activity recognition for three benchmark datasets, in (a) YouTube11 dataset, (b) HMDB51 dataset and (c) UCF101 dataset.

techniques obtained more than 80% accuracy. Additionally, P- RRNNs [50] gain 91.4% accuracy, even though our method achieved 91.79% and increase 0.3% accuracy for UCF101 dataset which demonstrate effectiveness of our method.

D. PERFORMANCE EVALUATION

We evaluate performance on test split of each dataset participated in the experiment. In order to show recognition performance of our proposed method for each dataset, confusion matrices presented in Fig. 4 (a) YouTube11 dataset, (b) HMDB51 dataset and (c) UCF101 dataset.

Confusion matrix for YouTube11 dataset can be visualized in Fig. 4 (a), where most of the classes surpass 90% recognition accuracy instead of class '1' and '11' which represents basketball and walking. These two classes have received 86.1% and 87.4% recognition accuracy respectively. The class 'basketball' perform poorly because some other classes interfere and has a tendency of predicting basketball as soccer and volleyball. In addition, class walking is misclassified as biking and horse riding.

The class-wise accuracy is obtained for each dataset used in our experiment and can be observed from Fig. 5. The graphs in Fig. 5 (a), (b) and (c) represent class-wise accuracy of YouTube11, HMDB51 and UCF101 datasets respectively.

The graph Fig. 5 (b) is acquired from test set of data during evaluation of our experiment for HMDB51 dataset. In graph Fig. 5 (b), many classes get moderate accuracy, for example, more than half classes exceed 55% recognition accuracy and some classes achieve the highest accuracy i.e., (above 90%). Moreover, few classes perform poorly and get the lowest accuracy i.e., (less than 50%). For instance, it can be seen classes at points 13, 18, 21, 22 and 43 in horizontal axis, obtain 32.91%, 48.48%, 39.01%, 44.28% and 43.45% accuracies (can be seen in vertical axis) of Fig. 5 (b) graph. These points represent classes 'fall on the floor', 'hit', 'kick', 'kick ball' and 'stand' in HMDB51 dataset respectively. The highest accuracy is achieved by class ' baby situp' which can be seen at point 39 in horizontal axis whereas the accuracy can be observe at vertical axis of the same graph.

Class-wise accuracy for UCF101 dataset is presented in Fig. 5 (c) graph, in which numerous classes receive more than 75.0% recognition accuracy. However, very few classes demonstrate unsatisfactory performance such as in points 40, 45 and 68 possess 49.7%, 55.95 and 54.3% recognition accuracy which express classes 'jumping high', 'metal-tipped javelin throwing' and 'vaulting pole' in UCF101 dataset respectively. The highest accuracy of 99.95% is achieved by class 'playing music with sitar ' which can observed at point 65 in horizontal axis of Fig. 5 (c) graph.

We also evaluate performance and time complexity of our proposed method by employing deep features of YouTube11, HMDB51 and UCF101 dataset to prove feasibility and significance of random forest features selection technique. We trained Bi-GRU upto 100 epochs and used same trainable parameters i.e., (described in section IV-A) for deep features and obtain 92.05%, 67.53% and 88.07% recognition accuracy while selected features gain 93.38%, 71.89% and 91.79% accuracy for YouTube11, HMDB51 and UCF101 datasets respectively. Furthermore, we also evaluate average time complexity of training the Bi-GRU by using deep features and selected important features of each dataset which shows that time complexity of our method significantly decrease by using random forest algorithm for important features selection which also helps in improvement of accuracy. In addition, [54] claimed 1.12 seconds average time complexity to train their model by using every six frames in the sequence of YouTube11 dataset while our proposed method consumes 1.03 seconds average time for training the Bi-GRU which shows the efficiency of our propose method. The acquired results of deep features and important selected features for each dataset are presented in the given Table. 3.

TABLE 3. Comparison of recognition performance and average time complexity between deep features and selected most important features for YouTube11, HMDB51 and UCF101 dataset.

Dataset	Feature Type	Epochs	Accuracy%	Average time
YouTube11	Selected Features	100	93.38	1.03 Sec
YouTube11	Deep Features	100	92.05	1.70 Sec
HMDB51	Selected Features	100	71.89	3.01 Sec
HMDB51	Deep Features	100	67.53	4.56 Sec
UCF101	Selected Features	100	91.79	5.62 Sec
UCF101	Deep Features	100	88.07	7.15 Sec

V. CONCLUSION

In this paper, we have proposed human activity recognition based on deep-temporal learning by using deep CNN features, then selected most important features among deep representations and feed selected features to learn temporal dynamics. We have discussed different problems faced by human activity recognition and mainly focused on improvement in accuracy, reduction computational complexity and learn long sequence temporal motions. To achieve this we have extracted intrinsic information of human activity frames sequences from pre-trained CNN which help the increasing performance, while, to reduce computational complexity, we have employed random forest algorithm to select most important features from deep intrinsic information of human activity. After this, we have proposed to use Bi-GRU and feed selected deep information at each time step t to improve temporal dynamics of long stream videos. The experimental results have shown that our proposed method perform well compared with other existing methods. The proposed method has some limitations, such as, it can recognize human activities when it runs on GPU, but its unable to predict human activity on internet of things (IOT) based devices.

REFERENCES

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2004, pp. 32–36.

[2] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1948–1955.

[3] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action Mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[4] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[6] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1996–2003.

[7] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[8] L. Wang and D. Suter, "Informative shape representations for human action recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 2, Aug. 2006, pp. 1266–1269.

[9] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," 2015, *arXiv:1502.04681*.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[12] 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," 2014, *arXiv:1411.4389*.

[13] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," 2015, *arXiv:1503.08909*.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016, *arXiv:1610.02357*.

[17] R. Couturier, H. N. Noura, O. Salman, and A. Sider, "A deep learning object detection method for an efficient clusters initialization," 2021, *arXiv:2104.13634*.

[18] S. Zhong, Z. Bao, S. Gong, and K. Xia, "Person reidentification based on pose-invariant feature and B-KNN reranking," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 5, pp. 1272–1281, Oct. 2021.

[19] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 357–360.

[20] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.

[21] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 275–1275.

[22] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[23] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 128–135.

[24] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4597–4605.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

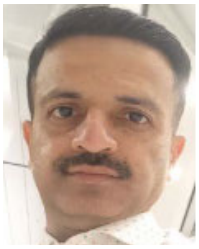
[26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1–9.
[27] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream ConvNets," 2015, *arXiv:1507.02159*.
[28] T. Ahmad, J. Wu, I. Khan, A. Rahim, and A. Khan, "Human action recognition in video sequence using logistic regression by features fusion approach based on CNN features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 18–25, 2021.
[29] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," 2015, *arXiv:1511.06432*.
[30] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream LSTM: A deep fusion framework for human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 177–186.
[31] K. Ranasinghe, M. Naseer, S. Khan, F. Shahbaz Khan, and M. Ryoo, "Self-supervised video transformer," 2021, *arXiv:2112.01514*.
[32] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "VideoMoCo: Contrastive video representation learning with temporally adversarial examples," 2021, *arXiv:2103.05905*.
[33] S. M. Shalmani, F. Chiang, and R. Zheng, "Efficient action recognition using confidence distillation," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3362–3369.
[34] Y. Liu, H. Zhang, Y. Li, K. He, and D. Xu, "Skeleton-based human action recognition via large-kernel attention graph convolutional network," *IEEE Trans. Vis. Comput. Graphics*, early access, Feb. 22, 2023, doi: 10.1109/TVCG.2023.3247075
[35] Z. Deng, Q. Gao, Z. Ju, and X. Yu, "Skeleton-based multi-features and multi-stream network for real-time action recognition," *IEEE Sensor J.*, early access, Feb. 23, 2023, doi: 10.1109/JSEN.2023.3246133.
[36] G. Zheng, "A novel attention-based convolution neural network for human activity recognition," *IEEE Sensors J.*, vol. 21, no. 23, pp. 27015–27025, Dec. 2021.
[37] S. An, Y. Li, and U. Ogras, "MRI: Multi-modal 3D human pose estimation dataset using mmWave, RGB-D, and inertial sensors," 2022, *arXiv:2210.08394*.
[38] T. Ahmad and J. Wu, "SDIGRU: Spatial and deep features integration using multilayer gated recurrent unit for human activity recognition," *IEEE Trans. Computat. Social Syst.*, early access, Mar. 9, 2023, doi: 10.1109/TCSS.2023.3249152.
[39] M. Woźniak, M. Wieczorek, and J. Siłka, *Deep Neural Network With Transfer Learning in Remote Object Detection From Drone*. New York, NY, USA: Association for Computing Machinery, 2022, doi: 10.1145/3555661.3560875.
[40] M. Woźniak, M. Wieczorek, and J. Siłka, "BiLSTM deep neural network model for imbalanced medical data of IoT systems," *Future Gener. Comput. Syst.*, vol. 141, pp. 489–499, Apr. 2023, doi: 10.1016/J.FUTURE.2022.12.004.
[41] A. Rahim, Y. Zhong, T. Ahmad, and U. Islam, "An intelligent approach for preserving the privacy and security of a smart home based on IoT using LogitBoost techniques," *J. Hunan Univ. Natural Sci.*, vol. 49, no. 4, pp. 372–388, Apr. 2022.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[43] A. Rahim, Y. Zhong, and T. Ahmad, "A deep learning-based intelligent face recognition method in the Internet of Home things for security applications," *J. Hunan Univ. Natural Sci.*, vol. 49, no. 10, pp. 39–52, Oct. 2022.
[44] M. A. Khan, N. S. Elmitwally, S. Abbas, S. Aftab, M. Ahmad, M. Fayaz, and F. Khan, "Software defect prediction using artificial neural networks: A systematic literature review," *Sci. Program.*, vol. 2022, pp. 1–10, May 2022.
[45] M. A. Ashraf, Y. D. Khan, B. Shoaib, M. A. Khan, F. Khan, and T. Whangbo, "β lact-pred: A predictor developed for identification of beta-lactamases using statistical moments and PseAAC via 5-step rule," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–10, Dec. 2021.
[46] F. Khan, I. Tarimer, H. S. Alwageed, B. C. Karadağ, M. Fayaz, A. B. Abdusalomov, and Y.-I. Cho, "Effect of feature selection on the accuracy of music popularity classification using machine learning algorithms," *Electronics*, vol. 11, no. 21, p. 3518, Oct. 2022.
[47] S. Ramasinghe and R. Rodrigo, "Action recognition by single stream convolutional neural networks: An approach using combined motion and static information," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 101–105.
[48] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.
[49] X. Xu, I. W. Tsang, and D. Xu, "Soft margin multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 749–761, May 2013.
[50] S. Yu, L. Xie, L. Liu, and D. Xia, "Learning long-term temporal features with deep neural networks for human action recognition," *IEEE Access*, vol. 8, pp. 1840–1850, 2020.
[51] W. Li, W. Nie, and Y. Su, "Human action recognition based on selected spatio-temporal features via bidirectional LSTM," *IEEE Access*, vol. 6, pp. 44211–44220, 2018.
[52] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1991–1999.
[53] A. Richard and J. Gall, "A bag-of-words equivalent recurrent neural network for action recognition," 2017, *arXiv:1703.08089*.
[54] 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, 2017.

TARIQ AHMAD received the M.Sc. degree in computer science from the University of Science and Technology Bannu, Pakistan, in 2011, and the M.Sc. degree in computer science, with a focus on software engineering, from IMSciences, Peshawar, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in communication and information systems, with a focus on artificial intelligence, with the Guilin University of Electronic Technology, Guilin, China, under the supervision of Prof. Jinsong Wu.

His research interests include human activity recognition, sequence learning for visual data, machine learning, deep learning, context-based image indexing and retrieval, facial expression prediction, and computer vision. He has been a Reviewer of machine learning and deep learning in IEEE Access, since December 2021.

JINSONG WU (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada, in 2006. He received the 2020 IEEE Green Communications and Computing Technical Committee Distinguished Technical Achievement Recognition Award, for his outstanding technical leadership and achievement in green wireless communications and networking, the 2017 IEEE Green Communications and Computing Technical Committee Excellent Services Award for his excellent technical leadership and services in the Green Communications and Computing Community, the 2017, 2019, and 2021 IEEE Systems Journal Best Paper Awards, and the 2018 IEEE Green Communications and Computing Technical Committee Best Magazine Paper Award. He was elected as the Vice Chair of Technical Activities of the IEEE Environmental Engineering Initiative (EEI) (2017–2022). He was the Founder and the Founding Chair (2011–2017) of the IEEE Technical Committee on Green Communications and Computing (TCGCC). Since 2022, he has been the Chair of the IEEE Technical Committee on Big Data (TCBD). He was the Co-Founder and the Founding Vice-Chair of the IEEE TCBD, in 2014 and from 2014 to 2022, respectively. He was the very first proposer of IEEE Green ICT Journals or Transactions, in 2012. He was a Proposer, in 2021, the Founder, in 2022, and the Founding Editor-in-Chief, since 2022, for the new international journal, *Green Technologies and Sustainability* (GTS), KeAi. He was the leading Editor and the coauthor of the comprehensive book, titled *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, September 2012).

HATHAL SALAMAH ALWAGEED received the Ph.D. degree in computer engineering from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2019. He has been with the College of Computer and Information Sciences, Jouf University, Aljouf, Saudi Arabia, since 2019. He is currently an Assistant Professor. His current research interests include machine learning, deep learning, the Internet of Things, and computer networks.



FAHEEM KHAN received the Ph.D. degree from the University of Malakand, Pakistan. He was an Assistant Professor, for four years, in Pakistan. He is currently an Assistant Professor with the Department of Computer Engineering, Gachon University, South Korea. His research interests include computer networking, wireless networks, MANET, VANET, the IoT, and artificial intelligence.

JAWAD KHAN received the master's degree in computer science from the Kohat University of Science and Technology (KUST), Pakistan, and the Ph.D. degree in computer engineering from Kyung Hee University (Global Campus), South Korea. He is currently a Postdoctoral Researcher with the Department of Robotics, Hanyang University (ERICA Campus), South Korea. His research interests include natural language processing, information retrieval, sentiment analysis/opinion mining, text processing, social media mining, and machine and deep learning.

YOUNGMOON LEE (Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University, South Korea, in 2014, and the M.S. and Ph.D. degrees in computer science and engineering from the University of Michigan, Ann Arbor, MI, USA, in 2016 and 2019, respectively. He is currently an Assistant Professor with the Department of Robotics, Hanyang University, South Korea. His research interests include cyber-physical systems, embedded systems, and mobile computing.