IEEE Open Journal of the Computer Society

Received 3 November 2023; accepted 12 November 2023. Date of publication 20 November 2023;
date of current version 28 December 2023. The review of this article was arranged by Associate Editor Xuan Song.

Digital Object Identifier 10.1109/OJCS.2023.3334528

A Real-Time 3-Dimensional Object Detection Based Human Action Recognition Model

CHHAYA GUPTA [1], NASIB SINGH GILL [1], PREETI GULIA [1], SANGEETA YADAV [1], GIOVANNI PAU [2] (Member, IEEE), MOHAMMAD ALIBAKHSHIKENARI [3] (Member, IEEE), AND XIANGJIE KONG [4] (Senior Member, IEEE)

[1]Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak 124001, India
[2]Faculty of Engineering and Architecture, Kore University, 94100 Enna, Italy
[3]Department of Signal Theory Communications, Universidad Carlos III Madrid, 28903 Getafe, Spain
[4]College of Computer Science and Technology, Zhejiang University, Hangzhou 310023, China

CORRESPONDING AUTHOR: GIOVANNI PAU (e-mail: giovanni.pau@unikore.it).

ABSTRACT Computer vision technologies have greatly improved in the last few years. Many problems have been solved using deep learning merged with more computational power. Action recognition is one of society's problems that must be addressed. Human Action Recognition (HAR) may be adopted for intelligent video surveillance systems, and the government may use the same for monitoring crimes and security purposes. This paper proposes a deep learning-based HAR model, i.e., a 3-dimensional Convolutional Network with multiplicative LSTM. The suggested model makes it easier to comprehend the tasks that an individual or team of individuals completes. The four-phase proposed model consists of a 3D Convolutional neural network (3DCNN) combined with an LSTM multiplicative recurrent network and Yolov6 for real-time object detection. The four stages of the proposed model are data fusion, feature extraction, object identification, and skeleton articulation approaches. The NTU-RGB-D, KITTI, NTU-RGB-D 120, UCF 101, and Fused datasets are some used to train the model. The suggested model surpasses other cutting-edge models by reaching an accuracy of 98.23%, 97.65%, 98.76%, 95.45%, and 97.65% on the abovementioned datasets. Other state-of-the-art (SOTA) methods compared in this study are traditional CNN, Yolov6, and CNN with BiLSTM. The results verify that actions are classified more accurately by the proposed model that combines all these techniques compared to existing ones.

INDEX TERMS CNN, feature extraction, human action recognition, multiplicative LSTM, skeleton articulation.

I. INTRODUCTION

Action identification in videos is one of the crucial ongoing issues in computer vision and artificial intelligence. For developing intelligent environments and cutting-edge security systems, action recognition in a live video is essential. It has uses in a variety of industries, including human-machine interface [1], monitoring systems [2], and visual comprehension [3]. Voluntary and non-voluntary activities taken by people are distinguished in human behavior [4]. Manually identifying these actions is challenging. For this reason, various strategies have been presented in the literature. The models suggested in the literature rely on conventional techniques, including geometric, point, texture, and shape features. Deep learning methods are employed to address the difficulty in HAR of distinguishing between various human actions. Layers such as convolution layers, pooling layers, ReLU, completely connected layers, dense layers, and SoftMax activation functions layers are just a few of the layers that deep learning uses to represent data [5] uniquely. Deep learning has many methods composed of supervised learning, unsupervised learning, hierarchical models, and probabilistic models. The training samples help evaluate the performance of any deep learning model [5].

Other significant challenges of HAR are: (i) focal point recognition in the current frame in a video sequence is a big challenge, (ii) lighting conditions in video sequences, shadows, occlusions, and background complexity impacts inefficient classification of actions, (iii) motion variations

capture wrong actions, and (iv) imbalanced datasets. Some of these challenges like occlusions, shadows in the videos, blurriness, background complexity and imbalanced datasets are considered in the present paper and rest are planned in our future work. The proposed model is evaluated on a real-time video taken from YouTube channel to classify the actions of people in an office and it shows that challenges like occlusions, blurriness, background complexity are taken care by the suggested model.

The research introduces a novel four-phase human action recognition model that utilizes object identification, skeleton articulation, and 3D convolutional network approaches that aid in resolving other key HAR difficulties, as mentioned above. The suggested approach creates 3DCNLSTM (3-dimensional Convolutional Network with LSTM) by combining multiplicative Long Short-Term Memory (LSTM) recurrent neural network with 3D CNN to process the videos. This method uses LSTM in conjunction with the multiplicative recurrent neural networks (mRNA) factorized hidden-to-hidden transition to assist in producing quick and effective results. In natural language processing, LSTM and mRNN are combined [6]. Classification is enhanced by incorporating feature extraction, object identification, and skeleton articulation techniques into the suggested model. The model's novelty lies in the combination of skeleton articulations of the person involved, classification of objects appearing in the scene, and extracted features from image sequences into a single neural network.

The KITTI dataset [7], the NTU-RGB-D and NTU-RGB-D 120 datasets [8], the UCF 101 dataset [9], and the fused dataset are used to assess the proposed model. The NTU-RGB-D dataset has 56880 videos of 60 classes, NTU-RGB-D 120 has 114480 video samples of 120 different classes, and UCF101 has 13320 videos of 101 different classes. The KITTI dataset has 7481 training images and 7518 validation images of 69 classes. The discussed datasets are combined to create a single set of data for training the suggested model. The suggested model is trained using 79282 pictures from 148 classes and 184680 video clips from 281 classes after the combined datasets. The motivation behind the work is to correctly classify input data from different video sequences into their activity category to enhance the video surveillance features and security systems. HAR plays a vital role in classifying various activities performed by subjects in videos.

The main objectives and concessions of the work are:

- Proposed a unique and novel four-phase model that combines four different modules into a single neural network. The model utilizes 3D CNN with multiplicative recurrent network LSTM and finetuned Yolov6 model for enhancing the classification and object detection process for actions in video sequences.
- Yolov6 itself is a novel model and not much research is executed on this, hence the finetuned and transfer learning-based Yolov6 model is used for object detection module in this proposed model along with multiplicative LSTM.
- Combining all the modules into a single neural network is a tedious task, hence the proposed model is a combination of data fusion with Dasarthy's technique, feature extraction with Xception V3 model merged with multiplicative LSTM, real-time object detection with finetuned Yolov6 [10] merged with multiplicative LSTM, and skeleton articulation technique.

The following is how the paper is set up: The relevant review is presented in Section II. Section III presents the proposed model design and deep learning models for feature extraction, object identification, and skeleton articulation techniques. The experiment's findings are presented in Section IV, and the study is wrapped up in Section V.

II. RELATED REVIEW

In recent years, HAR has emerged as a significant research area [11]. CCTV surveillance [12], the field of robotics [13], authentication [14], smart healthcare systems [15], and various other technologies are only a few of the many uses for HAR. For the recognition of human action, researchers created numerous deep learning models. The effectiveness and speed of deep learning attracted the attention of researchers. A deep neural network incorporating CNN and Bi-directional LSTM was proposed by Soni et al. [16] to recognize the human activity. The suggested model is tested on the UCI-HAR and UCI-WISDM datasets, and both datasets showed a 97.96% and 97.15% accuracy, respectively, for the model's performance. On the vanKasteren, CASAS Kyoto, and CASAS Aruba datasets, Patricia et al. [17] applied twelve classification approaches, including Logistic Regression, OneR, Attribute Selected, J48, Random Subspace, Random Forest, Random Committee, Bagging, JRip, Random Tree, and REP Tree. The study shows that logistic regression and OneR achieved an accuracy higher than 90%. Table 1 presents a detailed review of other methods used for HAR with their respective research gaps.

III. METHODOLOGY

A novel four-phase model has been proposed to improve Human Action Recognition considering existing methods. This approach's primary goal is to merge four distinct components into one integrated neural network. However, it is a tedious task to capture all human activities on one platform. Hence the proposed model tries to classify as many actions as it can. The proposed model's whole architecture is depicted in Fig. 1. Four phases make up the model. Data fusion is the first phase, wherein the already-existing datasets are combined to generate a new dataset. The second and third phase helps in extracting features and classifying those features according to skeleton articulations of the selected objects in an image. The fourth phase provides results. The proposed model combines four modules: data fusion, 3D CNN with multiplicative LSTM, object detection with finetuned Yolov6 and multiplicative LSTM, and skeleton articulation technique with multiplicative LSTM in a single neural network. The data fusion module is not shown in the figure as it is a step

TABLE 1. Summary of Literature Review With Techniques Used and Research Gaps

Reference	Publication Year	Title	Techniques Used	Research Gaps
[16]	2023	A Novel Smartphone-Based Human Activity Recognition Using Deep Learning in Healthcare	CNN, BiLSTM	Only two datasets are used to train and evaluate the suggested model.
[18]	2023	Evaluation of 2D and 3D posture for human activity recognition	Eight classification techniques, SVM, Random Forest, Decision tree, ANN, Naïve Bayes, KNN, Logistic Regression, Stochastic Gradient	The paper is more of a comparative analysis of different classification techniques on 2D and 3D human posture images. the techniques may not be applied to video sequences.
[17]	2023	Machine Learning Applied to Datasets of Human Activity Recognition: Data Analysis in Health Care	Logistic Regression, OneR, Attribute Selected, J48, Random SubSpace, Random Forest, RandomCommittee, Bagging, Random Tree, JRip, LMT, and REP Tree	These twelve classifiers are only effective for categorising human behaviours in picture datasets, not in video sequences.
[19]	2020	A Robust Feature Extraction Model for Human Activity Characterization Using 3-Axis Accelerometer and Gyroscope Data	Multi-class Support Vector Machine, Linear Discriminant Analysis, 3-fold cross-validation	Features are chosen at random, and datasets are not pre-processed.
[20]	2022	Human Activity Recognition Using an Ensemble Learning Algorithm with Smartphone Sensor Data	CNN, gated recurrent unit (GRU), categorical cross entropy	Activities like standing and sitting cannot be classified well. Time consumption is a limitation and hence the proposed algorithm cannot be used for real-time applications. The suggested algorithm is smartphone-dependent; HAR cannot be performed if the smartphone is charging or is not in the right location.
[21]	2022	Skeleton-based human activity recognition using ConvLSTM and guided feature learning	CNN+LSTM, Kinect (v2) sensor	Occlusions cannot be detected by the algorithm.
[22]	2022	A Real-Time Crowd Monitoring and Management System for Social Distance Classification and Healthcare Using Deep Learning	Yolov4 with Deepsort	The model is not able to detect occlusions and illusions.
[23]	2022	Human Activity Recognition Based on Residual Network and BiLSTM	Bi-directional LSTM, residual block	The accuracy of the model can be enhanced more in the future by using feature selection techniques.
[24]	2022	SSDT : Distance Tracking Model Based on Deep Learning	YoloV4, MFSORT, Kalman filter, the brute force technique	The model can be enhanced further for detecting occlusions and illusions.
[25]	2022	DTR-HAR: deep temporal residual representation for human activity recognition	Residual CNN, LSTM, Transfer learning,	For better architecture, the method can be used with several modalities including optical flow and depth mapping.
[26]	2022	Interpretable High-Level Features for Human Activity Recognition	Hidden Markov models (HMM), random forest, 10-fold cross-validation	Sequence extraction is the main challenge that may be overcome by using the sliding window approach in the future.
[27]	2022	3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information	3-dimensional fully connected CNN	The model can be enhanced for accuracy in the future.
[28]	2022	An information-rich sampling technique over spatiotemporal CNN for the classification of human actions in videos	3-dimensional CNN, spatiotemporal features, LSTM, Gaussian weighing function	The work is not tested on different datasets.
[29]	2022	View knowledge transfer network for multi-view action recognition	Conditional generative adversarial network (cGAN), transfer network	Although highly difficult to deploy, the network can enhance the efficiency of multi-view action recognition.
[30]	2022	Revisiting Skeleton-based Action Recognition Haodong	3-dimensional CNN, 3D heatmaps	The model is not tested on different datasets.
[31]	2021	A resource-conscious human action recognition framework using a 26-layered deep convolutional neural network	CNN, Extreme Learning Machine, Softmax	The hidden layer of ELM requires a complex structure due to the initialization of parameters that impacts the efficiency of the model.
[32]	2020	LSTM-CNN Architecture for Human Activity Recognition	CNN+LSTM, Average pooling	The model is not capable of reducing the dimensions of the input parameters.

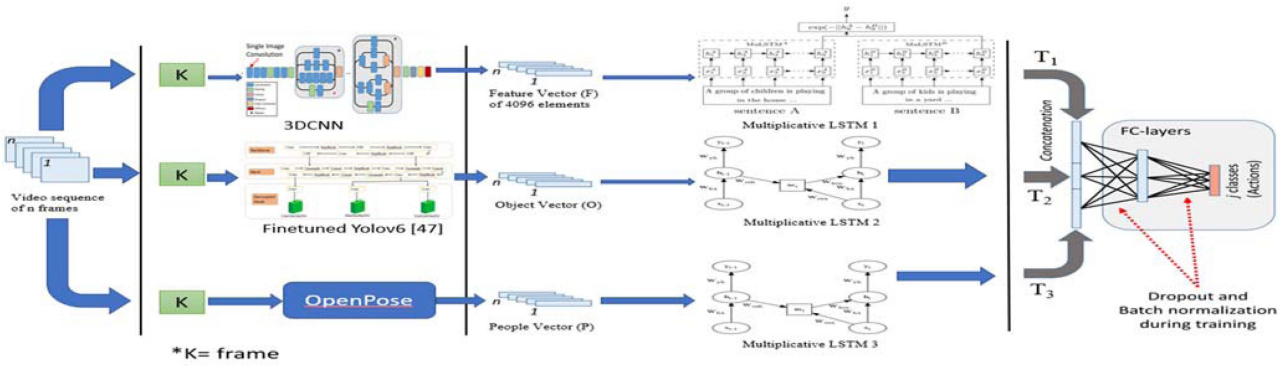

FIGURE 1. Architectural design of the suggested model 3DCNLST.

before pre-processing of data, where different datasets are merged to form a bigger dataset. All these modules, except data fusion, have their multiplicative LSTM layers. Single CNN with the LSTM model is not able to recognize specific actions. Hence, the need arises to merge four modules in a single neural network to identify human actions accurately. Video sequences with n frames, shown in Fig. 1, are divided into k frames and are passed to three different modules. The first k frames are passed to 3DCNN. In 3DCNN, the Xception module has been implemented with transfer learning for classification purposes. Xception neural network (XNN) is an 'extreme inception' model, which is a 71-layer deep neural network and is more efficient than the inception v3 model. Instead of compressing input data into discrete chunks before performing a 1×1 convolution to determine cross-channel correlation, XNN translates the spatial correlation for every output channel separately. Hence, Xception is a combination of depth-wise separable convolution and pointwise convolution. Transfer learning is a technique in which a model is initialized with weights from a pre-trained model like Xception and uses the model either as a feature extractor or a fine tuner for the last layers. In this study, Xception with transfer learning is used for better results [33]. The Xception module obtains convolution using the 1×1, 3×3, and 5×5 filter sizes. Convolutions are computed in parallel for all of them. Two further layers, max pooling, and average pooling, come before the completely connected layer. Utilizing the weights acquired during ImageNet training, the Xception module uses transfer learning. The Xception module consists of 4096 feature vectors. In this approach, 90 frames are obtained for each video with a frame rate of 35 Hz. Feature vectors are obtained for each frame, and the first multiplicative LSTM layer receives an input vector (F) of 90×4096 values. The first multiplicative LSTM provides an output in the form of a vector (T1) with other 4096 values that might represent the value of a sequence in input, but this output vector will be concatenated with outputs of other modules before pushing it to a fully connected layer. Feature maps obtained at different layers during Xception inference are shown in Fig. 2. In the second module, the finetuned YOLOv6 is used for object detection [10]. Yolov6 helps balance between speed and

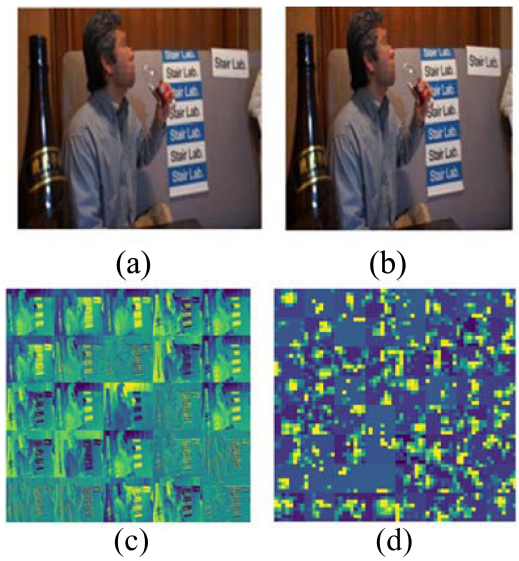

FIGURE 2. Feature maps of input images obtained during Xception with Transfer Learning (a) are the input image from the Fused dataset, (b) batch normalized image, (c) feature map at first layer, and (d) feature map at last 48th layer.

accuracy. Yolov6 works on Anchor-free paradigm that helps in increasing the speed by 51%. The finetuned Yolov6 has been used in this study that helps in reducing the challenges of HAR like occlusions, background complexity, blurriness to some extent [10]. Gupta et al. in [10] proposed a novel finetuned Yolov6 object detection model whose parameters are finetuned that helps in dimension reduction. Once the parameters are tuned, this reduces the model's accuracy. In order to enhance the reduced accuracy, transfer learning algorithm is proposed which enhances the model accuracy. This object detection method is given k frames as input. The vector in this module comprises 61 objects, each with six parameters that allow for the detection of confidence and bounding box positions. In this module, if comparable objects emerge, just one object is chosen, reducing the redundancy of objects in a single frame. In order to reduce redundancy, the object with the highest confidence score is chosen. These output values

undergo batch normalization, and the resulting output is supplied as input to the second multiplicative LSTM layer.

Before being sent to the fully connected layer, an output vector (T2) from the second multiplicative LSTM layer is concatenated with the outcomes of other modules.

In the third module, skeleton articulations of persons selected in the scene are computed. The k frames are provided as input to the OpenPose module. OpenPose module is a Python library embedded with the CNN module and trained with the COCO dataset. OpenPose returns output in the form of a heatmap, Part Confidence Maps (PCMs), and Part Affinity Fields (PAFs). The 90 frames are passed to this module, and OpenPose returns 18 coordinates with 135 key points of the two tallest persons selected in the given frame, forming an output vector (P) of $90\times2\times18$ values. This output forms the input for the third multiplicative LSTM layer. The output vector (T3) from the third multiplicative LSTM layer will be combined with the outputs of earlier modules before being pushed to the fully connected layer. Concatenating all of the multiplicative LSTM layers' output vectors ($T_1\|T_2\|T_3$), which is then used as the input to a fully connected layer of CNN. The action is classified into many classifications in the final stage. This section also provides a thorough explanation of each module utilised in the suggested model.

A. DATA FUSION

Data are raw facts that are not processed. After processing the data, it converts into helpful information. When working on HAR, it is impossible to classify each action a human performs. Humans are constantly performing some activities, such as if a person is sitting idle doing nothing, he is also conducting an act of sitting idle or standing idle. Hence, datasets are fused to form a larger dataset to analyze many actions. While merging datasets, it is necessary to understand that there will be redundant data and different data types. Using merge technique, datasets have been merged into two datasets, one with images of humans performing different actions and another with videos of human actions. The video database includes 184680 videos across 281 classes, and the picture dataset has 79282 images from 148 classes. This study uses Dasarathy's data fusion technique [34]. The technique is categorized into five categories:

- DAI – DAO (Data In – Data Out): In this, raw data is input and output. After the data is collected from the sensors, data fusion is carried out. The algorithms used are based on single-image processing.
- DAI – FEO (Data In – Feature Out): Features are extracted from the raw data which help describe the data.
- FEI – FEO (Feature In - Feature Out): In this, features are input and features are output. The data fusion technique is applied to features to refine them or to obtain new features.
- FEI – DEO (Feature In - Decision Out): At this level, characteristics are used as input, and the result is a set of judgements.
- DEI – DEO (Decision In – Decision Out): This level is known as decision-based fusion. This level helps in fusing decisions.

By using the data fusion technique, one of the challenges of HAR like imbalanced dataset is reduced to some extent. Imbalanced dataset is not helpful in class separation and evaluation and also results in poor model performance. Hence, fused dataset approach is used in this situation.

B. 3DCNN WITH MULTIPLICATIVE LSTM

The first module is 3D CNN with multiplicative LSTM (Fig. 1), a model that extracts features with CNN and provides them as input to layers of multiplicative LSTM. LSTM learns long-term and short-term dependencies. Before the dense layer, the LSTM layer gets the final result of the pooling layer as input. The CNN model performs convolutions on input with three filters: 1×1, 3×3, and 5×5. A 3-dimensional convolutional neural network (3DCNN) is similar to a 2-dimensional convolutional neural network, except in a 3DCNN, the kernel can slide in three directions, whereas in a 2DCNN, the kernel slides in two directions only. 3DCNN has two parts, a feature extractor and a classifier. 3DCNN uses a 3D filter to perform convolutional tasks, unlike the 2DCNN and produces 3D volume as the output of convolutions. By shifting filters vertically, horizontally, and across the depth of the input video frame or 3D picture, the layers of 3DCNN convolve the input. Multiplicative LSTM layers get the outcome of the classification layer as input. A multiplicative recursive neural network (mRNN) and LSTM are combined to create multiplicative LSTM [7]. Input gate a, Output gate b, and Forget gate c make up the three gates of an LSTM. Previously hidden state h_{t-1} and input layer x_t provide input to the next hidden states h_t of the LSTM, which is shown as:

$$\hat{h}_t = W_{hx}x_t + W_{hh}h_{t-1} \tag{1}$$

where, $\hat{h}_t$ = current hidden state, $W_{hx}x_t$ = weight of hidden state in input layer x, $W_{hh}h_{t-1}$ = weight of previous hidden state.

The three gates of LSTM, input gate a, output gate b, and forget gate c are stated as:

$$a_t = \sigma\left(W_{ax}x_t + W_{ah}h_{t-1}\right) \tag{2}$$

$$b_t = \sigma\left(W_{bx}x_t + W_{bh}h_{t-1}\right) \tag{3}$$

$$c_t = \sigma\left(W_{cx}x_t + W_{ch}h_{t-1}\right) \tag{4}$$

where σ = sigmoid function, W = weight vector.

The relationship between the components of the input gate and output gate determines what data should be stored and what data should be deleted at each transition. The input gate creates an internal state vector called d_t and decides how much input should be sent to each hidden unit. Forget gate c determines the amount of how much previous internal state d_{t-1} is preserved. The internal state is stated as:

$$d_t = c_t \odot d_{t-1} + b_t \odot \tanh\left(\widehat{h_t}\right) \tag{5}$$

The output gate b helps in preserving the relevant information which may not be helpful for the recent output but will be useful later. In (5), internal state vector d_t is XNOR operation of forget gate c and output gate b with previous internal state d_{t-1} and current hidden state $\hat{h}_t$. An intermediate state m_t from a multiplicative recurrent neural network is combined with each gate of LSTM forming multiplicative LSTM as:

$$m_t = (W_{mx}x_t) \odot (W_{mh}h_{t-1}) \tag{6}$$

$$\hat{h}_t = W_{hx}x_t + W_{hm}m_t \tag{7}$$

$$a_t = \sigma\,(W_{ax}x_t + W_{am}m_t) \tag{8}$$

$$b_t = \sigma\,(W_{bx}x_t + W_{bm}m_t) \tag{9}$$

$$c_t = \sigma\,(W_{cx}x_t + W_{cm}m_t) \tag{10}$$

The max pooling layer has been applied to the model before all of these convolutions are concatenated to improve the feature extraction strategy. The average pooling layer presents the main features of images before classification. Instead of using the output from the entire network for processing, this layer's output is utilized. An activation Softmax function, denoted as ş, is used to minimize the output vectors to real numbers between 0 and 1. This activation function helps in obtaining normalized distribution as shown in (11) and (12):

$$ş : \mathbb{R}^c \rightarrow [0, 1]^c \tag{11}$$

$$a = (a_1, \ldots.., a_c) \rightarrow ş\,(a) = \left(ş_1, \ldots\ldots, ş_c\right) \tag{12}$$

where c is classes and a is actions.

To obtain probabilities of each applied action a in classes c, using softmax function is depicted in the formula as:

$$ş_j = \frac{e^{a_j}}{\sum_{k=1}^{c} e^{a_k}} 1 \leq j \leq c \tag{13}$$

Each video sequence in the proposed model is processed at a frame rate of 35 Hz over the course of 90 frames. Each frame's 4096-element feature vector is obtained before being sent to the LSTM. The input is normalized using the batch normalization method to scale the pixel values between -1 and 1.

C. OBJECT DETECTION WITH FINETUNED YOLOV6 AND MULTIPLICATIVE LSTM

A real-time object identification technique called You-Only-Look-Once (YOLO) version 6 uses CNN for identifying objects in pictures and videos. YOLOv6 is a high-performing, single-stage detector with an effective design. YOLOv6 performs better than all of the earlier iterations of YOLO in terms of both accuracy and inference speed. This paper uses a hidden layer pruning approach that reduces the total number of parameters and the network depth of YOLOv6, making it a lightweight network. Following model pruning, there is a decrease in detection accuracy. Using the optimized YOLOv6 network, a transfer learning technique was applied to improve the detection accuracy [10]. On YOLOv6, the head is detached. A network with a decoupled head signifies that the head part has more layers, contributing to improved performance. The decoupled head section receives the neck information directly and uses it for simultaneous objectness, classification, and regression tasks. Three components comprise the YOLOv6 model: the neck, the decoupled head, and the backbone. During the training phase, YOLOv6 employs reparameterized VGG blocks with skip conditions. The COCO dataset is used to train YOLOv6 [35]. Unlike YOLO, the finetuned Yolov6 model uses two loss functions as Verifocal Loss and Distribution Loss. The varifocal loss function is used for classification and for box regression, distribution loss function is used. Verifocal loss function uses BCE (Binary Cross Entropy). Distribution Loss depends on the probability of the target box as discussed in [10].

FIGURE 3. Different frames from video sequences.

The finetuned YOLOv6 processes at a speed of 70 Hz, and the video sequences are limited to 32 FPS, but this does not slow the process. This module forms a vector of 61 objects with six parameters per object. If objects of the same type with different confidence scores appear in an image, the object with the highest confidence score is selected. For humans appearing in the image, the tallest humans are chosen. After processing 90 frames with finetuned YOLOv6, the total processing is calculated as $90*(61*6) = 90*366$ values. These data are batch-normalized to an image's height and width. Fig. 3 shows some of the frames of videos representing different actions, such as PullUps, Biking, HulaHoop, Skiing, Playing Guitar, JavelinThrow, and others.

D. SKELETON ARTICULATION TECHNIQUE WITH MULTIPLICATIVE LSTM

The last module corresponds to the skeletons of the tallest humans involved in a scene. The structure in this study uses 18 coordinates which OpenPose returns [36]. A real-time, multi-user human pose identification toolkit written in Python called OpenPose uses 135 key points to identify the human body.OpenPose comprises a CNN model trained on the COCO dataset. These skeletons are helpful when there is a need to see the movement of one or two people in 90-frame sequences. The frames help diagnose the bounding boxes around people moving in a scene. Hence there is no need to translate the image. For instance, it makes no difference where a person walks—in the middle of the room or close to a window—because a bounding box will follow him wherever

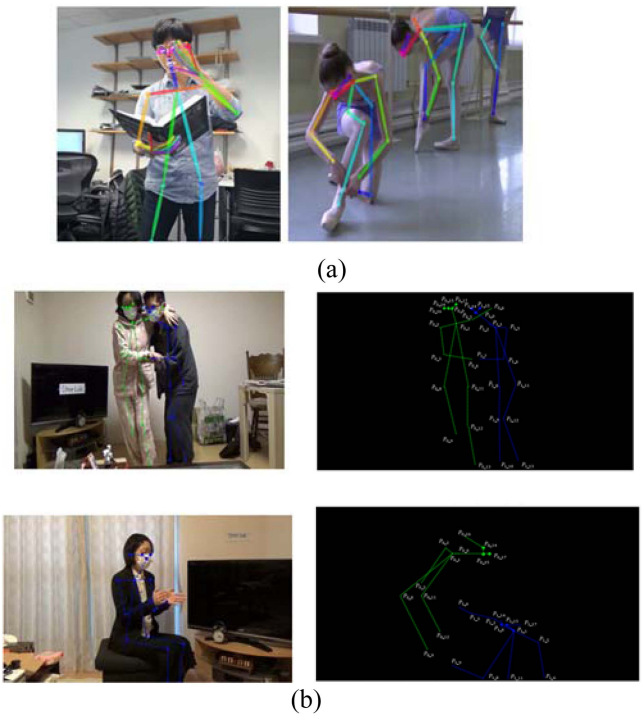

FIGURE 4. (a) OpenPose skeletons for different frames of different video sequences from the fused dataset, (b) estimation of pose on fused dataset.

he goes. The skeleton joints are adjusted to the bounding box's height and width. Fig. 4(a) shows frames with skeleton articulations corresponding to different action classes from videos fused datasets, such as reading a book or belle dance. Fig. 4(b) depicts the skeleton articulations of fused datasets when executed through the proposed model.

In the first step, the image is passed via 3D CNN architecture, which extracts feature maps. Part Confidence Maps (PCMs) and Part Affinity Fields (PAF) are created by further processing these feature maps [37]. Finally, PCMs and PAFs are further processed by a bipartite algorithm that helps to generate the skeletons.

Any body part that can be found in any pixel is represented in a 2D confidence map. Confidence map (C) is computed as:

$$\mathrm{C} = \left(\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 \ldots .., \mathrm{C}_j\right) \text{ where } \mathrm{C_j} \in \mathrm{R}^{\mathrm{w*h}}, \mathrm{j} \in 1 \ldots .\mathrm{j} \tag{14}$$

where j = the no. of body parts locations.

PAF is computed as follows:

$$\mathrm{P} = (\mathrm{P}_1, \mathrm{P}_2, \ldots \ldots, \mathrm{P}_\mathrm{x}) \text{ where} \mathrm{P_x} \in \mathrm{R}^{\mathrm{w*h*x}}, \mathrm{x} \in 1 \ldots .\mathrm{x} \tag{15}$$

The difference in loss among PCM and PAF is also calculated using an L2-Loss function as:

$$f_L = \sum_{t=1}^{P} f_C^t + \sum_{t=P+1}^{P+C} f_P^t \tag{16}$$

$$f_C^{t_i} = \sum_{C=1}^{C} \sum_{p} W\,(p)\,.\,||L_C^{t_i}\,(p) - \,L_C^*\,(p)\,||_2^2 \tag{17}$$

$$f_P^{t_k} = \sum_{j=1}^{J} \sum_{P} W\,(p)\,.\,||\,P_j^{t_k}\,(p) - \,P_j^*||_2^2 \tag{18}$$

where L_C^* = ground truth value for Part Affinity Field, P_j^* = ground truth value for Partial Confidence Map, W = binary mask with W(p) = 0 and it helps in preventing the extra loss.

E. COMBINING DATA FUSION, 3DCNN WITH MULTIPLICATIVE LSTM, OBJECT DETECTION WITH FINETUNED YOLOV6 AND MULTIPLICATIVE LSTM AND SKELETON ARTICULATION TECHNIQUE WITH MULTIPLICATIVE LSTM

Datasets are combined with the help of data fusion techniques, as discussed. The proposed model 3DCNLSTM receives input with 90 frames covering all the activities. Each module processes each frame to obtain a set of features. All the relevant features are filtered and normalized and have 61 objects with coordinates of two humans and feature vectors of 4096 elements which CNN obtains. LSTM helps in processing the generated features and helps in reducing the dimensions of the data. LSTM provides output in the form of three vectors, that is, F(f0, …..,f2047), O0(x0, y0, x1, y1) … …O60(x0, y0, x1, y1), and P(x0, y0, ….., x17, y17) where F is feature vector, O is movement vector, and P is person vector. As indicated in (21), all of these results are concatenated and given as a single input T to a completely connected layer. Dropout layers are also added to the model to avoid the problem of overfitting. The rest of the section describes the working of the model mathematically:

$$\mathrm{T} = \sum \mathrm{T_i} \quad \text{where i} = 1, 2, 3 \tag{19}$$

Let Df be a vector produced by a 3D CNN model for f frames with dimensions (8∗8∗2048). The vector is pooled by the average pooling layer (Ff = avg(Df)). Each video sequence is divided into 90 frames. Hence these 90 frames are transformed into vectors of size (90∗2048), that are given as an input to the LSTM layer. This first LSTM layer produces an outcome vector as:

$$\mathrm{T_1} = \mathrm{mLSTM}\,(\mathrm{vector}\,(\mathrm{F_0}, \ldots ., \mathrm{F_{89}})) \tag{20}$$

The object detection module provides an output vector Zf with dimensions (b, rx, ry, x), where b represents the bounding boxes, rx, and ry are several grids that consist of objects, and x is the output from finetuned YOLOv6. This module also provides another vector as [ax, ay, w, h, con], where ax and ay are positions of an object, and con is the confidence score concerning width (w) and height (h). this vector helps in identifying the things with the highest confidence score. This vector output acts as a filtering process that helps to produce 61 other object vectors as O0(x0, y0, x1, y1) … …O60(x0, y0, x1, y1) for a single frame 'f'. This second LSTM layer

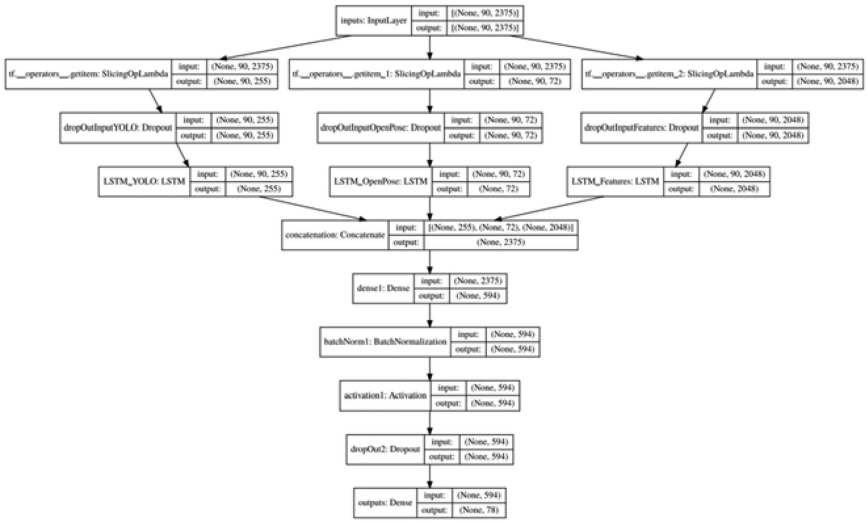

FIGURE 5. Proposed 3DCNLSTM architecture.

TABLE 2. Outcomes on the UCF101 Dataset [38]

Metrics	CNN	YOLOv6	CNN+BiLSTM	DTR-HAR	**Proposed 3DCNLSTM**
Training Accuracy	0.9657	0.9619	0.9645	0.9034	**0.9678**
Training Loss	0.0187	0.0234	0.0056	0.0067	**0.0034**
Validation Accuracy	0.9445	0.9534	0.9534	0.9045	**0.9545**
Validation Loss	0.6876	0.7135	0.6978	0.7256	**0.6645**
Test Accuracy	0.9356	0.9467	0.9465	0.9004	**0.9578**

produces another vector with dimensions as:

$$T_2 = \text{mLSTM}(\text{vector}\,(O_0, \ldots., O_{89}) \tag{21}$$

The Skeleton articulation module produces a vector Sf having dimensions (h, k, t), where h is the number of humans detected, k is the number of joints in a body part, and t is data, [x, y, con], where (x, y) are coordinates of joints, and con is confidence score value. This vector is batch normalized to produce vectors of two persons as Hf = [person0(x0,y0, …..x17,y17), person1(x0,y0, … …x17,y17)]. The third and final LSTM layer processes this input and produces a vector as:

$$T_3 = \text{mLSTM}(\text{vector}\,(P_0, \ldots., P_{89}) \tag{22}$$

All these vectors generated by LSTM are concatenated (T1|T2|T3) and then given as input to a fully connected model layer. Fig. 5 depicts the data flow through the suggested paradigm. In this paper, the proposed model is tested, and all the other models are tested separately to check the efficiency of each module used.

TABLE 3. Outcomes on the KITTI Dataset [7]

Metrics	CNN	YOLOv6	CNN+BiLSTM	DTR-HAR	**Proposed 3DCNLSTM**
Training Accuracy	0.8680	0.4456	0.9796	0.9123	**0.9953**
Training Loss	0.0189	0.0234	0.0234	0.0345	**0.0156**
Validation Accuracy	0.9234	0.9567	0.9715	0.9134	**0.9834**
Validation Loss	0.7865	0.3456	0.4434	0.4891	**0.2567**
Test Accuracy	0.9767	0.9887	0.9786	0.9154	**0.9823**

TABLE 4. Outcomes on the NTU-RGB-D Dataset [8]

Metrics	CNN	YOLOv6	CNN+BiLSTM	DTR-HAR	**Proposed 3DCNLSTM**
Training Accuracy	0.8767	0.6789	0.9754	0.8923	**0.9898**
Training Loss	0.0234	0.0456	0.0345	0.0567	**0.0123**
Validation Accuracy	0.9456	0.9556	0.9678	0.8991	**0.9763**
Validation Loss	0.3456	0.5678	0.3245	0.5467	**0.3156**
Test Accuracy	0.9434	0.9578	0.9654	0.8945	**0.9765**

TABLE 5. Outcomes on the NTU-RGB-D 120 Dataset [8]

Metrics	CNN	YOLOv6	CNN+BiLSTM	DTR-HAR	**Proposed 3DCNLSTM**
Training Accuracy	0.8767	0.6789	0.9787	0.9001	**0.9898**
Training Loss	0.0234	0.0456	0.0145	0.0567	**0.0123**
Validation Accuracy	0.9454	0.9553	0.9765	0.9023	**0.8963**
Validation Loss	0.3454	0.5678	0.2345	0.5546	**0.3456**
Test Accuracy	0.9434	0.9578	0.9715	0.8990	**0.9876**

TABLE 6. Outcomes on the Fused Dataset

Metrics	CNN	YOLOv6	CNN+Bi LSTM	DTR-HAR	**Proposed 3DCNLSTM**
Training Accuracy	0.7767	0.7789	0.9745	0.9017	**0.9898**
Training Loss	0.0334	0.0446	0.0231	0.0534	**0.0123**
Validation Accuracy	0.9456	0.9656	0.9556	0.8990	**0.9663**
Validation Loss	0.3356	0.5578	0.4535	0.5674	**0.3256**
Test Accuracy	0.9434	0.9578	0.9712	0.9012	**0.9765**

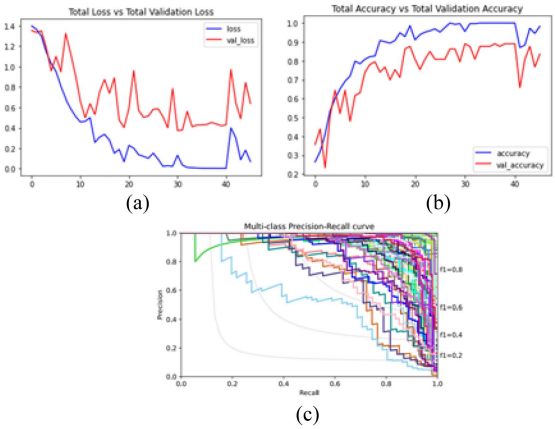

FIGURE 6. (a) Total loss and total validation loss of proposed model (b) total accuracy and total validation accuracy of the proposed mode (c) precision-recall curve for multiple-classes.

TABLE 7. Performance Evaluation of All the Models Over Different Datasets

Model	Avg. Accuracy(%)	Avg. Precision (%)	Avg. F1-score(%)	Avg. Recall(%)
CNN+BiLSTM	95.78	92.34	92.45	93.24
CNN	84.85	87.67	84.56	76.78
YOLOv6	95.56	93.45	95.76	94.78
DTR-HAR	90.23	89.97	89.92	88.79
Proposed 3DCNLSTM	**96.78**	**93.78**	**95.89**	**95.45**

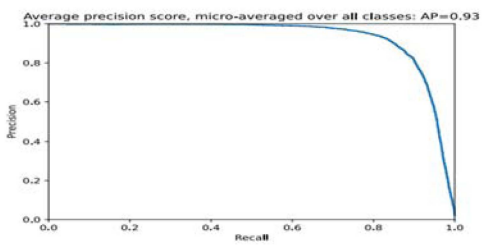

FIGURE 7. Precision-recall curve with average precision and recall scores of the suggested model.

TABLE 8. Classes Considered for Evaluation

Class	Action
Assisting	Getting up, walking, holding another person, passing something
Baby Crying/Crawling	Baby is crawling or crying
Drinking	Drinking something
Eating	Eating or cutting food or meal
Exercising	clapping, bowing, jumping, boxing, shaking, bending
Massage	Massaging head, body
Moving	Entering or moving out of room or lawn
Opening something	Opening or closing something
Personal care	Brushing, drying hair, manicuring
Playing	Playing some game
Putting cloths	Putting on or off clothes
Reading/writing	Reading or writing on paper
Watching TV/Rest	Doing rest or watching TV
Sewing	Sewing something
Standing/smoking	Smoking while standing
Throwing	Throwing something
Listening music	Listening music on headphone
Washing something	Washing dish, washing face
Wearing	Wearing glasses, gloves, shoes

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this study, different datasets NTU-RGB-D, KITTI, NTU-RGB-D 120, and UCF 101, have been explored and fused to form a new dataset. The NTU-RGB-D dataset has 56880 videos of 60 classes, NTU-RGB-D 120 has 114480 video samples of 120 different classes, and the KITTI dataset has 7481 images for training and 7518 testing images of 69 classes. The UCF101 dataset has 13320 videos of 101 different classes. All the datasets have other action classes, and each class has approximately 800 videos featuring each class, which is an excellent number to train any model. Fusion techniques combined these datasets to form a single dataset with 79282 images belonging to 148 classes and 184680 videos of 281 classes. The model has just one fully linked layer, which is adequate to produce good results. The inputs, as well as the outputs of this model, affect how the neurons in this layer act. Layers with dropouts were employed to prevent the issue of overfitting. The experiment was run for 50, 100, 200, and 500 epochs, with 500 epochs producing outstanding results. Table 2 shows the outcomes of different models used in this study compared with other models for 500 epochs on the UCF101 dataset. Additionally, the suggested model is contrasted with the cutting-edge CNN with the Bi-directional LSTM model, and the findings demonstrate that the proposed model outperformed it. The work cited in this paper is based on IoT modules and smartphones. In this study,

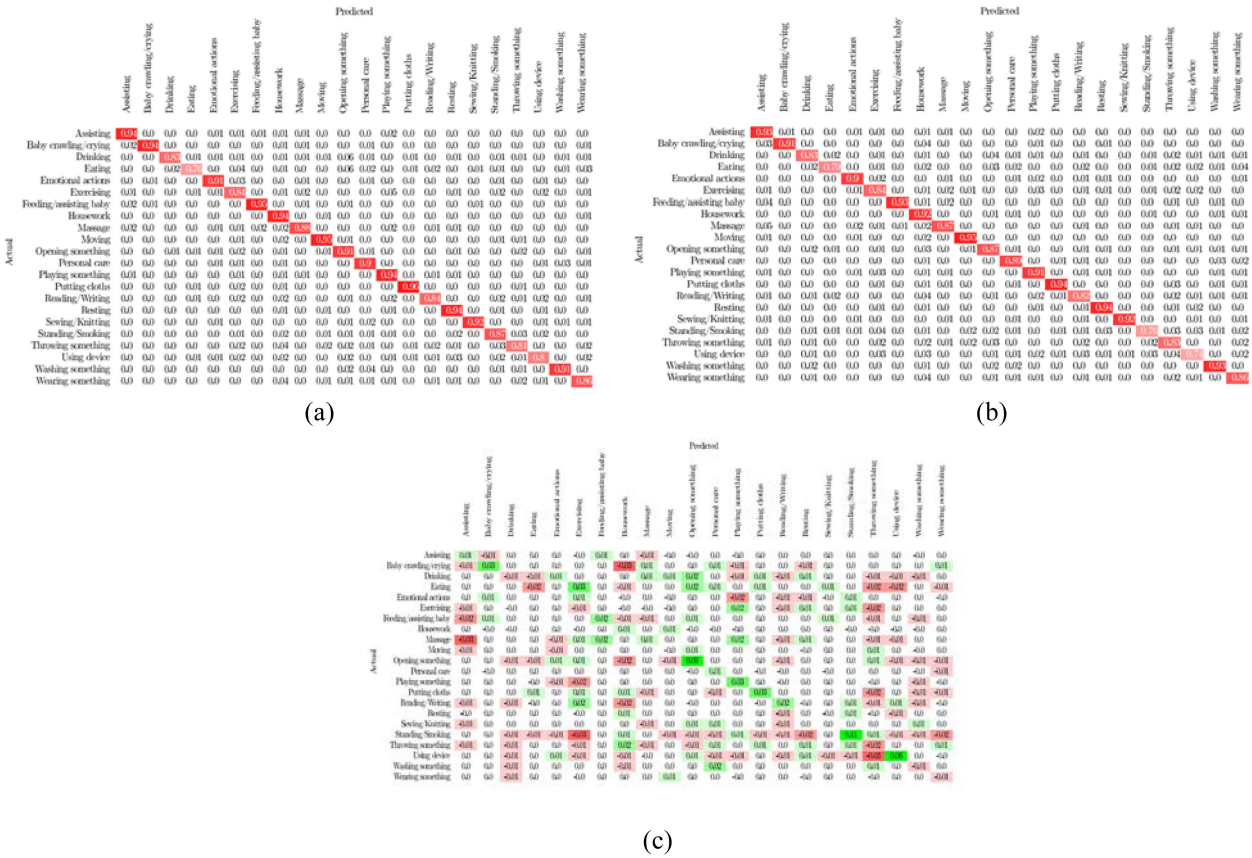

FIGURE 8. **(a) Confusion matrix of the proposed model for training, (b) confusion matrix of the proposed model for testing, and (c) confusion matrix showing improvement in each class prediction.**

only the neural network (composed of CNN and BiLSTM) is implemented with the different datasets used for the experiment, and the results are captured. The proposed model is also compared with a deep residual convolutional neural network for human activity recognition [25]. This DTR-HAR model is trained on all the described datasets in this study and results showed that DTR-HAR achieved an average accuracy of 90%. Table 3 depicts the results of various models used in this study compared with other models for 500 epochs on the KITTI dataset. Table 4 depicts the results of various models used in this study compared with other models for 500 epochs on the NTU-RGB-D dataset. Table 5 depicts the results of various models used in this study compared with other models for 500 epochs on the NTU-RGB-D 120 dataset. Table 6 shows the outcomes of various models used in this study compared with other models for 500 epochs on the Fused dataset.

Fig. 6(a), (b), and (c) presents the proposed model's total loss vs. total validation loss and total accuracy vs. total validation accuracy as well as the precision-recall curve for multiple classes.

All the evaluations have been obtained at each activity level. An average of all the accuracies, precision, F1 score, and Recall is calculated for evaluating the performance of different models on different datasets. Table 7 shows the results, and Fig. 7 shows the Precision-Recall curve with precision and recall scores.

A few classes are taken into consideration in order to streamline the outcomes and make them clear, making the confusion matrix simple to understand. The classes considered are shown in Table 8. To appropriately display the confusion matrix, classes are condensed. The confusion matrix is shown in Fig. 8(a) at training time and in Fig. 8(b) at testing time. The confusion matrix shown in Fig. 8(c) shows the improvement in each class prediction with the proposed model.

The proposed model is tested in real-time environment also. For real-time analysis, 3DCNLSTM is tested on YouTube video and the evaluated results in the form of action recognition are shown in Fig. 9. The figure depicts the pose estimations along with action recognitions in different situations. Action set has been selected randomly for testing in real-time environment. The video belongs to a workspace where three people are walking, running, bending, jumping and kicking sometimes and it is clear from Fig. 9 that the proposed model identifies the human actions with pose estimations well. The results showed that the proposed model

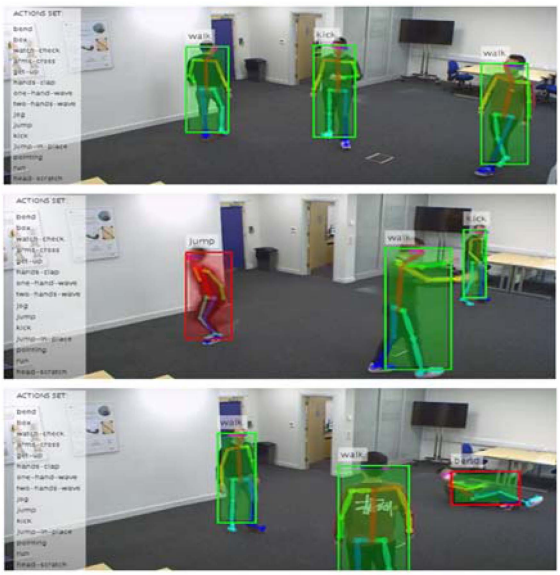

FIGURE 9. Real-time analysis of proposed 3DCNLSTM model on a random video from YouTube.

has overcome some of the challenges of HAR such as occlusions, blurriness and background complexity. It can be seen that the objects present in background are not creating confusion between human action, hence improving the background complexity problems. Further, it can be seen clearly that the person bending in one of the images is blurred to naked eyes, but the proposed model is able to detect the exact action of the person, hence reducing occlusion as well as blurriness problem.

V. CONCLUSION

The article presents a combination of different techniques in a single neural network named 3DCNLSTM. The study analyses other techniques as well. Several datasets have been used, namely, UCF101, NTU-RGB-D, KITTI, and NTU-RGB-D 120, for comparing various techniques with the proposed model. The data fusion techniques are also used to form a fused dataset that consists of 79282 images belonging to 148 classes and 184680 videos of 281 classes. Classic CNN, YOLOv6, CNN with BiLSTM, and DTR-HAR are also trained on these datasets and are compared with the proposed model. The suggested model outperforms existing models with the highest accuracy. The proposed model combines four techniques, namely, data fusion, feature extraction, object detection, and skeleton articulation techniques in a single neural network. Multiplicative LSTM has been applied to improve the suggested model's effectiveness. In the future, the model can be enhanced further by using deeper convolutional networks for better feature extraction, and this model may be integrated with humanoids. This model can be used to track the activities of old age people living alone in their homes and hence can be used as an assistant for them.

REFERENCES

[1] Y. Liu, C. B. Sivaparthipan, and A. Shankar, "Human–computer interaction based visual feedback system for augmentative and alternative communication," *Int. J. Speech Technol.*, vol. 25, no. 2, pp. 305–314, 2022, doi: 10.1007/s10772-021-09901-4.

[2] A. Ravat, A. Dhawan, and M. Tiwari, *Advances in VLSI, Communication, and Signal Processing*. Berlin, Germany: Springer-Verlag, 2021.

[3] M. A. Khan et al., "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimedia Tools Appl.*, vol. 79, pp. 1–27, 2020, doi: 10.1007/s11042-020-08806-9.

[4] B. E. Arikan, B. M. van Kemenade, B. Straube, L. R. Harris, and T. Kircher, "Voluntary and involuntary movements widen the window of subjective simultaneity," *i-Percep.*, vol. 8, no. 4, 2017, Art. no. 2041669517719297, doi: 10.1177/2041669517719297.

[5] S. Kiran et al., "Multi-layered deep learning features fusion for human action recognition," *Comput., Mater. Continua*, vol. 69, no. 3, pp. 4061–4075, 2021, doi: 10.32604/cmc.2021.017800.

[6] B. Krause, I. Murray, S. Renals, and L. Lu, "Multiplicative LSTM for sequence modelling," in *Proc. 5th Int. Conf. Learn. Representations*, 2019, pp. 1–11.

[7] "kitti dataset | Kaggle." Accessed: Mar. 01, 2023. [Online]. Available: https://www.kaggle.com/datasets/klemenko/kitti-dataset

[8] "ROSE lab." Accessed: Nov. 25, 2022. [Online]. Available: https://rose1.ntu.edu.sg/dataset/actionRecognition/

[9] "CRCV | center for research in computer vision at the University of Central Florida." Accessed: Mar. 01, 2023. [Online]. Available: https://www.crcv.ucf.edu/data/UCF101.php

[10] C. Gupta, N. S. Gill, P. Gulia, and J. M. Chatterjee, "A novel fine-tuned YOLOv6 transfer learning model for real-time object detection," *J. Real-Time Image Process.*, vol. 20, no. 3, 2023, Art. no. 54, doi: 10.1007/s11554-023-01299-3.

[11] M. A. Khan, M. Sharif, T. Akram, M. Yasmin, and R. S. Nayak, "Stomach deformities recognition using rank-based deep features selection," *J. Med. Syst.*, vol. 43, no. 12, 2019, doi: 10.1007/s10916-019-1466-3.

[12] C. She, R. Zheng, Q. Yang, and S. Liang, "Research of intelligent video surveillance system based on artificial neural network," *J. Phys. Conf. Ser.*, vol. 2181, no. 1, 2022, Art. no. 012057, doi: 10.1088/1742-6596/2181/1/012057.

[13] M. Hellou, J. Y. Lim, N. Gasteiger, M. Jang, and H. S. Ahn, "Technical methods for social robots in museum settings: An overview of the literature," *Int. J. Soc. Robot.*, vol. 14, no. 8, pp. 1767–1786, 2022, doi: 10.1007/s12369-022-00904-y.

[14] G. Aquino, M. G. F. Costa, and C. F. F. C. Filho, "Explaining one-dimensional convolutional models in human activity recognition and biometric identification tasks," *Sensors*, vol. 22, no. 15, 2022, Art. no. 5644, doi: 10.3390/s22155644.

[15] F. Serpush, M. B. Menhaj, B. Masoumi, and B. Karasfi, "Wearable sensor-based human activity recognition in the smart healthcare system," *Comput. Intell. Neurosci.*, vol. 2022, 2022, Art. no. 1391906, doi: 10.1155/2022/1391906.

[16] V. Soni, H. Yadav, V. B. Semwal, B. Roy, D. K. Choubey, and D. K. Mallick, "A novel smartphone-based human activity recognition using deep learning in health care," in *Proc. Mach. Learn., Image Process., Netw. Secur. Data Sci.: Select 3rd Int. Conf. MIND*, 2023, pp. 493–503, doi: 10.1007/978-981-19-5868-7_36.

[17] A.-C. P. Patricia et al., "Machine learning applied to datasets of human activity recognition: Data analysis in health care," *Curr. Med. Imag. Former. Curr. Med. Imag. Rev.*, vol. 19, no. 1, pp. 46–64, 2023, doi: 10.2174/1573405618666220104114814.

[18] M. A. Hossen, A. G. Naim, and P. E. Abas, "Evaluation of 2D and 3D posture for human activity recognition," in *Proc. Amer. Inst. Phys. Conf.*, 2023, vol. 2643, no. 1, Art. no. 040013, doi: 10.1063/5.0111224.

[19] R. A. Bhuiyan, N. Ahmed, M. Amiruzzaman, and M. R. Islam, "A robust feature extraction model for human activity characterization using 3-axis accelerometer and gyroscope data," *Sensors*, vol. 20, no. 23, pp. 1–17, 2020, doi: 10.3390/s20236990.

[20] T. H. Tan, J. Y. Wu, S. H. Liu, and M. Gochoo, "Human activity recognition using an ensemble learning algorithm with smartphone sensor data," *Electronics*, vol. 11, no. 3, pp. 1–17, 2022, doi: 10.3390/electronics11030322.

[21] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "Skeleton-based human activity recognition using ConvLSTM and guided feature learning," *Soft Comput.*, vol. 26, no. 2, pp. 877–890, 2022, doi: 10.1007/s00500-021-06238-7.

[22] S. Yadav, P. Gulia, N. S. Gill, and J. M. Chatterjee, "A real-time crowd monitoring and management system for social distance classification and healthcare using deep learning," *J. Healthcare Eng.*, vol. 2022, 2022, Art. no. 2130172, doi: 10.1155/2022/2130172.
[23] Y. Li and L. Wang, "Human activity recognition based on residual network and BiLSTM," *Sensors*, vol. 22, no. 2, pp. 1–18, 2022, doi: 10.3390/s22020635.
[24] C. Gupta, N. S. Gill, and P. Gulia, "SSDT : Distance tracking model based on deep learning," *Int. J. Elect. Comput. Eng. Syst.*, vol. 13, no. 5, pp. 339–348, 2022, doi: 10.32985/ijeces.13.5.2.
[25] H. Basly, W. Ouarda, F. E. Sayadi, B. Ouni, and A. M. Alimi, "DTR-HAR: Deep temporal residual representation for human activity recognition," *Vis. Comput.*, vol. 38, no. 3, pp. 993–1013, 2022, doi: 10.1007/s00371-021-02064-y.
[26] Y. Hartmann, H. Liu, S. Lahrberg, and T. Schultz, "Interpretable high-level features for human activity recognition," *Biosignals*, vol. 4, pp. 40–49, 2022, doi: 10.5220/0010840500003123.
[27] A. Sánchez-Caballero et al., "3DFCNN: Real-time action recognition using 3D deep neural networks with raw depth information," *Multimedia Tools Appl.*, vol. 81, no. 17, pp. 24119–24143, 2022, doi: 10.1007/s11042-022-12091-z.
[28] S. H. S. Basha, V. Pulabaigari, and S. Mukherjee, "An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos," *Multimedia Tools Appl.*, vol. 81, no. 28, pp. 40431–40449, 2022, doi: 10.1007/s11042-022-12856-6.
[29] Z. Liang, M. Yin, J. Gao, Y. He, and W. Huang, "View knowledge transfer network for multi-view action recognition," *Image Vis. Comput.*, vol. 118, 2022, Art. no. 104357, doi: 10.1016/j.imavis.2021.104357.
[30] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2022, vol. 2022, no. 1, pp. 2959–2968, doi: 10.1109/CVPR52688.2022.00298.
[31] M. A. Khan, Y. D. Zhang, S. A. Khan, M. Attique, A. Rehman, and S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network," *Multimedia Tools Appl.*, vol. 80, no. 28/29, pp. 35827–35849, 2021, doi: 10.1007/s11042-020-09408-1.
[32] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020, doi: 10.1109/ACCESS.2020.2982225.
[33] Rismiyati, S. N. Endah, Khadijah, and I. N. Shiddiq, "Xception architecture transfer learning for garbage classification," in *Proc. IEEE 4th Int. Conf. Inform. Comput. Sci.*, 2020, pp. 1–4, doi: 10.1109/ICICoS51170.2020.9299017.
[34] F. Castanedo, "A review of data fusion techniques," *Sci. World J.*, vol. 2013, 2013, Art. no. 704504, doi: 10.1155/2013/704504.
[35] C. Li et al., "YOLOv6: A single-stage object detection framework for industrial applications," Sep. 2022, *arXiv:2209.02976*.
[36] C. C. Huang and M. H. Nguyen, "Robust 3D skeleton tracking based on openpose and a probabilistic tracking framework," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2019, pp. 4107–4112, doi: 10.1109/SMC.2019.8913977.
[37] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE 30th Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310, doi: 10.1109/CVPR.2017.143.
[38] "CRCV | center for research in computer vision at the University of Central Florida." Accessed: Nov. 25, 2022. [Online]. Available: https://www.crcv.ucf.edu/data/UCF101.php

CHHAYA GUPTA received the M.C.A. degree from Guru Gobind Singh Indraprastha University, Delhi, India. She is currently working toward the Ph.D. degree in computer science with the Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India. She is currently an Assistant Professor with the Vivekananda Institute of Professional Studies-TC, Delhi. Her main research interests include machine learning, deep learning, object detection, and pattern recognition.

NASIB SINGH GILL received the M.B.A. degree and the Ph.D. degree in computer science in 1996. From 2001 to 2002, he held Postdoctoral research in computer science with Brunel University, West London, London, U.K. He is currently the Head of the Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India. He is also the Director of the Directorate of Distance Education and Director of the Digital Learning Centre, Maharshi Dayanand University. He has authored or coauthored more than 304 research papers indexed in SCI, SCIE, and Scopus and authored five popular books He has guided so far 12 Ph.D. scholars and guiding about five more scholars. His research interests primarily include IoT, machine and deep learning, information and network security, data mining and data warehousing, NLP, and measurement of component-based systems. He was the recipient of the Commonwealth Fellowship Award of the British Government for the Year 2001. He is an active professional member of IETE, IAENG, and CSI.

PREETI GULIA received the Ph.D. degree in computer science in 2013. She is currently an Associate Professor with the Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India. She has been serving the Department since 2009. She has authored or coauthored more than 80 research papers indexed in SCI, SCIE, and Scopus and presented papers at national and international conferences. She has guided four scholars and guided five more scholars. Her research interests include data mining, big data, machine learning, deep learning, IoT, and software engineering. She is an active professional member of IAENG, CSI, and ACM. She is also an Editorial Board Member and Active Reviewer of International/ National Journals.

SANGEETA YADAV, photograph and biography not available at the time of publication.

GIOVANNI PAU (Member, IEEE) received the bachelor's degree in telematic engineering from the University of Catania, Catania, Italy, and the master's (*cum Laude*) and the Ph.D. degree in telematic engineering from the Kore University of Enna, Enna, Italy. He is currently an Associate Professor with the Faculty of Engineering and Architecture, Kore University of Enna. He is the author/coauthor of more than 95 refereed papers published in journals and conference proceedings. His research interests include wireless sensor networks, fuzzy logic controllers, intelligent transportation systems, the Internet of Things, smart homes, and network security. He has been involved in several international conferences as the session co-chair and a technical program committee member. He is/was the leading guest editor for special issues of several international journals. He is an Editorial Board Member and an Associate Editor for several journals, such as IEEE ACCESS, *Wireless Networks* (Springer), *EURASIP Journal on Wireless Communications and Networking* (Springer), *Wireless Communications and Mobile Computing* (Hindawi), *Sensors* (MDPI), and *Future Internet (MDPI)*, to name a few.

MOHAMMAD ALIBAKHSHIKENARI (Member, IEEE) was born in Mazandaran, Iran, in February 1988. He received the Ph.D. degree (Hons.) with a European Label in electronics engineering from the University of Rome "Tor Vergata," Rome, Italy, in February 2020. He was a Ph.D. Visiting Researcher with the Chalmers University of Technology, Göteborg, Sweden, in 2018. His training during the Ph.D. study included a research stage in the Swedish company Gap Waves AB. He is currently with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid (uc3m), Getafe, Spain, as the Principal Investigator of the CONEX (CONnecting EXcellence)-Plus Talent Training Program and Marie Skłodowska-Curie Actions. From 2021 to 2022, he was a Lecturer of the electromagnetic fields with Electromagnetic Laboratory, Department of Signal Theory and Communications. He is spending an industrial research period with SARAS Technology Ltd., Company, Leeds, U.K., defined as his secondment plan by CONEX-Plus Program and Marie Skłodowska-Curie Actions. His research interests include electromagnetic systems, antennas and wave propagations, metamaterials, and metasurfaces, synthetic aperture radars, multiple input–multiple output systems, RFID tag antennas, substrate-integrated waveguides, impedance-matching circuits, microwave components, millimeter-waves and terahertz integrated circuits, gap waveguide technology, beamforming matrices, and reconfigurable intelligent surfaces. He was the recipient of the three-year research grant funded by Universidad Carlos III de Madrid and the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant in 2021, the two-year research grant funded by the University of Rome "Tor Vergata" in 2019, the three years Ph.D. Scholarship funded by the University of Rome "Tor Vergata" in 2016, and the two Young Engineer Awards of the 47th and 48th European Microwave Conference held in Nuremberg, Germany, in 2017, and Madrid, Spain, in 2018, respectively. His research article titled High-Gain Metasurface in Polyimide On-Chip Antenna Based on CRLH-TL for Sub Terahertz Integrated Circuits (Scientific Reports) was awarded as the Best Month Paper at the University of Bradford, U.K., in 2020. He was also the recipient of the "Teaching Excellent Acknowledgement" Certificate for the course of electromagnetic fields from the Vice-Rector of Studies of uc3m. He also acts as a referee in several highly reputed journals and international conferences. He is an Associate Editor for *Radio Science* and *IET Journal of Engineering*.

XIANGJIE KONG (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, and the School of Software, Dalian University of Technology, Dalian, China. He has authored or coauthored more than 160 scientific papers in international journals and conferences. His research interests include mobile computing, network science, and data science.

Open Access funding provided by 'Università degli Studi di Enna "KORE"' within the CRUI CARE Agreement