Spatio-Temporal Features based Human Action Recognition using Convolutional Long Short-Term Deep Neural Network

A F M Saifuddin Saif, Ebisa D. Wollega, Sylvester A. Kalevela
School of Engineering, Colorado State University Pueblo, CO 81001, USA

***Abstract*—Recognition of human intention is crucial and challenging due to subtle motion patterns of a series of action evolutions. Understanding of human actions is the foundation of many applications, i.e., human robot interaction, smart video monitoring and autonomous driving etc. Existing deep learning methods use either spatial or temporal features during training. This research focuses on developing a lightweight method using both spatial and temporal features to predict human intention correctly. This research proposes Convolutional Long Short-Term Deep Network (CLSTDN) consists of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). CNN uses Inception-ResNet-v2 to classify object specific class categories by extracting spatial features and RNN uses Long Short-Term Memory (LSTM) for final prediction based on temporal features. Proposed method was validated on four challenging benchmark dataset, i.e., UCF Sports, UCF-11, KTH and UCF-50. Performance of the proposed method was evaluated using seven performance metrics, i.e., accuracy, precision, recall, f-measure, error rate, loss and confusion matrix. Proposed method showed better results comparing with existing research results. Proposed method is expected to encourage researchers to use in future for real time implications to predict human intentions more robustly.**

***Keywords*—Convolutional neural network; recurrent neural network; long short-term memory; human action recognition**

I. Introduction

The recognition and understanding of human activities on video streams have now become crucial in many applications for instance smart video monitoring, automobiles driving, somatic gaming, etc. This task is extremely difficult if both accuracy and robustness are taken into consideration. In the field of human behavior recognition, substantial research is needed. Recognition and prediction of actions are two key tasks in the field of computer vision and action recognition. Understanding the actions of others is the foundation of human social interaction. It is difficult to predict the intent of others from their acts but it's necessary. Recognition of human activities plays a significant role in people's daily lives, for example in applications for medical, protection, and law enforcement fields. Observation of human intentions through motions was introduced to many instances of Human-Robot Interaction. Recognition of actions and prediction of actions can be very different among the different classes of action. Intention prediction, however, infers from the subtle motion patterns of a series of action evolutions of the same action. This makes the prediction of intention a more challenging task. Recognition of action can essentially be categorized at various abstraction levels depending on the nature of the visual information. It varies from simple actions like activity concepts, interaction with objects and human beings to complex acts as a group activity.

In recent years, one of the popular deep learning models, Convolutional Neural Networks (CNNs), has shown great success in many computer vision's tasks, such as image recognition, image segmentation, object tracking and so on [1] [2]. CNN appears to learn a hierarchy of features from low-level to high-level and researchers find the features that CNN automatically learns are typically better than the handcrafted features. Researchers have put a great effort to develop neural networks capable of capturing spatial-temporal features in recognition of human activities. Most of the researchers have taken advantage of the deep learning approach for the recognition of human activities. Because deep learning techniques allow automated extraction and learning of hierarchical features by human behavior recognition systems, several of these systems have been developed and have shown promising results. Deep Learning (DL) methods have gained significant attention recently due to their remarkable performance in various fields. It is, therefore, not surprising that DL-based methods for identification, prediction and prediction of intention have also increased. Particularly Deep Learning models based on the Recurrent Neural Network have brought much success in the field of behavior analysis in recent years. The most widely used model in RNN is usually the Long Short-Term Memory (LSTM). It is an extension of the RNN structure that essentially allows long-term temporal dependencies by replacing hidden nodes with gated memory cells.

Focusing on that, this research proposes a Deep Learning approach for recognizing human intention. In videos with complex scenarios, proposed method called Convolutional Long Short-Term Deep Network (CLSTDN) can identify human intentions with a good amount of precision. Pre-trained convolutional neural network, Inception-ResNet-v2 is used for extracting spatial features from video sequence and then using LSTM temporal features are extracted from the video sequence. With some more dense layers, training and classification are done to recognize intention from videos. In this task, this research evaluates proposed method on the human action datasets such as UCF Sports dataset, UCF-11 dataset, UCF-50 dataset and KTH dataset where variety of actions of different situations was found. These datasets are

very complex and challenging because of wide variations in camera motion, conditions of lighting, object appearance and posture, size of object, view, cluttered background, etc. Regardless of the point of view, background, inter-class and intra-class similarities are found in image frames, proposed method performed very well with very small data sequences in all the datasets. Experiment results show that the proposed method achieved state-of-the-art performance with low computational resources for action intentions recognition from human behavior. Overall contribution of this research is summarized as below.

1) This research proposes a news method called Convolutional Long Short-Term Deep Network (CLSTDN) consists of a Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for efficient human intention prediction. Proposed CLSTDN serves following two purposes:

a) Convolutional Neural Network (CNN) uses Inception-ResNet-v2 to classify scenes to categorize class specific objects by extracting spatial features.

b) Recurrent Neural Network (RNN) uses Long Short-Term Memory (LSTM) for final prediction based on temporal features.

Thus, this research ensures to use spatial and temporal features for improved human intention prediction. Previous research methods used either spatial or temporal features for human intention prediction; however, proposed CLSTDN used spatial-temporal features for human intention prediction.

2) Massive experimental results are demonstrated on four benchmark datasets to validate the proposed method. Four publicly available datasets were used for validation, i.e., UCF Sports, UCF-11, KTH and UCF-50. Seven evaluation metrics were estimated for each dataset, i.e., accuracy, precision, recall, f-measure, error rate, loss and confusion matrix.

3) Proposed CLSTDN performed efficiently based on all the evaluation metrics with limited number of data sequence irrespective of viewpoint, background, inter-class and intra-class similarities present in the image frames in all the datasets which is very promising comparing with previous research methods.

4) Previous research methods provide the evidence that deep learning architectures require huge computational power and GPU clusters which imply the fact that it is wise to consider not only the accuracy but also the cost of a method for implementation. By considering this fact to implicate new research methodology with limited computation resources in lieu of achieving satisfactory computation time, this research proposed Convolutional Long Short-Term Deep Network (CLSTDN) to recognize the human action based on intention.

Rest of this research is organized as follows, Section II illustrates existing research methods, Section III reflects proposed research methodology, Section IV reflects comprehensive experimental results for robust validation of the proposed methodology and finally, conclusion section presents concluding remarks.

II. Previous Research Study

There are essentially two types of human activity recognition approaches, i.e., video-based models and sensor-based models. High-dimensional features are derived from videos or images [3][5][10][13][14][35], whereas sensor-based systems rely on motion information captured by sensing devices [44]. Researchers have focused on analyzing human behavior through radar backscattering echoes with the development of new sensing technology. It is understood that the carrier frequency of the radar signal would be changed when reflected from any moving target which is known as the Doppler Effect. A point network can learn more efficiently about structural features from the micro motion trajectory than it can directly process the raw point cloud and it also shows that the temporal range-Doppler PointNet approach performs better on most behaviors [18]. This uses radar to emit signals, while very high frequency reflected signal analysis captures very fine dynamic behavior. Radar can be used for extreme climates, meaning that it is immune to light and weather conditions, as this system uses radar, human beings can be identified via walls, making radar useful in more situations. For researchers, the hierarchical model is superior to its base counterpart (P-Net). Sensor-based technologies may be a trigger for mental and physical discomfort [43] and wearable technology is not ideal for applications where complex motor activity needs to be monitored and interpreted [15][23][24][25][26][27].

YOLOv3 object detector [8] and SSD [29] can accurately detect a group of people with a high degree of confidence by only taking a few frames from video-based models. There is also a pipeline of Generative Adversarial Networks (GANs), which jointly learns latent information for estimating and group identification of pedestrian trajectories [2]. GAN suggests a learning mechanism for task-specific loss function where a minimum game between the generative and discriminatory models is an objective of training. In the area of computer vision, significant progress has been made concerning recent developments in deep learning methodologies due to the advancement of deeper CNN, parallel computer hardware and wide-ranged annotated datasets. Various approaches have been put on places to address the recognition of action problems. Actions recognition systems can be classified into representations based on shape and appearance, representations based on optical flow and representations based on the point of interest depending on the representations of the feature. The most successful methods of classifying the action involve using CNNs. Deep CNN models achieved state-of-the-art results with tasks of object identification, classification, generation and segmentation. In deep learning, there are two primary strategies used to extract features from the video frames. First, by expanding traditional 2D CNN architectures to 3D, 3D CNN learns convolution kernels in space and time domains [17]. The second strategy is a two-stream CNN [55] that delivers state-of-the-art results in [56][57][58]. A Convolutional Neural Network (CNN) with YOLO is used together for an IoT detection that enables suspicious human behavior with minimal effort [20]. A Deep Neural Network (DNN) is followed by a time-domain classifier method for automatic violent behavior detection designed for video sensor

networks [40]. One distinguishing characteristic of the proposed solution is that it relies entirely on motion characteristics by completely disregarding appearance details. Asymptotically the tests are like optical flow-based approaches and often better than the others. This method is ideally suited for low computing devices and used by researchers for improvement in the future. Researchers achieved state-of-the-art efficiency, using Raspberry-Pi sensor nodes to run on low computational embedded architecture. This approach works perfectly well with low-computing devices and a potential area for enhancement work. Research in [22] compared state-of-the-art machine learning and deep learning approaches suitable for detecting early changes in human behavior, where support vector machine (SVM) and Convolutional neural network (CNN) from the supervised method, One-class support vector machine (OCSVM) and Stacked auto-encoders (SAE) from Semi-Supervised and K-means clustering (KM) and Convolutional auto-encoder (CAE) from Unsupervised is used. Methods based on CNN take many parameters for model learning. This takes hours to months to optimize. In the meantime, using a GPU with parallel architecture significantly reduces the processing time. The more GPUs, the less computational time it takes to train. These factors have given much importance to the transition of transfer learning [11] [12] [32] [38] [41]. Researchers proposed a pre-trained model of a convolutional neural network (CNN) based on Visual Geometry Group network 19 (VGGNet-19) which is used to extract descriptive features [41]. An abnormal event detection system using pre-trained VGGNet-19 and Binary Support Vector Machine (BSVM) videos demonstrates higher detection accuracy than other pre-trained networks: GoogleNet, ResNet50, AlexNet, and VGGNet-16. BSVM needs only a few abnormal detection training patterns and provides reasonable detection accuracy for new patterns with the same features. Several researchers use Deep convolutional neural networks (CNN) to detect violent scenes by transfer learning to classify aggressive human behaviors [28][31][50]. Research in [30] recognized basic human activities using the Deep Belief Network (DBN) method which is a good candidate to the model activity recognition system. DBN is a robust, deep learning method used during training using Restricted Boltzmann Machines (RBMs). At first, they extracted efficient features from the raw data. Then they used kernel principal component analysis (KPCA) and linear discriminant analysis (LDA) to make the data more robust. Research in [16] proposed a Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields (PAFs). The OpenPose program recognizes the key human points better than other methods. The program identifies and generates a student report every three seconds in the classroom. It utilizes a non-parametric representation called PAFs to learn to connect the body components with individuals in the image. Via this approach, the OpenPose system recognizes the human key points better than other approaches. This also operates on the low computational device slowly, leading to a lack of input in real-time and considering only six gestures. Long-term memory networks (LSTM) are another technology which is used in various approaches like in [32] to predict human movements accurately. To send visual signals, they used a deep learning model that explored combining CNN with LSTM. LSTM is used to derive temporal patterns of human motion outputting the prediction result immediately before movement occurs. LSTM is used by research in [45] for the simulation of spatial-temporal sequences obtained from smart home sensors. They mentioned that approaches based on LSTM yield higher results than the existing DL and ML methods. For pattern recognition, LSTM is also used for anomaly detection with the Recurrent Neural Network (RNN) and Multi-Layer Perceptron (MLP) [21,60]. Different researchers used models to explain human emotion, sentiment, stress, and fatigue using GRU [6][9] and LSTM [19]. GRU is LSTM-like, which has shown that it operates well on smaller datasets. GRU has less operation compared to LSTM and thus it can be trained much faster than LSTMs, while LSTM is more accurate in the long sequence datasets. A method was proposed by research in [43] to identify such behaviors in which humans interact with various objects, considering object-oriented knowledge of the operation, using a hybrid approach to combine deep convolutional neural networks with multi-class support vector machines (multi-class SVM). An Adaptive feature recalibration residual network (AFRRNet) and Quaternion Spatio-Temporal Convolutional Neural Network (QST-CNN) based model is proposed by research in [47] to recognize human behavior. In predicting human action or behavior, the temporal network works much better than other related networks. It also makes use of optical flow representation for the input flow. The pre-processing of the optical flow image corresponds to the spatial flow. Optical flow suppresses horizon details and displays series motion fields. The motion fields highlighted encode motion information between adjacent frames which contribute significantly to the prediction of intention. This essentially improves the ability of the network to derive functionality and accuracy of recognition of behaviors.

Research in [35] proposed a multi-stream model for a better understanding of human activities through 3-channel Depth MHI and Skeleton based ST-GCN. The model defines contextual awareness, global and local intervention recognition motions. These two models have a fusion performance that exceeds each model and is comparable with state-of-the-art results. Researchers proposed an architecture with the multi-stream convolutional neural network (HR-MSCNN) based on a human-related region that encodes the presence, motion and captured tubes of regions with human relations [36]. The improved version of B-RPCA (IB-RPCA) can be defined reliably as the main actor in complex realistic circumstances including vibration, specific luminous conditions and partial occlusions. Human emotions also make a significant contribution to understand human intention. A deep learning approach is focused on the multimodal detection of stress through Convolutional Autoencoders and Recurrent Neural Networks. This also contained a recurrent unit called Gated Recurrent Unit (GRU) [9]. Some researchers use a lightweight CNN model for recognition of facial expression from a given input image [7]. CNN generates a matrix using the input image. After that the pooling process takes place then, the flattening process starts. Finally, the system predicts the facial expression. Using Convolutional Neural Network (CNN) fused with Extreme Learning Machines (ELMs), it can

classify emotions where two layers of the ELM to the fusion make calculation fast. It is found that the fusion based on ELM performed better than the combination of the classifiers [17]. With respect to local feature descriptors, there has been a lot of work to extract and explain useful and robust information. Several feature descriptors have been successfully adapted to enhance the accuracy of human action recognition from the image domain to the video domain. In addition to human-robot interaction, prediction of human intention plays an important role in a wide variety of applications, such as driving assistance systems on cars to predict the lane changes intentions of drivers [4] [34], prediction of pedestrian or cyclist intention [49], as well as monitoring to predict the intentions underlying detected suspicious activities and giving security [20][21].

There were many methods and frameworks with various types of limitations or drawbacks that are needed to overcome to develop a perfect method for human activity recognition. Many of the methods and frameworks were simple to implement and develop but offered considerable computational time. The more GPUs there are, the less computational time it needs for training. However, the solution is costly. It is wise to consider not only the accuracy but also the cost of a method when choosing a method for any implementation, as this can influence the time to build a production-ready method and the operating costs for running it.

III. Proposed Research Methodology

Previous research provides the evidence that deep learning architectures require huge computational power and GPU clusters to perform in low computational time. By considering this fact to implicate new research methodology with limited computation resources in lieu of achieving satisfactory computation time, this research proposed Convolutional Long Short-Term Deep Network (CLSTDN) to recognize the Human Action based Intention. Proposed method used both spatial and temporal features from a video and predicted intentions robustly. Pre-trained convolutional neural network (CNN), Inception-Resnet-v2 extracts spatial features from a video frame and makes a feature sequence. After that, features are given as input into the LSTM network. LSTM network extracts the temporal features from the sequence and with dense layer and SoftMax activation function, proposed method predicts an intention from the video. Overall proposed method was very resource friendly as this research did not use any external GPUs and only 8 gigabytes of RAM. Later, proposed method was validated with real-world environment videos and performed robustly to identify human intentions.

Proposed CLSTDN by this research consists of a convolutional neural network and a Long-short-term Memory network. Proposed methodology consists of three main parts, i.e., data preprocessing, feature extraction and intention recognition. This research used video-based datasets to train and test the method. In the data processing section, frames are extracted from the video data first and then reshaped those images to 224 x 224. For KTH dataset, this research calibrated images into 120 x 120 as the image dimension of the KTH dataset was smaller than the other datasets used by this research. Second significant part of the proposed methodology is feature extraction. This research used a pre-trained Inception-ResNet-v2 model to extract an image encoding. In this section, this research stacks the frames to provide a three-channel image to meet the specifications of Inception-ResNet-v2 and after that, this research reshaped images into 299 x 299 and made sequences by extracting features from the image frames for training and testing purpose. The third part of the proposed methodology is training and testing the model from the extracted features using LSTM deep learning method. Here, this research fed the extracted spatial feature into the LSTM model and extracts temporal features and the last layer output using the SoftMax activation function. Overall proposed methodology is shown in Fig. 1 and comprehensively explained in the subsequent sections.

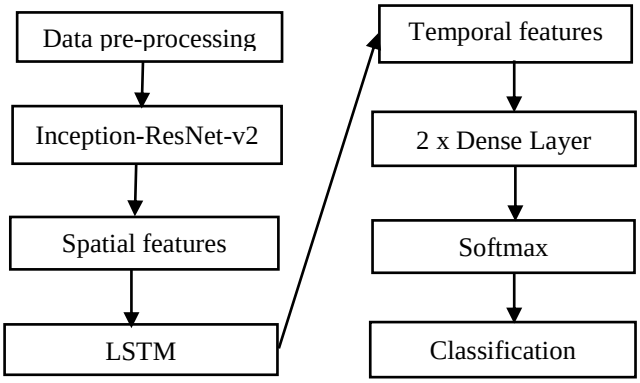


Fig. 1. Components of proposed methodology.

A. Data Processing

This research used UCF Sports, UCF11, UCF50 and KTH dataset to train and test proposed CLSTDN. As these datasets are video-based dataset so this research first extracted image frames from these datasets followed by splitting the data into train and test data and saved each video's number of frames into a CSV file by which extraction of spatial features were done by Inception-ResNet-v2 model. As imbalanced dataset may lead to wrong accuracy and over-fitting problems, this research removed blank image frames from the dataset. After that, this research reshaped images size to decrease computation time. Next, these images are used as input to Inception-Resnet-v2 to extract features.

B. Backbone

To utilize depth features as guidance of 2D convolutions, this research formulates backbone as a two-branch network: the first branch is the feature extraction network using RGB images and the other is the filter generation network to produce convolutional kernels for feature extraction network using the estimated depth as input. These two networks process the two inputs separately and their outputs of each block are merged by the depth guided filtering approach.

The backbone of the feature extraction network is Inception-ResNet-v2[54] without its final FC and pooling layers and is pre-trained on the ImageNet classification dataset [63]. To obtain a larger field-of-view and keep the network stride at 16, this research finds the last convolutional layer (conv5 1, block4) that decreases resolution and set its stride to

1 to avoid signal decimation and replace all subsequent convolutional layers with dilated convolutional layers (the dilation rate is 2). For the filter generation network, we only use the first three blocks of Inception-ResNet-v2 to reduce computational costs. Note the two branches have the same number of channels of each block for the depth guided filtering task. More about Inception-ResNet-v2 is illustrated in the next section.

C. Inception-ResNet-V2

For extracting spatial features from input images, this research used pre-trained Inception-ResNet-v2 model. This architecture is trained on more than a million images from the ImageNet database. The network is 164 layers deep and can classify images into 1000 object categories. Inception-ResNet-v2[54] is from the Inception family convolutional neural network (CNN) architecture which gradually evolved from GoogleNet and was the first Inception neural network through Inception v2 and Inception v3, and finally, Inception v4 [54]. Every one of the architectures brought changes primarily with respect to the Inception module, Inception networks building block. Research in [54] motivated by the results obtained by residual neural networks, experimented with integrating residual connections, key features of residual networks, with the Inception module. The result, Inception-ResNet-v2 is a very deep network, which in a lower number of epochs can achieve high accuracy which motivates proposed method by this research to use Inception-ResNet-v2 model for spatial feature extraction.

Inception-ResNet-v2 architecture contains three separate types of Residual Inception modules, called A, B, and C, and two distinct blocks of reduction. In Residual Inception module, residual connections play significant role for overall manipulation. Residual connection is a simple concept that has been invented as a way of solving the problems of the vanishing gradient and exploding gradient that may appear in a very deep neural network. The theory behind is that residual block only measures the adjustments that will make the input perfect, apply them to the input, and present it as its output, instead of calculating the output from scratch. Residual Inception blends residual learning with the Inception modules by incorporating a residual connection to it. The concept on which the Inception module operates is to expand rather than deeper neural network. The expanded width enables complex patterns to be recorded at different scales. The initial 1x1 convolutions are only used to reduce dimensionality on the axis of the channel to lighten the following convolutions. All convolutions use zero padding for the preservation of height and width. This is important since the outputs are concatenated along the depth dimension after each separate calculation in the module, which would not be possible if the heights and widths differed. The split of a n×n convolution into two -1×n and n×1 convolution stored one on top of each other is another feature of the Inception module. An important feature of the Residual Inception module is that after the concatenation, there is a 1x1 convolution to allow the addition of identity passed by the residual relation and the actual output by aligning its dimensions [54]. Inception-ResNet-v2 contains reduction modules which are slightly changed Inception modules. Unlike those mentioned earlier, these modules contain an average pooling on one of the branches followed by a 1x1 convolution. This research extracted spatial features from the Average Pooling layer of Inception-ResNet-v2 architecture (Fig. 1 Top) and fed it as an input to the LSTM model for training the model.

D. Training LSTM Model

After extracting the spatial features, this research uses Long Short-Term Memory (LSTM) network to extract temporal features from a sequence. LSTM is a recurrent neural network (RNN) architecture designed more accurately than conventional RNNs for modeling temporal sequences and long-term dependencies. In the recurring hidden layer, LSTM has special units known as memory blocks. Fig. 2 shows that memory blocks contain a memory cell with self-connections to store the network's temporal state, as well as special multiplication units called information flow gates. An input gate and an output gate were included in each memory block of the initial architecture. Input Gate controls the memory cell flow of the activations. Output gate controls cell activation output flow into the rest of the network. Forget gate was added to the memory block later. The forgotten gate measures the inner cell state before connecting it to the cell through the self-recurring connection of the cell, thereby forgetting or resetting the cell's memory adaptively.

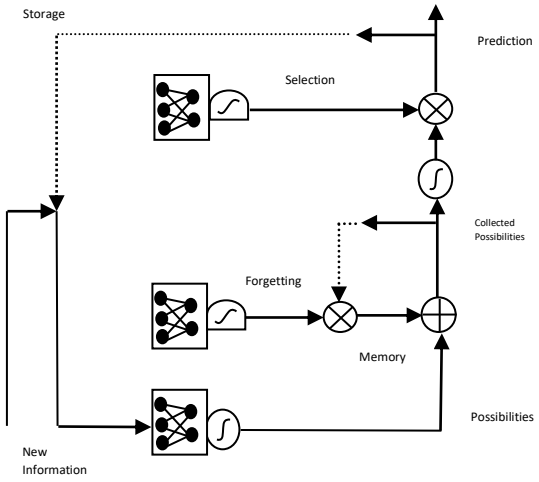


Fig. 2. Long Short-Term Memory (LSTM) network.

For training purpose this research used Long Short-Term Memory Deep Learning approach. During implementation, this research changed learning rate, decay rate, dropout and momentum to find the optimal value of these hyper parameters. A low learning rate helps overall methodology to avoid overfitting problem, where proposed method performs well on both training and test data. In this context, this research tuned parameters into four steps, i.e., tuning LSTM size, tuning batch size, tuning number of epochs and tuning for LSTM and Dense layers for four different datasets.

1) Tuning for LSTM size: This research initially fixed number of LSTM and Dense layers to 3, batch size to 32 and number of epochs to 50 and tuned LSTM size over these parameters. This research trained LSTM model with size 512,

1024 and 2048 for every dataset used for experimentation. With size 2048, proposed method generated the highest accuracy for tuning LSTM size. This research also used a fixed dense layer of size 512 and a classification dense layer to classify overall result.

2) Tuning batch size and no. of epochs: After tuning LSTM size, this research fixed its size to 2048 and number of LSTM and Dense layers to 3 and tuned batch size and number of epochs. This research performs training with a batch size of 8, 16, 32, 64, 128, 256 and number of epochs of 20, 50, 75 and 100. This research performs training for each set of batch sizes and number of epochs respectively. Here, this research found different batch sizes and number of epochs sets for different datasets which are explained in more details in experimental results section.

3) Tuning No. of LSTM and dense layers: For the final step of parameter tuning, proposed method fixed LSTM size to 2048 and different batch sizes and number of epochs for the different datasets and tuned number of LSTM and Dense layers. In this context, this research performs training with one fixed LSTM layers and 2, 3, and 4 Dense layers. This research also tuned the model by changing dense layer size of 256, 512 and 1024. Proposed method generated the best result for 3 layers with different sizes for the different datasets.

In summary, this research proposed a new method called Convolutional Long Short-Term Deep Network (CLSTDN) to recognize human action-based intention. Proposed CLSTDN consists of convolutional neural network and recurrent neural network. For convolutional neural network, this research used pre-trained network which was Inception-ResNet-v2 which is trained on more than a million images from ImageNet database. This network contains 164 layers to classify images into 1000 object categories. For recurrent neural network, this research used Long Short-Term Memory (LSTM) network. Overall proposed methodology consists of three main parts, i.e., data preprocessing, feature extraction, intention recognition. At first, video frames were extracted followed by preprocessing frames according to Inception-ResNet-v2. Then, images are passed to the Inception-ResNet-v2 network to extract the spatial features and create a sequence of the video. After that, the sequences are passed to the LSTM network which extracts temporal features from the sequence. Then dense layers and SoftMax activation function are used in the proposed method to predict the intention from the video. This research performs training with low computational resources by using both spatial-temporal features. Proposed method predicted intentions accurately which was revealed robustly during experimental validation demonstrated in the next section. In addition, implication of the proposed method is resource friendly.

IV. Experimental Results and Analysis

Evaluation for the proposed Convolutional Long Short-Term Deep Network (CLSTDN) was done by experimenting four publicly available datasets, i.e., UCF Sports, UCF-11, KTH and UCF-50. This research plotted actual and predicted data for each intention into a confusion matrix after experimentation. From the confusion matrix, this research estimated True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) which helped to find out the evaluation metrics. This research estimated accuracy, precision, recall, f-measure and error rate for each dataset. After 62 epochs, this research received best result for UCF Sports dataset which gives an accuracy of 95.74%, precision of 95.83%, recall of 97.49%, F-measure of 96.23%, and an error rate of 4.26%. After 43 epochs, proposed method received best results for UCF 11 dataset which gives an accuracy of 95.44%, precision of 95.3%, recall of 95.33%, F-measure of 95.26%, and error rate of 4.66%. After 45 epochs, proposed method found best result for KTH dataset which gives an accuracy of 90.1%, precision of 90.1%, recall of 90.54%, F-measure of 90.2% and an error rate of 9.9%. After 20 epochs, proposed method received best results for UCF 50 dataset which gives an accuracy of 80.80%, precision of 80.27%, recall of 80.27%, F-measure of 79.68%, and error rate of 19.2%.

A. Hardware and Software Set Up

1) Hardware set up: For experimental validation, this research used two different computers with similar configurations. Both computers were running on windows 10x64 platforms with an Intel Core i5-7200U CPU 2.5GHz with turbo boost up to 3.1 GHz and both with 8 Gigabytes of RAM and with no external Graphics Processing Unit (GPU).

2) Software set up: This research used python 3.7.1 on Jupyter Notebook 6.0.3 Integrated Development Environment (IDE) on Anaconda Navigator 2, Atom 1.51.0 x64 IDE. This research used different types of python libraries OpenCV 3.4.2, Tensorflow 2.1.0, Keras 2.3.1, MatplotLib 3.2.2 for line graph, Seaborn 0.10.1 for confusion matrix, FFMpeg 4.2.2, NumPy 1.19.1, Pandas 1.0.5, Scikit Learn 0.23.1 for evaluation metrices, Tqdm 4.47.0, CSV, and Glob.

B. Evaluation Parameters

To evaluate the performance of the proposed method, evaluation was done by finding the accuracy, precision, recall, f-measure, and error rate on various datasets. In this context, TP, FP, FN, TN are used to calculate accuracy, precision, recall, f-measure and error rate [67]. True Positive (TP) is the number of correct predictions that an instance is negative. False Positive (FP) is the number of incorrect predictions that an instance is positive. False Negative (FN) is the number of incorrect of predictions that an instance negative. True Negative (TN) is the number of correct predictions that an instance is positive [69].

1) Accuracy: Accuracy is the most natural measure of performance which refers a ratio between correctly predicted observation and total observations expressed using following equation [63, 67].

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \quad (1)$$

2) Precision: Precision is the proportion of positive observations accurately predicted to the overall predicted

positive observations which is expressed using following equation [64, 67].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

3) Recall: Recall is the proportion of positive observations accurately predicted to all observations in actual class which is expressed using following equation [65, 67].

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4) F-Measure: F-Measure is the weighted average of Recall and Precision which takes both false positives and false negatives into account and is expressed using following equation [67, 68].

$$\text{F Measure} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

5) Error rate: Error rate is simply a ratio of wrongly predicted observation to the total observations which is expressed using following equation [67].

$$\text{Error rate} = \frac{FP+FN}{TP+FP+FN+TN} \quad (5)$$

6) Loss: Loss is a number that indicates how poor a prediction of the model was in one particular case [59,60,61,62,66]. Loss is zero if the assumption of the model is right or else the loss is greater. Loss functions are intended to measure how much a method requires to minimize value in training. This research used the categorical_crossentropy as the loss function. The value measured by the loss function is simply called a loss which is usually used when there are two or more than two label classes.

7) Confusion matrix: For deep learning classification perspectives, confusion matrix is a performance measurement where two or more classes generate output [59, 60, 61, 62, 66, 69]. Confusion matrix is a table with four different predicted and actual value combinations which is very useful for measuring accuracy, precision, recall, f- measure, and error rate. This research found predicted percentage value comparing to the actual value and inserted the value for that class in the confusion matrix.

C. Datasets

In order to evaluate the performance of the proposed Convolutional Long Short-Term Deep Network (CLSTDN) for Intention recognition, this research evaluated the performance of the network with the publicly accessible datasets like: UCF-Sports [31, 36, 60, 66], UCF-11 [60, 66], KTH [8,31,47,62], and UCF-50 [31,36]. Such challenging datasets are commonly used for benchmarking. Later, this research compared experimental results with state-of-the-art approaches. Details for each of the datasets are illustrated below.

1) UCF-Sport: UCF Sport is one of the oldest datasets for action's recognition which consists of a sequence of acts from different sports activities. The dataset consists of 150 videos for 10 human actions with a resolution of 720x480. This dataset's frame numbers vary from video to video. Frame rate of the UCF Sports dataset was 10 frames per second and on average each video contains 30 to 130 frames in total.

2) UCF-11: UCF-11 dataset includes 11 types of action categories. Due to broad variations in camera movement, object appearance and posture, object size, view, cluttering background, lighting conditions, etc. This dataset is very complex and challenging comparing than YouTube video-based dataset.

3) KTH: The KTH is the most widely used human behavior public dataset which contains 6 different types of video action with a resolution of 160x120. 25 participants in four different scenarios performed actions multiple times. For each combination of 25 subjects, 6 actions, and 4 scenarios, there are total of 600 video files. All sequences with a static camera with a 25fps frame rate had been taken over homogenous backgrounds.

4) UCF-50: UCF50 is a collection of action recognition data with 50 categories of action, consisting of realistic videos taken from YouTube. This dataset is an extension of the YouTube Action dataset (UCF-11). The videos are grouped into 25 groups for all the 50 categories, where each group consisting of more than four action clips with a resolution of 320x240.

This research splits all datasets into 70% for training and 30% for testing. The training and testing data split are shown in Table I.

TABLE I. DESCRIPTION OF TRAIN AND TEST SPLIT OF VIDEOS OF DATASETS

Dataset	Training Data	Testing Data
UCF Sports	103	47
UCF-11	1084	439
KTH	407	192
UCF-50	4636	1839

In the UCF Sports dataset, this research extracted 15 frames from per second of the video. For UCF-11 dataset, this research extracted 10 frames per second of the video as there are more videos. Also, for the UCF-50 dataset, this research extracted 10 frames from per second of the video as the dataset is very big in size and the images are loaded into a resolution of 224x224x3. For the KTH dataset, this research extracted 25 frames per second of the video and images are loaded into a resolution of 120x120x3 as the videos were of low resolution. This research ignored the frames mentioned to be ignored from the official site of the KTH dataset so there is no ambiguous data. This research reshaped images into a resolution of 299x299 because pre-trained Inception-ResNet-v2 network accepts an input of size 299x299 only. Inception-ResNet-v2 is a pre-trained convolutional neural network which was used to extract the spatial features from frames of a video.

To train the model, this research used batch size of 32 for the UCF Sports dataset, 32 for the UCF-11 dataset, 64 for KTH and 128 for the UCF-50 dataset. This research used

LSTM of size 2048 and two more dense layers size of 1024 and 512 with the most common Rectified Linear Unit (ReLU) activation function. This research used dropout probability of 0.1 to reduce any overfitting issues in all layers for UCF sports and UCF-11 dataset. In addition, this research used a dropout of 0.1 on the dense layer size of 512 for KTH and UCF-50 dataset and no dropout on other layers. Finally, this research used SoftMax activation function on the last dense layer of the size of the output class to achieve output. This research used Adam optimizer with a learning rate of 1x10^-5, a decay rate of 1x10^-6 and a momentum of 0.2 for all the datasets. This research sets the model to train for 100 epochs but if the validation loss of the model does not improve for 10 consecutive epochs, this research used early stopper function to stop epochs.

D. Experimental Results

Proposed method received best results for UCF Sports dataset after 62 epochs which needed 2336 seconds with accuracy of 95.74%, precision of 95.83%, recall of 97.49%, F measure of 96.23% and error rate of 4.26% shown in Table II. After 43 epochs that needed 4970 seconds, proposed method received best results for UCF 11 dataset which gives an accuracy of 95.44%, precision of 95.3%, recall of 95.33%, F-measure of 95.26%, and error rate of 4.66%. After 45 epochs that needed 10556 seconds, proposed method achieved best results for the KTH dataset which gives an accuracy of 90.1%, precision of 90.1%, recall of 90.54%, F-measure of 90.2%, and error rate of 9.9%. After 20 epochs which needed 5146 seconds, proposed method received best results for UCF 50 dataset which gives an accuracy of 80.80%, precision of 80.27%, recall of 80.27%, F-measure of 79.68% and an error rate of 19.2%.

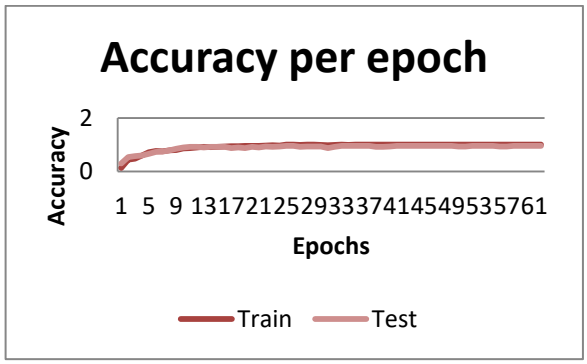
TABLE II. EXPERIMENTAL RESULTS FOR THE PROPOSED METHOD IN FOUR DATASETS

Dataset	Accuracy	Precision	Recall	F-Measure	Error rate
UCF Sports	95.74	95.83	97.49	96.23	4.26
UCF-11	95.44	95.3	95.33	95.26	4.66
KTH	90.1	90.1	90.54	90.2	9.9
UCF-50	80.8	80.27	80.27	79.68	19.2

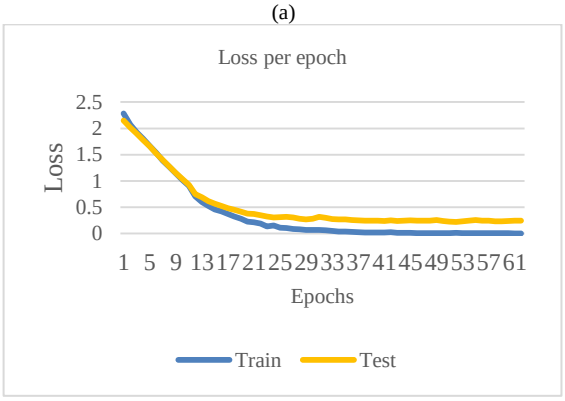
Line graph as Accuracy per epochs of the UCF-Sports dataset is shown in Fig. 3(a) where yellow line indicates testing accuracy and blue line indicates training accuracy. As this research uses deep learning architecture to train and test data, proposed method learns gradually from the train data. From test data proposed method validates performance in each epoch. Fig. 3(a) states that after 62 epochs, training accuracy increased gradually and finally reached 1.0 which is 100%. Test accuracy also increased gradually with each epoch and reached a maximum value of 0.9574 which is 95.74%.

Line graph as Loss per epochs of the UCF Sports dataset is shown in Fig. 3(b) where yellow line is the test loss and blue line is the train loss. Loss is zero if the assumption of the proposed method is right or else the loss is greater. Loss will gradually decrease with each epoch. If test loss value does not decrease for 10 consecutive epochs, then this research stopped the epoch. Following this, this research receives best result with minimum test loss which is targeted for minimizing overfitting. Fig. 3(b) states that minimum train loss is 0.0042 and the minimum test loss is 0.22254.

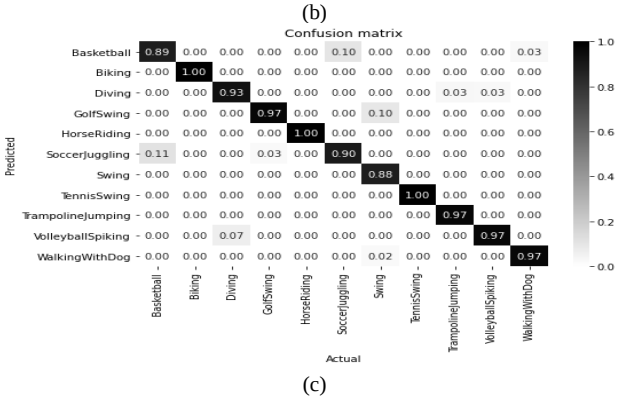
Confusion matrix for UCF Sports dataset is shown in Fig. 3(c) which states that proposed method faced problems in predicting actions like Golf swing and Run side. Running sidewise is one time predicted as kicking by the proposed method. On the other hand, Golf-swing is predicted as kicking for one time by the proposed method. For the rest of the activities, proposed method model predicted intentions accurately. After analyzing various actions in videos for which proposed method predicted wrong, this research observed that some frames from the video of golf swing make the intention of kicking due to similarity of patterns. Also, for the running activity, the videos were a bit unclear and made the proposed method predicting it as kicking due to the same reason.



(a)



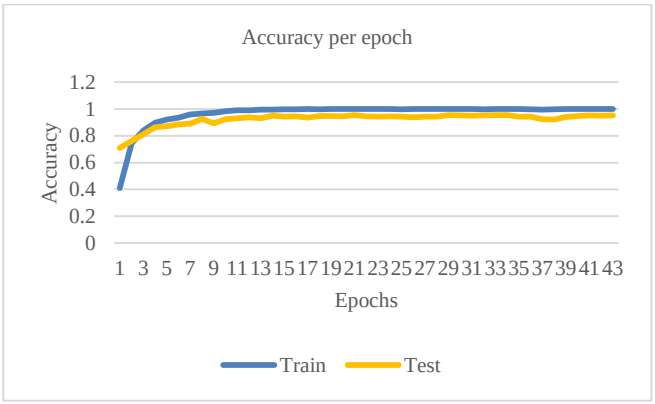
(b)



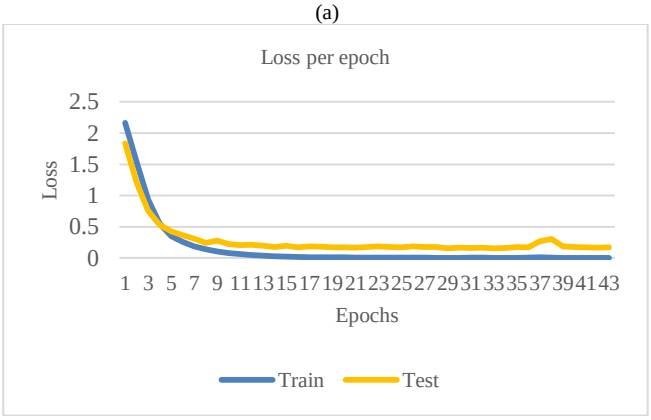
(c)

Fig. 3. (a). Line graph as accuracy of UCF sports dataset, (b). Line graph as loss of UCF sports dataset, (c) Confusion matrix of UCF sports dataset.

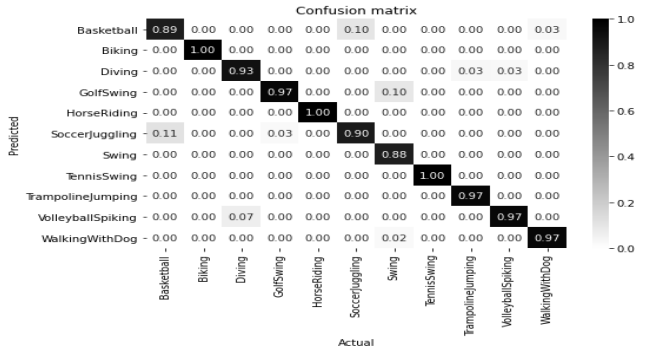
Line graph as accuracy per epochs of the UCF-11 dataset is shown in Fig. 4(a) where yellow line indicates testing accuracy and blue line indicates training accuracy. Fig. 4(a) states that after 43 epochs, training accuracy increased gradually and finally reached 1.0 which is 100%. Testing accuracy also increased gradually with each epoch and reached a maximum value of 0.9544 which is 95.44%. Line graph as the Loss per epochs of the UCF-11 dataset is shown in Fig. 4(b) where yellow line is the test loss and the blue line is the train loss. Loss is zero if the assumption of the proposed method is right or else loss is greater. Best results were achieved with minimum test loss which is targeted for minimizing overfitting. Fig. 4(b) states that the minimum train loss is 0.0012 and minimum test loss is 0.1557.



(a)



(b)



(c)

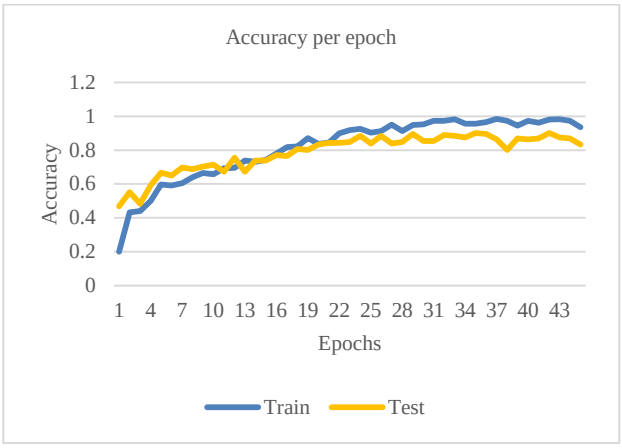
Fig. 4. (a) Line graph as accuracy of UCF-11 dataset, (b) Line graph as loss of UCF-11 dataset, (c) Line graph as confusion matrix of UCF-11 dataset.

Confusion matrix of UCF-11 dataset is shown in Fig. 4(c). From the confusion matrix, this research observed that proposed method almost predicted all the intentions perfectly where for almost all the activities prediction was robust. In some activities like Basketball playing, proposed method predicted some video sequences as soccer as there is a common object which is a ball. In this context, accuracy can be increased if more training is done with these activities. Also, for swing, there were some video angles in which proposed method was not trained which causes wrong prediction for those angles as golf swings.

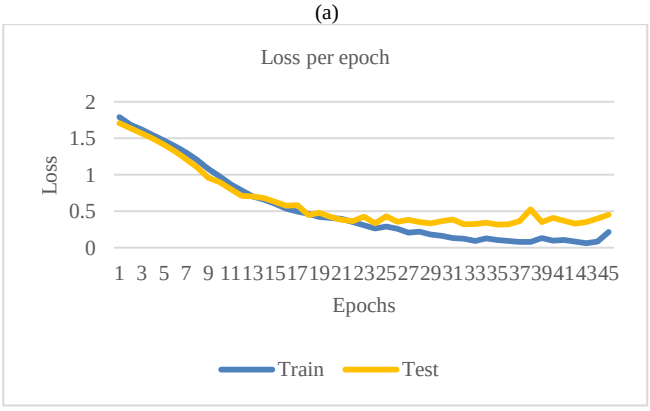
Line graph as the Accuracy per epochs of the KTH dataset is shown in Fig. 5(a) where yellow line indicates testing accuracy and blue line indicates training accuracy. Fig. 5(a) states that after 45 epochs, training accuracy increased gradually and finally reached 0.9853 which is 98.53%. Test accuracy also increased gradually with each epoch and reached a maximum value of 0.9010 which is 90.10%.

Line graph as the Loss per epochs of the KTH dataset is shown in Fig. 5(b) where yellow line indicates test loss and blue line indicates train loss. Proposed method receives best results with minimum test loss which was targeted for minimizing the overfitting. Fig. 5(b) shows that the minimum train loss is 0.0645 and the minimum test loss is 0.3179. Confusion matrix of the KTH dataset is shown in Fig. 5(c). KTH dataset is robust datasets which contains similar kinds of activities like jogging, running and walking. Confusion matrix states that proposed method by this research predicted the intentions accurately enough. Although, proposed method faced problems for running activity which was predicted as jogging, which is also tough for a normal human being to differentiate between, them due to the similarity of the pattern.

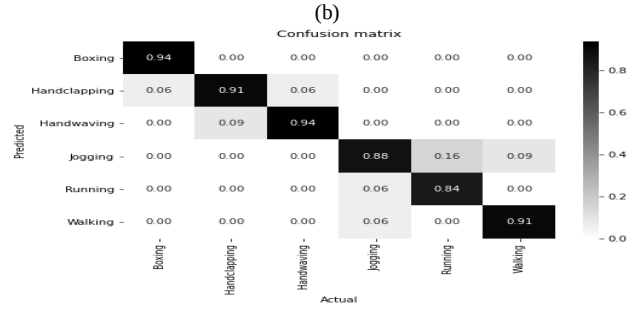
Line graph as the Accuracy per epochs of the UCF-50 dataset is shown in Fig. 6(a) where yellow line indicates testing accuracy and blue line indicates training accuracy. Fig. 6(a) states that after 20 epochs, training accuracy increased gradually and finally reached 1.0 which is 100%. Testing accuracy also increased gradually with each epoch and reached a maximum value of 0.8113 which is 81.13%. Line graph as the Loss per epochs of the UCF-50 dataset is shown in Fig. 6(b) where yellow line indicates test loss and blue line indicates train loss. Proposed method received accuracy was 81.13%, best test loss was achieved with accuracy of 80.8%. Fig. 6(b) states that minimum train loss is 0.0098 and the minimum test loss is 0.7413. Confusion matrix of UCF-50 dataset is shown in Fig. 6(c). Confusion matrix for UCF-50 datasets states that proposed method faced problems in predicting the actions like nunchucks, jumping jack, and javelin throw. In these actions, there are some positions or frames which are like other activities. Nunchucks is often considered as golf swing or swing because it has some position like these two activities. For jumping jacks, it has some frames which make the proposed method predicting other activities like lunges, jumping rope or pullups. In javelin throwing, the actor often jumps to throw the javelin which was predicted as high jump by our proposed method often due to similarity of the pattern.



(a)

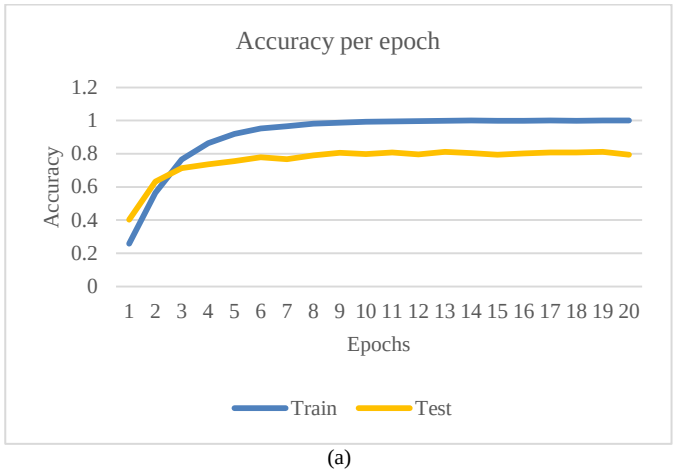


(b)

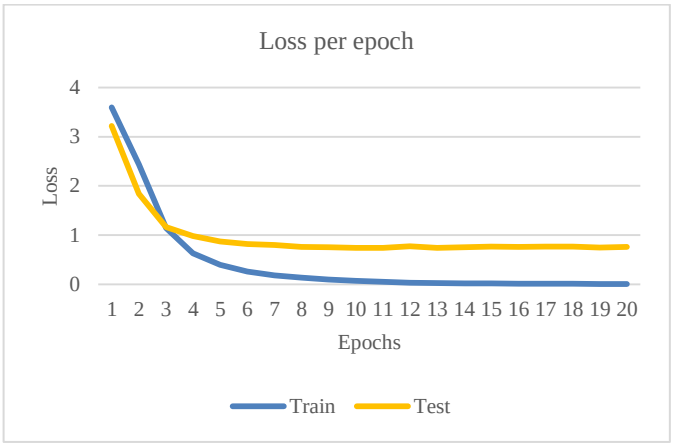


(c)

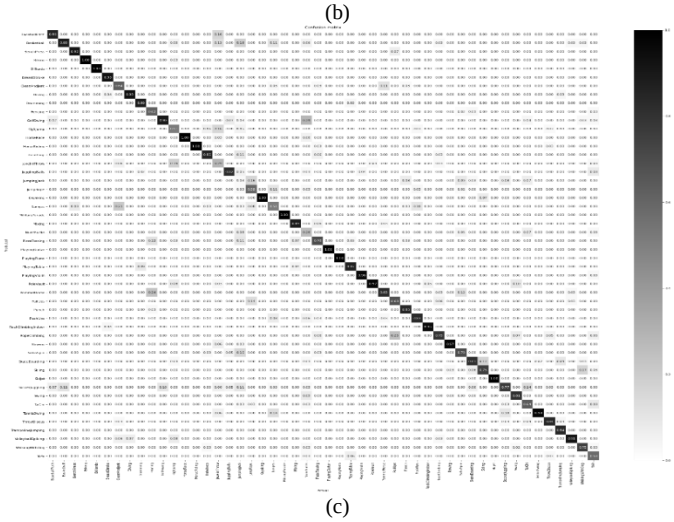
Fig. 5. (a) Line graph as accuracy of KTH dataset, (b) Line graph as loss of KTH dataset, (c) Confusion matrix of KTH dataset.



(a)



(b)



(c)

Fig. 6. (a) Line graph as accuracy of UCF-50 dataset, (b) Line graph as loss of UCF-50 dataset, (c) Confusion matrix of UCF-50 dataset.

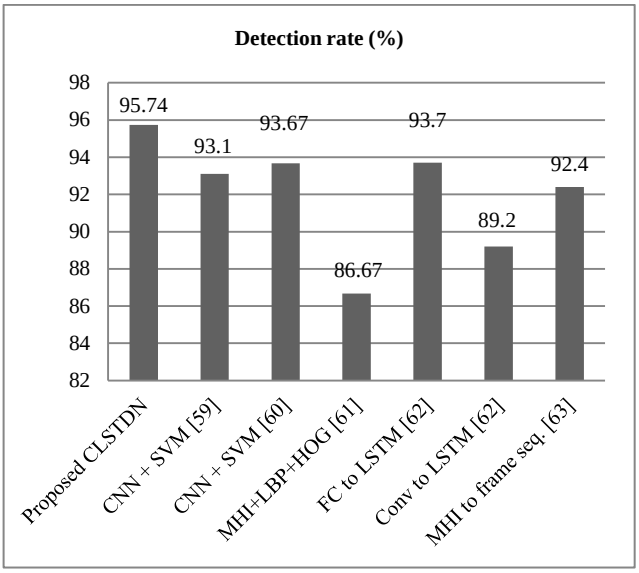
E. Comparison with Previous Research Results

After experimenting on four different datasets, this research compared proposed method with state-of-the-art methods and found that proposed method performed very well using spatial-temporal features compared to other method with a good amount of accuracy, precision, recall and f-measure.

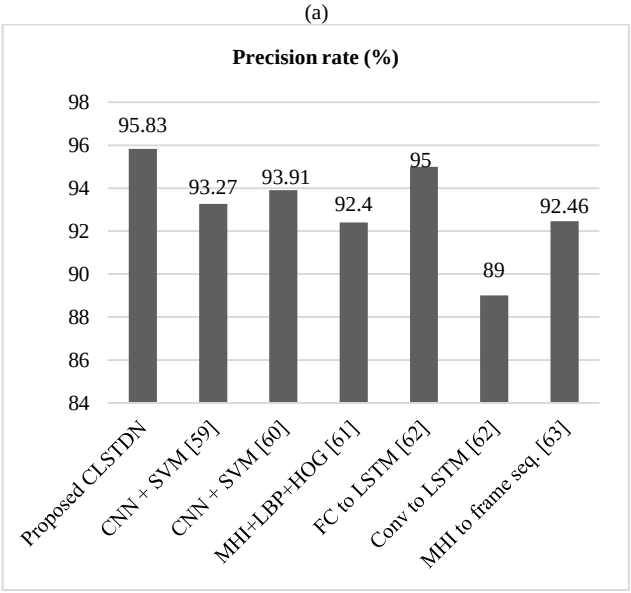
Proposed CLSTDN received accuracy rate of 95.74% which is higher than previous research methods shown in Fig. 7(a). Research in [59] received accuracy of 93.1% using Convolutional Neural Network and Support Vector Machine. They used three frames of a video instead of the all the frames to understand the human action. Besides, they extracted conceptual features to recognize objects and worked with sports-based dataset only. Whereas proposed CLSTDN used spatial-temporal features caused better performance than research in [59]. Research in [60] received accuracy rate of an accuracy of 93.67% using same method as research in [60]. They used only first and last frames of a video instead of the all the frames to understand human action. Like in research [59], they used conceptual features to recognize objects and used SVM to classify high level and stationary features obtained from CNN; whereas usage of spatial-temporal features by the proposed CLSTDN resulted better performance than research in [60]. Research in [61] received

accuracy rate of 86.67% using action-history and histogram of oriented gradient. They used Motion History Image (MHI), Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG) approaches for action recognition. Proposed CLSTDN by this research pre-trained CNN used Inception-ResNet-v2 to extract features from deep inside the image to create sequence. For this reason, proposed CLSTDN achieved better performance comparing with research in [61]. Research in [62] received accuracy rate of 93.7% using fully connected-to-LSTM. They used VGG-16 as pre-trained CNN network and fused the result with LSTM to recognize human actions whereas proposed CLSTDN extract spatial features from average pooling layer and passed the data sequence to LSTM to extract temporal features to recognize human intentions. Also, research in [62] used spatial features only to recognize human actions whereas proposed CLSTDN used both spatial and temporal features to recognize human intention and achieved better accuracy. Research in [66] received accuracy rate of 92.4% using motion history images of frame sequences with spatial information extraction. They used Motion History frame sequence to understand the temporal changes. However, proposed CLSTDN passed spatial feature sequences to LSTM for temporal understanding of the whole video caused higher accuracy than research in [66]. In overall, previous methods used spatial or temporal data for understanding video whereas proposed CLSTDN uses both spatial-temporal understanding of a video which helps for a better understanding of human intention.

Proposed CLSTDN received precision rate of 95.83% which is higher than previous methods shown in Fig. 7(b). Research in [59] received precision rate of 93.27% using Convolutional Neural Network and Support Vector Machine. Precision rate indicates the proportion of positive observations accurately predicted to the overall predicted positive observations. As research in [59] used only three frames of a video instead of the all the frames to understand the human action, their overall positive classifications were less than proposed method by this research. Research in [60] received precision rate of 93.91% using Convolutional Neural Network and Support Vector Machine. As they used first and last frames only of a video instead of the all the frames to understand the human action caused their positive classifications less than proposed CLSTDN by this research. As accurately predicted positive observation by research in [60] is less than proposed CLSTDN, precision is also less than proposed CLSTDN. They used spatial-temporal understanding of an image but they are still unable to learn from an image due to lack of depth features which causes lower accurately predicted positive observations than proposed CLSTDN and precision is also higher than research in [61]. Research in [62] received precision rate of 95% and 89% using fully connected-to-LSTM and Convolutional-to-LSTM respectively. They used VGG-16 as pre-trained CNN network caused number predicted positive observations less than proposed method and finally resulted in better precision rate than in research [62]. Research in [66] received precision rate of 92.46% using frame sequences and spatial features extraction. Like other previous methods, research in [66] used the VGG-16 pre-trained network which caused less accurate predictions than proposed method by this research and resulted in lower precision rate than by the proposed CLSTDN.



(a)



(b)

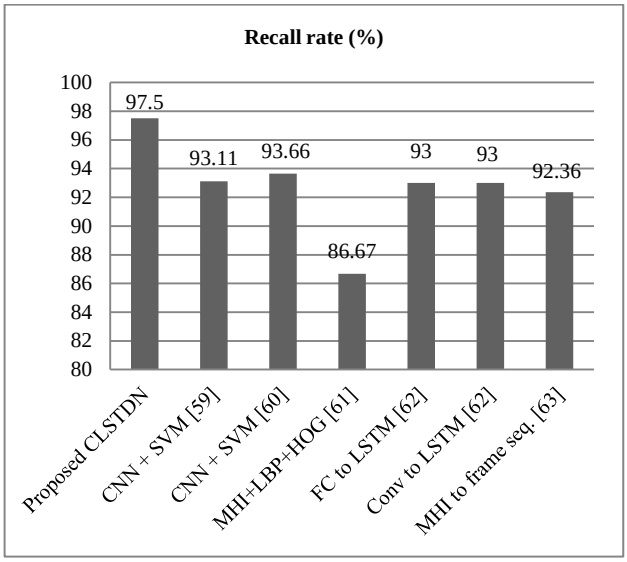
Fig. 7. (a) Detection rate comparison with previous research., (b) Precision rate comparison with previous research.

Proposed CLSTDN received recall rate of 97.5% which is higher than previous methods shown in Fig. 8(a). Research in [59] received recall rate of 93.11% using Convolutional Neural Network and Support Vector Machine which is lower than proposed CLSTDN due to usage of randomly three frames only to understand the overall video. Research in [60] received recall rate of 93.66% using Convolutional Neural Network and Support Vector Machine. However, due to usage of first and last frame to understand overall video without Spatio-temporal features, research in [60] produced lower recall rate comparing with the proposed CLSTDN. Research in [61] received recall rate of 86.67%% using Binary patterns of action-history and histogram of oriented gradient. Similarly, research in [62] received recall rate of 93% using

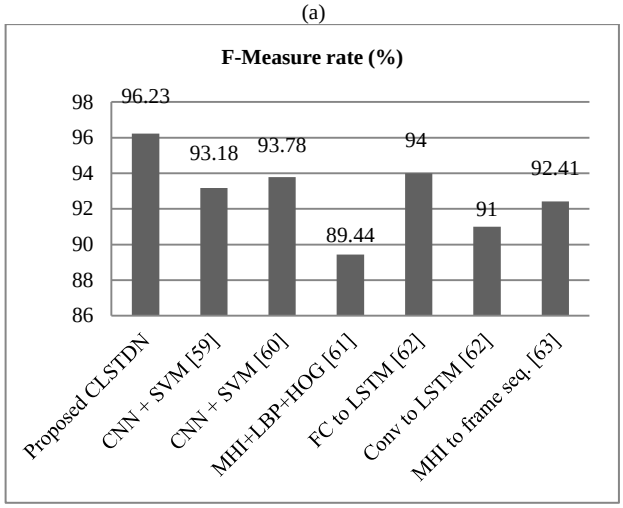
both fully connected-to-LSTM and Convolutional-to-LSTM and research in [66] received recall rate of 92.36% using frame sequences and spatial features extraction. Proposed CLSTDN achieved better recall rate than research in [61], [62] and [66] due to usage of Inception-ResNet-v2 for deep Spatio-temporal features extraction instead of VGG-16.

Proposed CLSTDN received F-Measure value of 96.23% which is higher than previous methods shown in Fig. 8(b). Research in [59] received F-Measure value of 93.18% using Convolutional Neural Network and Support Vector Machine. Similarly, research in [60] received F-Measure value of 93.78% using Convolutional Neural Network and Support Vector Machine. F measure indicates the weighted average of recall and precision which considers both false positives and negative into account. As proposed CLSTDN understands spatial-temporal features of the video sequence and human motion in video, better F-Measure was achieved by this research comparing with research in [59] and [60]. Research in [61] received F-Measure value of 89.44% using Binary patterns of action-history and histogram of oriented gradient. As proposed method used Inception-ResNet-V2 to understand video sequence, better F-Measure rate was achieved than research in [61]. Research in [62] received F-Measure rate of 94% and 91% using fully connected-to-LSTM and Convolutional-to-LSTM respectively. As number of features extraction by VGG-16 is less and not deep as Inception-ResNet-V2 which used by proposed CLSTDN, F measure by research in [62] is not promising like proposed method by this research. Research in [66] received recall rate of F-Measure rate of 92.41% using frame sequences and spatial features extraction. They used VGG-16 pre-trained network to understand temporal changes from Motion History Images. However, proposed CLSTDN used Inception-ResNet-V2 to model spatial features which were passed to LSTM and caused understanding patterns better than research in [66]. For this reason, better F-Measure was achieved by this research comparing with research in [66].

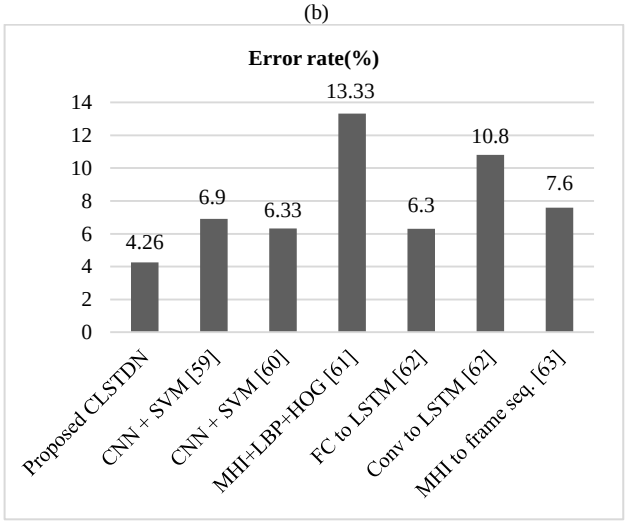
Proposed CLSTDN received error rate of 4.26% which is lower than previous methods shown in Fig. 8(c). Research in [59] and [60] received error rate of 6.9% and 6.33% respectively using Convolutional Neural Network and Support Vector Machine. Error rate by research in [59] and [60] is higher than proposed CLSTDN due to usage of random frames to identify objects in the video scenes. Research in [61] received error rate of 13.33% which is very high comparing with the proposed CLSTDN due to learn features more deeply using Inception-ResNet-V2. Research in [62] received error rate of 6.3% and 10.8% using fully connected-to-LSTM and Convolutional-to-LSTM respectively which made their proposed approach computationally heavy and resulted in higher error rate comparing with the proposed CLSTDN. Research in [66] received error rate of 7.6% using frame sequences and spatial features extraction which is also higher than proposed CLSTDN due to extract more depth features using Inception-ResNet-V2.



(a)



(b)



(c)

Fig. 8. (a) Recall rate in comparison with previous research, (b) F- Measure rate in comparison with previous research., (c) Error rate in comparison with previous research.

In overall, results from experimenting on different datasets showed that proposed Convolutional Long Short-Term Deep Network (CLSTDN) using both spatial and temporal features can lead to a viable solution for significantly improving recognition performance of human intentions of various activities. Proposed method performed well on UCF Sports, KTH, and UCF-11 dataset with low computation power. However, proposed method lacks in performance in the UCF-50 dataset due to the number of frames extracted from the video which was very low.

Proposed CLSTDN was evaluated on four publicly available datasets where accuracy, precision, recall, f-measure, and error rate were estimated for each dataset. To validate the proposed method, satisfactory accurate results were achieved on determining the intention on real time videos. After 62 epochs that needed 2336 seconds, proposed method received best results for UCF Sports dataset which gives an accuracy of 95.74%, precision of 95.83%, recall of 97.49%, f-measure of 96.23%, and an error rate of 4.26%. Previous methods used either spatial or temporal data for understanding the scene [33] [37][39][42][46][48][51][52][53], whereas proposed method used both spatial-temporal understanding of a video which helps for a better understanding of an intention. To extract spatial features, proposed method used pre-trained convolutional neural network which was Inception ResNet V2 and passed the sequence to long short-term memory for the temporal understanding of the sequence. Also, this research dealt with limited data sequences whereas other methods used full data which causes the proposed method be faster comparing with previous methods. In addition, another benefit of the proposed method is that it runs on low computational power and resources. As this research illustrated earlier, because of taking both spatial and temporal features, proposed method gives more accurate results.

V. Conclusion

This research proposed Convolutional Long Short-Term Deep Network (CLSTDN) to recognize human action-based intention. Proposed CLSTDN consists of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Inception-ResNet-v2 was used as pre-trained network for CNN which contains 164 deep layers deep to classify class specific object categories. Long Short-Term Memory (LSTM) network was used for Recurrent Neural Network. Proposed CLSTDN extracted spatial features by Convolutional Neural Network and temporal features by Recurrent Neural Network to ensure the usage of Spatio-temporal features for efficient human intention prediction. Overall proposed methodology consists of three main phases, i.e., data preprocessing, feature extraction and final classification. Data preprocessing involves reshape, handling of blank image frames and overfitting to prepare the data to feed into Inception-ResNet-v2. Feature extraction phase involves implication of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Finally, dense layers and SoftMax activation function predicts human intention based on spatial-temporal features. Training of the proposed methodology was done on low computation resources and achieved better performance comparing with existing research results. Four publicly available datasets were used for validation, i.e., UCF Sports, UCF-11, KTH and UCF-50. Validation of the proposed CLSTDN was done based seven evaluation metrics, i.e., accuracy, precision, recall, f-measure, error rate, confusion matrix and loss. Previous research methods used either spatial or temporal features for human intention prediction, however, proposed method used spatial-temporal features for human intention prediction caused more improved performance than previous research methods. In addition, proposed CLSTDN performed efficiently based on all the evaluation metric with a limited number of data sequence irrespective of viewpoint, background, inter-class and intra-class similarities present in the image frames with very small data sequences in almost all the datasets. However, proposed method produced some errors for several activities, i.e., jogging and running in KTH dataset, basketball and soccer game in UCF-11 dataset, nunchucks, jumping jack, and javelin throw in UCF-50 dataset due to similarity of patterns. In future, proposed CLSTDN will be investigated more comprehensively for more similar types of human actions.

Acknowledgment

Authors would like to thank Artificial Intelligence and Data Science Program under Mentoring Access & Platforms in STEM (MAPS) in Colorado State University in Pueblo, USA for financial support for this research. Authors also would like to thank School of Engineering, Colorado State University in Pueblo for various research supports.

References

[1] T.D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, A.-R. Sadeghi, DÏoT: A federated self-learning anomaly detection system for IoT, in: 2019 IEEE 39th International conference on distributed computing systems (ICDCS), IEEE, 2019, pp. 756-767.

[2] T. Fernando, S. Denman, S. Sridharan, C. Fookes, Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds, in: Asian conference on computer vision, Springer, 2018, pp. 314-330.

[3] G.K. Gudur, P. Sundaramoorthy, V. Umaashankar, ActiveHARNet: Towards on-device deep Bayesian active learning for human activity recognition, in: The 3rd International Workshop on Deep Learning for Mobile Systems and Applications, 2019, pp. 7-12.

[4] Z. Wei, C. Wang, P. Hao, M.J. Barth, Vision-based lane-changing behavior detection using deep residual neural network, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE, 2019, pp. 3108-3113.

[5] M. Munir, M.A. Chattha, A. Dengel, S. Ahmed, A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, 2019, pp. 561-566.

[6] C. Soh, S. Yu, A. Narayanan, S. Duraisamy, L. Chen, Employee profiling via aspect-based sentiment and network for insider threats detection, Expert Systems with Applications, 135, 2019, pp.351-361.

[7] C.-M. Kuo, S.-H. Lai, M. Sarkis, A compact deep learning model for robust facial expression recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2121-2129.

[8] B. Padmaja, M.B. Myneni, E.K.R. Patro, A comparison on visual prediction models for MAMO (multi activity-multi object) recognition using deep learning, Journal of Big Data, 7,2020, pp.1-15.

[9] C.-P. Bara, M. Papakostas, R. Mihalcea, A Deep Learning Approach Towards Multimodal Stress Detection, in: AffCon@ AAAI, 2020, pp. 67-81.

[10] Z. Tariq, S.K. Shah, Y. Lee, Speech emotion detection using iot based deep learning for health care, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 4191-4196.

[11] Y. Gao, Y. Zhang, H. Wang, X. Guo, J. Zhang, Decoding behavior tasks from brain activity using deep transfer learning, IEEE Access, 7 (2019) 43222-43232.

[12] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, F.-Y. Wang, Driver activity recognition for intelligent vehicles: A deep learning approach, IEEE transactions on Vehicular Technology, 68 (2019) 5379-5390.

[13] P. Venuprasad, T. Dobhal, A. Paul, T.N. Nguyen, A. Gilman, P. Cosman, L. Chukoskie, Characterizing joint attention behavior during real world interactions using automated object and gaze detection, in: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, 2019, pp. 1-8.

[14] W. Shi, J. Li, Y. Yang, Face fatigue detection method based on MTCNN and machine vision, in: International Conference on Applications and Techniques in Cyber Security and Intelligence, Springer, 2019, pp. 233-240.

[15] W. Fang, Y. Ding, F. Zhang, J. Sheng, Gesture recognition based on CNN and DCGAN for calculation and text output, IEEE access, 7 (2019) 28230-28237.

[16] T. Zhang, Gesture Recognition Based on Deep Learning, Journal of Physics: Conference Series, 1449 (2020).

[17] M.S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, Information Fusion, 49 (2019) 69-78.

[18] A. Saif, Z.R. Mahayuddin, Robust Drowsiness Detection for Vehicle Driver using Deep Convolutional Neural Network, International Journal of Advanced Computer Science and Applications, (2020).

[19] A. Saif, Z.R. Mahayuddin, Fast and Effective Motion Model for Moving Object Detection Using Aerial Images, International Journal of Advanced Computer Science and Applications, (2018).

[20] V.R. Mali, A.R. Surve, V. Ghorpade, IoT Enabled Detection of Suspicious Human Behavior for ATM Environment, in: Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies, Springer, 2020, pp. 269-277.

[21] A.A. Sukor, A. Zakaria, N.A. Rahim, L. Kamarudin, H. Nishizaki, Abnormality detection approach using deep learning models in smart home environments, in: Proceedings of the 7th International Conference on Communications and Broadband Networking, 2019, pp. 22-27.

[22] G. Diraco, A. Leone, A. Caroppo, P. Siciliano, Deep Learning and Machine Learning Techniques for Change Detection in Behavior Monitoring, in: AI* AAL@ AI* IA, 2019, pp. 38-50.

[23] B. Rezaei, Y. Christakis, B. Ho, K. Thomas, K. Erb, S. Ostadabbas, S. Patel, Target-specific action classification for automated assessment of human motor behavior from video, Sensors, 19 (2019) 4266.

[24] A.A.Q. Mohammed, J. Lv, M. Islam, A deep learning-based end-to-end composite system for hand detection and gesture recognition, Sensors, 19 (2019) 5282.

[25] Y. Gu, H. Zhang, S. Kamijo, Multi-person pose estimation using an orientation and occlusion aware deep learning network, Sensors, 20 (2020) 1593.

[26] K. Slimani, K. Lekdioui, R. Messoussi, R. Touahni, Compound facial expression recognition based on highway cnn, in: Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society, 2019, pp. 1-7.

[27] V. Raudonis, A. Paulauskaite-Taraseviciene, K. Sutiene, D. Jonaitis, Towards the automation of early-stage human embryo development detection, Biomedical engineering online, 18 (2019) 1-20.

[28] A. Saif, Z.R. Mahayuddin, Vision based 3D Gesture Tracking using Augmented Reality and Virtual Reality for Improved Learning Applications, International Journal of Advanced Computer Science and Applications, (2021).

[29] I. Condés, J.M. Cañas, Person Following Robot Behavior Using Deep Learning, in: Workshop of Physical Agents, Springer, 2018, pp. 147-161.

[30] M.M. Hassan, M.Z. Uddin, A. Mohamed, A. Almogren, A robust human activity recognition system using smartphone sensors and deep learning, Future Generation Computer Systems, 81 (2018) 307-313.

[31] N. Jaouedi, N. Boujnah, M.S. Bouhlel, A new hybrid deep learning model for human action recognition, Journal of King Saud University-Computer and Information Sciences, 32 (2020) 447-453.

[32] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, J. Chen, Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing, Procedia CIRP, 83 (2019) 272-278.

[33] X. Qin, Y. Ge, J. Feng, D. Yang, F. Chen, S. Huang, L. Xu, DTMMN: Deep transfer multi-metric network for RGB-D action recognition, Neurocomputing, 406 (2020) 127-134.

[34] H. Fujiyoshi, T. Hirakawa, T. Yamashita, Deep learning-based image recognition for autonomous driving, IATSS research, 43 (2019) 244-252.

[35] Y. Gu, X. Ye, W. Sheng, Y. Ou, Y. Li, Multiple stream deep learning model for human action recognition, Image and Vision Computing, 93 (2020) 103818.

[36] Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, J. Yuan, Multi-stream CNN: Learning representations based on human-related regions for action recognition, Pattern Recognition, 79 (2018) 32-43.

[37] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J.C.S.J. Junior, X. Baró, H. Demirel, Dominant and complementary emotion recognition from still images of faces, IEEE Access, 6 (2018) 26391-26403.

[38] A. Othmani, A.R. Taleb, H. Abdelkawy, A. Hadid, Age estimation from faces using deep learning: A comparative analysis, Computer Vision and Image Understanding, 196 (2020) 102961.

[39] F. You, Y. Gong, H. Tu, J. Liang, H. Wang, A fatigue driving detection algorithm based on facial motion information entropy, Journal of advanced transportation, 2020 (2020).

[40] M. Baba, V. Gui, C. Cernazanu, D. Pescaru, A sensor network approach for violence detection in smart cities using deep learning, Sensors, 19 (2019) 1676.

[41] A. Al-Dhamari, R. Sudirman, N.H. Mahmood, Transfer deep learning along with binary support vector machine for abnormal behavior detection, IEEE Access, 8 (2020) 61085-61095.

[42] S. Ruan, C. Tang, X. Zhou, Z. Jin, S. Chen, H. Wen, H. Liu, D. Tang, Multi-pose face recognition based on deep learning in unconstrained scene, Applied Sciences, 10 (2020) 4669.

[43] C.N. Phyo, T.T. Zin, P. Tin, Complex human–object interactions analyzer using a DCNN and SVM hybrid approach, Applied Sciences, 9 (2019) 1869.

[44] D. Liciotti, M. Bernardini, L. Romeo, E. Frontoni, A sequential deep learning application for recognising human activities in smart homes, Neurocomputing, 396 (2020) 501-513.

[45] P. Wang, H. Liu, L. Wang, R.X. Gao, Deep learning-based human motion recognition for predictive context-aware human-robot collaboration, CIRP annals, 67 (2018) 17-20.

[46] L. Zhang, S. Li, H. Xiong, X. Diao, O. Ma, An application of convolutional neural networks on human intention prediction, Int. J. Artif. Intell. Appl., 10 (2019) 1-11.

[47] F. Yao, RETRACTED ARTICLE: Deep learning analysis of human behaviour recognition based on convolutional neural network analysis, Behaviour & Information Technology, 40 (2021) LXXVI-LXXXIV.

[48] Z. Liu, J. Hao, Intention recognition in physical human-robot interaction based on radial basis function neural network, Journal of Robotics, 2019 (2019).

[49] Z. Fang, A.M. López, Intention recognition of pedestrians and cyclists by 2d pose estimation, IEEE Transactions on Intelligent Transportation Systems, 21 (2019) 4773-4783.

[50] W. Huang, X. Liu, M. Luo, P. Zhang, W. Wang, J. Wang, Video-based abnormal driving behavior detection via deep learning fusions, IEEE Access, 7 (2019) 64571-64582.

[51] S. Mirjalili, H. Faris, I. Aljarah, Introduction to evolutionary machine learning techniques, in: Evolutionary Machine Learning Techniques, Springer, 2020, pp. 1-7.

[52] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, M. Hasan, B.C. Van Essen, A.A. Awwal, V.K. Asari, A state-of-the-art survey on deep learning theory and architectures, Electronics, 8 (2019) 292.

[53] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks, 61 (2015) 85-117.

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[55] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in neural information processing systems, 27 (2014).

[56] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4305-4314.

[57] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933-1941.

[58] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L.V. Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European conference on computer vision, Springer, 2016, pp. 20-36.

[59] G. Shamsipour, J. Shanbehzadeh, H. Sarrafzadeh, Human action recognition by conceptual features, (2017).

[60] A. Saif, Z.R. Mahayuddin, Moving Object Detection Using Semantic Convolutional Features, Journal of Information System and Technology Management, (2022), pp. 24-41.

[61] M. Rahman Ahad, M. Islam, I. Jahan, Action recognition based on binary patterns of action-history and histogram of oriented gradient, Journal on Multimodal User Interfaces, 10 (2016) 335-344.

[62] H. Gammulle, S. Denman, S. Sridharan, C. Fookes, Two stream lstm: A deep fusion framework for human action recognition, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 177-186.

[63] A. Saif, Z.R. Mahayuddin, Moment features based violence action detection using optical flow, International Journal of Advanced Computer Science and Applications, 11 (2020).

[64] A. Saif, Z.R. Mahayuddin, Crowd Density Estimation from Autonomous Drones Using Deep Learning: Challenges and Applications, Journal of Engineering and Science Research, (2021), pp.01-06.

[65] A. Saif, Z.R. Mahayuddin, An Efficient Method for Hand Gesture Recognition using Robust Features Vector, Journal Information System and Technology Management (JISTM), (2021), pp.25-35.

[66] S. Zebhi, S. Almodarresi, V. Abootalebi, Human activity recognition by using MHIs of frame sequences, Turkish Journal of Electrical Engineering & Computer Sciences, 28 (2020) 1716-1730.

[67] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, International journal of data mining & knowledge management process, 5 (2015) 1.

[68] A. Saif, Z.R. Mahayuddin, Vision based 3D Object Detection using Deep Learning: Methods with Challenges and Applications towards Future Directions, International Journal of Advanced Computer Science and Applications, (2022).

[69] A. Santra, C.J. Christy, Genetic algorithm and confusion matrix for document clustering, International Journal of Computer Science Issues (IJCSI), 9 (2012) 322.