Computers, Materials & Continua
DOI: 10.32604/cmc.2023.035512
Article

Leveraging Transfer Learning for Spatio-Temporal Human Activity Recognition from Video Sequences

Umair Muneer Butt[1,2,*], Hadiqa Aman Ullah[2], Sukumar Letchmunan[1], Iqra Tariq[2], Fadratul Hafinaz Hassan[1] and Tieng Wei Koh[3]

[1]School of Computer Sciences, Universiti Sains Malaysia, Penang, 1180, Malaysia
[2]Department of Computer Science, The University of Chenab, Gujrat, 50700, Pakistan
[3]Department of Software Engineering and Information System, Universiti Putra Malaysia, Selangor, 43400, Malaysia
*Corresponding Author: Umair Muneer Butt. Email: umair@student.usm.my
Received: 24 August 2022; Accepted: 26 October 2022

Abstract: Human Activity Recognition (HAR) is an active research area due to its applications in pervasive computing, human-computer interaction, artificial intelligence, health care, and social sciences. Moreover, dynamic environments and anthropometric differences between individuals make it harder to recognize actions. This study focused on human activity in video sequences acquired with an RGB camera because of its vast range of real-world applications. It uses two-stream ConvNet to extract spatial and temporal information and proposes a fine-tuned deep neural network. Moreover, the transfer learning paradigm is adopted to extract varied and fixed frames while reusing object identification information. Six state-of-the-art pre-trained models are exploited to find the best model for spatial feature extraction. For temporal sequence, this study uses dense optical flow following the two-stream ConvNet and Bidirectional Long Short Term Memory (BiLSTM) to capture long-term dependencies. Two state-of-the-art datasets, UCF101 and HMDB51, are used for evaluation purposes. In addition, seven state-of-the-art optimizers are used to fine-tune the proposed network parameters. Furthermore, this study utilizes an ensemble mechanism to aggregate spatial-temporal features using a four-stream Convolutional Neural Network (CNN), where two streams use RGB data. In contrast, the other uses optical flow images. Finally, the proposed ensemble approach using max hard voting outperforms state-of-the-art methods with 96.30% and 90.07% accuracies on the UCF101 and HMDB51 datasets.

Keywords: Human activity recognition; deep learning; transfer learning; neural network; ensemble learning; spatio-temporal

1 Introduction

Human Activity Recognition (HAR) automatically detects people's daily physical activities. A HAR system recognizes a person's actions and gives feedback for intervention [1]. Moreover, it assigns

activity labels to video sequences. The main obstacles to activity recognition systems are the complexity of activities performed and the number of subjects engaging in the activity. Activity recognition systems can be used for people detection, motion tracking, and person identification [2], as shown in Fig. 1. An activity recognition system's purpose is to recognize common human activities. However, human activity is dynamic and diverse, making accurate activity identification difficult. Practical applications such as smart homes, human-computer interaction, automated surveillance, and human healthcare are increasing the need for HAR research.

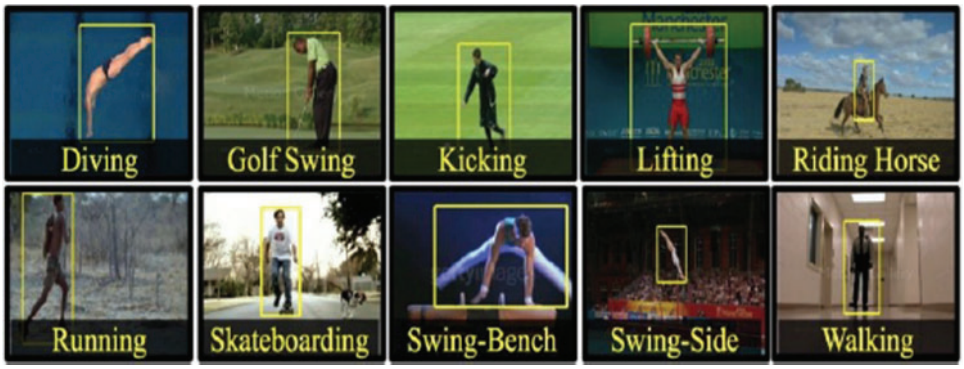

Figure 1: Diverse human activities in the real-world [3]

Several researchers considered HAR research a pattern recognition problem [4]. Primarily, they focused on traditional machine learning approaches such as Support Vector Machine (SVM), Hidden Markov Models (HMM) and deep learning approaches. Nevertheless, conventional machine learning methods, usually referred to as shallow learning, rely on expert human knowledge to extract data characteristics, restricting the architecture designed for one environment to solving issues in another [5].

On the other hand, deep learning enables direct feature extraction from data without the requirement for expert knowledge or the best feature selection. In contrast, standard manual approaches depend on accurate feature extraction [6]. Chen et al. [7] presented a two-stream CNN and estimated homography between two successive frames without the aid of a human. Given the approximate homography caused by camera motion, background motion may be avoided via the warped optical flow. Two-stream CNN-based video dynamics mining techniques are presented by Liu et al. [8] to extract temporal data. These methods generate action data features by measuring the degree of motion in each video stream frame. At various time scales, the n-skip optical flow extraction approach is used.

The innate ability of humans to do daily activities creates numerous challenges. A person may be engaged in many tasks simultaneously [9]. Individual differences in performance and anthropometrics make action recognition more challenging. Lighting, occlusion, background clutter, and viewpoint alteration all contribute to action recognition difficulty. The speed at which an activity is executed substantially impacts its duration. Temporal variance is another critical issue in detecting human behavior. Researchers are working on solutions to these difficulties and the importance of action recognition in various applications [10].

Recently, Yadav et al. [10] introduced the C2-LSTM, an enhanced version of the LSTM units that perceives motion data together with spatial properties and temporal relationships. UCF101 and HMDB51, well-known benchmarks, are used to assess the network. It supports C2LSTM's ability to capture motion in addition to spatial characteristics and temporal dependencies. Similarly, Gammulle et al. [11] concentrate on LSTM networks for mapping the temporal connection between

prominent spatial elements after learning them using CNN. They assess four fusion techniques for integrating LSTM and convolutional neural network outputs and demonstrate that these techniques beat state-of-the-art algorithms on three UCF11, UCFSports, and jHMDB. Finally, several researchers are studying Recurrent Neural Networks (RNN), LSTM, 3D-ConvNet, and two-Stream networks to extract representational from RGB data and temporal features from optical flow images when handling spatial representations in sequential video data [12].

1.1 Problem Statement

This study identified the following challenges in state-of-the-art research.

- The Primary concern in video data is to capture spatial and temporal information.
- The other challenges are visual appearance variation regarding subjective and objective factors and intra-class and inter-class activity variation.
- Other challenges for building a successful deep learning model include frame redundancy, varied length activity clips, compactness, and discriminative video representations.

1.2 Contribution

This study contributed to the following dimensions for HAR in video sequences.

- First, six state-of-the-art pre-trained models are exploited to find the best model for spatial feature extraction. Second, this study uses dense optical flow following the two-stream ConvNets and BiLSTM to capture long-term dependencies for temporal feature extraction.
- This study proposes a two-stream ConvNet to extract features from video sequences and fine-tunes a deep neural network.
- Transfer learning is applied to extract features while reusing object identification information.
- Seven state-of-the-art optimizers are used to fine-tune the proposed network parameters. Furthermore, two state-of-the-art datasets, UCF101 and HMDB51, are used for evaluation purposes. Finally, an ensemble is presented using max hard voting.

This study is organized as follows: Section 2 focuses on state-of-the-art techniques. Section 3 discusses the methodology in different experimental setups. Section 4 discusses the empirical results achieved. Finally, the paper concludes with the conclusion and future directions in Section 5.

2 Related Work

Recognizing human activity from a video is an essential use of computer vision. Human motion research began in the 1850s with E. J. Marey and E. Muybridge shooting moving people [13]. Various taxonomies for recognizing movement have been proposed [6,14]. The following sections discuss HAR's two cutting-edge learning approaches (Shallow and Deep learning).

2.1 Shallow Learning

An effective generalized predictive model can be produced via a machine learning approach known as model learning with as few compositional layers, as shown in Fig. 2. It needs samples subjected to extensive research and extracted expert characteristics. Shallow learning has been widely used in literature for HAR [15].

Feature representation and categorization are two key components of action recognition systems. Features are extracted from a video to reflect its appearance and distinguish it from other videos

containing distinct behaviors. A multi-class classifier can be learned using feature vectors from training films of all. The extracted feature vector from a test video can be passed to the learned classifier to identify it as one of the actions. Faridee et al. [16] propose DeepconvLSTM with 1-layer shallow for HAR. They validate the proposed architecture on five state-of-the-art HAR datasets. Although methodology decreases training time significantly, to improve performance, at least two-layer LSTM should be used. Moreover, it cannot perform well on sensor-based HAR.

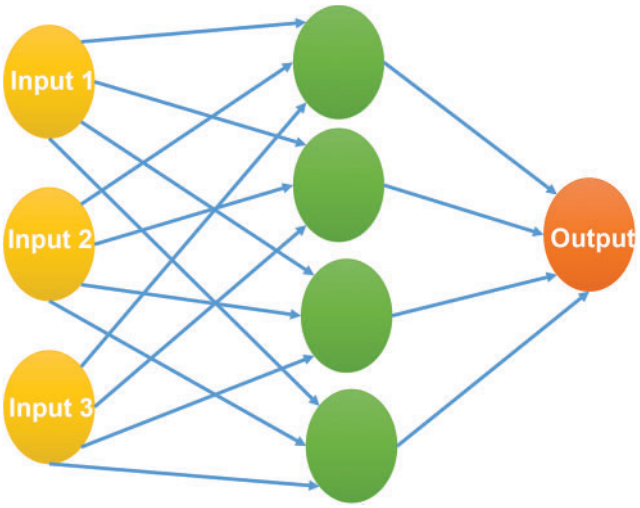

Figure 2: A one-layer shallow learning architecture [14]

2.2 Deep Learning

Recently, deep learning has gained popularity, and several studies have been conducted employing learning in augmented reality. Deep neural networks (DNN), CNN, RNN, and LSTM networks are some of the well-known deep learning techniques used in augmented reality. Zhang et al. [17] investigated DNN, CNN, and RNN for activity datasets and concluded that the RNN outperformed the findings from state-of-the-art techniques.

In deep learning, features are extracted hierarchically using non-linear transformations, as shown in Fig. 3. CNN, RNN, and LSTM networks are all deep neural architectures. In 2014, two seminal research papers were presented by SravyaPranati et al. [18] and Simonyan et al. [12]. They reintroduced deep learning by focusing on single and two-stream deep neural networks for action recognition. In addition, significant research on the identification of human activity was carried out and validated in 2014, utilizing video-based datasets such as UCF101, Sports1M, and HMDB51 [19].

Recently, a CNN-based method was exploited to extract features through a series of max-pooling layers on the DMLSmartActions dataset [20]. Zheng et al. [21] proposed a spatiotemporal module for integrating multi-level CNN features into a hierarchical representation. While the temporal pyramid module pools frames to create several time-grained representations of snippets, the spatial pyramid module extracts multiscale appearance features from video frames. The discriminative hierarchical pooling design incorporates a powerful temporal pooling layer that functions with any CNN architecture [22]. Learning informative dynamics from beginning to end is achievable, thanks to the rich frame-based video representations.

2.2.1 Two-Stream ConvNets

Muhammad et al. [23] perform significant experiments to extend the architecture of CNN with LSTM for large-scale video categorization. They incorporate two separate streams of processing, i.e., they combine features from both streams using three methods. To help, late fusion joins two distant frames with flattened deep hierarchical features. Early fusion stacks images as channels and learns video representations using 2D convolution. However, the suggested fusion networks have a high cost due to several factors.

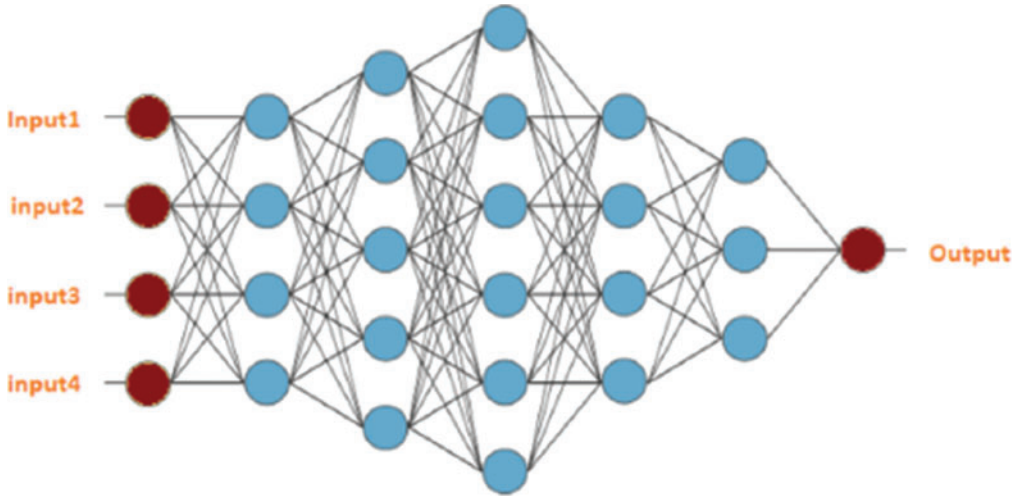

Figure 3: A glimpse of deep learning architecture [14]

Simonyan et al. [12] used two-stream ConvNets to recognize the video-based activity. A proposed two-stream design separates spatial and temporal features. Using 2D convolution, one stream processes an RGB image from a video clip while the other processes a stack of optical flow frames. Following that, the anticipated activity class from both streams is merged. The issue with this network topology is the extra preprocessing and computing required to calculate optical flow pictures. Transfer learning works on spatial streams from pre-trained huge image datasets but not on temporal streams. Despite the increasing computing cost, this approach's strength can be seen by comparing it to advanced techniques like iDT [10]. Zhu et al. [24] identified that every volume, or spatiotemporal video segment, is not identical to the action class. Critical video partitions are identified and excluded from the analysis, allowing substantial video chunks to be omitted. In addition, it incorporates a novel convolutional and temporal fusion layer at later levels of the network to increase performance without raising parameters.

2.2.2 ConvNets with RNNs

The datasets used to train and validate HAR are distinct. HAR video clips vary in length and require extensive sliding window searching. Many researchers now use RNNs and LSTMs to simulate activity changes over time. LRCNs can learn features from variable-length inputs, unlike traditional input methods [25]. Donahue et al. [26] use encoder-decoder architecture to extract video representations. This model uses LSTM cells to build a recurrent neural network from frames extracted at each time step. This approach failed to beat the state-of-the-art but presents a significant single-frame architecture. A hierarchical multiscale RNN can learn the hierarchical temporal structure from data.

Shou et al. [27] proposed two methods for processing full-length video clips by modeling the video as an ordered sequence of frames with RNN and LSTM cells. Efficacious and efficient temporal action localization algorithms can locate action segments of various lengths using exhaustive searching at multiple temporal scales. Deep action proposals (DAPs) extract C3D features from every 16-frame chunk and use LSTM to predict class labels for those activity segments for temporal action localization. Buch et al. [28] designed a single-pass network that directly outputs an activity event's starting and ending bounds and its action class. They utilized a segment CNN framework based on 3D ConvNets. This method divides the localization and classification networks. First, an action recognition classification network is proposed as initialization for the localization network, which then localizes action occurrences in time and predicts scores for action labels.

In a paradigm for HAR presented by Ullah et al. [29], saliency representations are first recovered from continuous video streams after they have been separated into their fundamental shots. The CNN method suggests significant shots. To express temporal aspects, FlowNet2 is then employed. Finally, a multilayer LSTM is used to learn temporal sequences. Khan et al. [30] proposed a 3-stream CNN and achieved state-of-the-art results on the Kinetics-400 dataset. They extract salient patches from appearance and motion representations, three-stream CNN abstract features, integrate frame-level descriptors into an RNN, and combine three classification scores to predict the final class labels.

2.2.3 3D ConvNets

Crasto et al. [31] use the C3D network to handle temporal activity localization. This model improves activity detection by combining an optical flow-based motion stream with the original RGB stream. In addition, an online hard mining strategy improves performance and speed detection.

He et al. [32] proposed Factorized Spatio-Temporal Convolutional Networks (FSTCN) to simplify 3D convolutional kernel learning. With this technique, you can learn spatial and temporal features by learning spatial representations from 2D spatial kernels in the lower layers and temporal representations from 1D temporal kernels in the upper layers. 3D convolution improves feature learning but increases parameter costs. With MicroNets and asymmetric one-directional 3D convolutions, Ullah et al. [5] overcame this problem. They proposed that micro nets improve feature learning by incorporating multiscale 3D convolution branches to handle the different scales of convolutional features in videos.

Fin et al. [33] propose an enhanced CNN based on group and depth convolution. This method can produce 3D convolutions for action recognition in video classification. 3D Channel Separated Networks (CSN) use depth-wise convolutions for local spatiotemporal connections. It reduces network parameters and introduces a strong form of regularization. Discriminative spatial and temporal features can be learned in 14 layers using these channel-separated blocks. 3D ResNet employs RGB images with motion and scenes to create patch features.

3 Methodology

The primary objective of this study is to classify human activities from video sequences. In addition to a fixed, clean background, video recording of human activities includes varying camera motions, lighting conditions, and low-quality frames. This study uses spatial and temporal information to recognize activities effectively. Moreover, a transfer learning paradigm is adopted to extract features from varied and fixed frames. With the success of LSTM and its variants, this study fine-tunes BiLSTM on extracted features for learning long-term dependencies. Finally, ensemble learning is utilized to combine all these streams. Fig. 4 provides a comprehensive explanation of the process. In addition,

accuracy is used as a performance measure to evaluate the proposed methodology. The following sections discuss the methodology comprehensively.

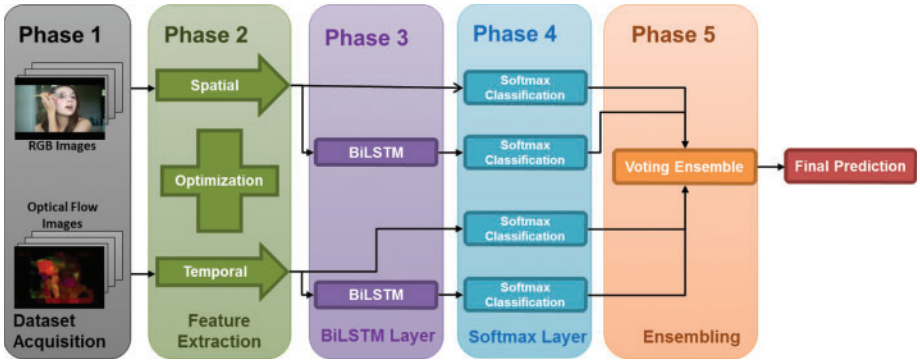

Figure 4: A proposed approach for our four-stream ensemble

3.1 Dataset Acquisition

This study uses two state-of-the-art HAR datasets UCF101 and HMDB-51. UCF101 is a collection of real-world action videos from YouTube that have been categorized into 101 different action types. The most difficult data set to date, UCF101, has 13320 videos from 101 action categories and offers the widest variety of actions. It also has significant variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, and other factors.

UCF101 intends to promote further action recognition research by learning and examining new realistic action categories because most of the action recognition data sets that are currently accessible are not realistic and are staged by actors. The videos in the 101 action categories are divided into 25 groups, each of which may contain 4–7 action videos. Table 1 shows the characteristics of the UCF101 dataset.

Table 1: UCF101 dataset characteristics [34]

Parameter	Values
Number of actions	101
Clips	13320
Groups per action	25
Clips per group	4–7
Mean clip length	7.21 s
Total duration	1600 mins
Min clip length	1.06 s
Max clip length	71.04 s
Frame rate	25 fps
Resolution	320 ∗ 240
Audio	Yes (51 actions)

Most of the HMDB-51 [35] data was gathered from films, with a tiny amount coming from public databases like the Prelinger archive, YouTube, and Google videos. The dataset comprises 6849 clips, classified into 51 action categories, each of which has at least 101 clips, as shown in Table 2.

Table 2: HMDB-51 dataset characteristics [35]

Parameter	Values
Number of actions	51
Clips	6849
Background	Dynamic
Camera motion	Yes
Release year	2011
Source	Movies, YouTube, Web
Clips per action	Min. 101

3.2 Feature Extraction

The feature extraction method can be quite helpful when analyzing massive datasets with only limited computational resources. The process of extracting essential data features from data and transforming them into condensed representations is referred to as feature extraction. The extracted characteristics must be simple to process, discriminatory, and accurately describe the real data set. In addition to this benefit, extracting features helps reduce redundant data. The goal of this research is to derive Spatio-temporal information from video data.

This research uses off-the-shelf feature extractors. These layers are combined to get the final prediction. Pre-trained networks without their final layer can be fixed feature extractors. Raw pixel data is used to extract features. This study uses the pre-trained model's weighted layers to extract features but doesn't change their weights during training. Instead, unsupervised, image-class-unrelated feature extraction is used. Training will change weights from generic feature maps to dataset-specific features.

This technique replaces not just the final layer (for classification and regression) but also certain prior levels. Initial layers record generic features, whereas later ones focus on the specific task. This study uses six popular off-the-shelf pre-trained networks to extract features. All these models are evaluated on the UCF101 and HMDB51 datasets. Table 3 shows the pre-trained model specifications.

Table 3: State-of-the-art pre-trained models specifications

Model	Total parameters	Feature shape
VGG16	14,714,688	$7 * 7 * 512$
InceptionV3	21,802,784	$5 * 5 * 2048$
ResNet101V2	42,626,560	$7 * 7 * 2048$
InceptionResNetV2	54,336,736	$5 * 5 * 1536$
DenseNet121	7,037,504	$7 * 7 * 1024$
NASNetMobile	4,269,716	$7 * 7 * 1056$

4 Result and Discussion

This section focuses on human activities captured with an RGB camera with many real-life applications. The main challenge with video data is capturing both spatial and temporal information. In addition, there is intra-class and inter-class activity variation. Frame redundancy, variable-length activity clips, compactness, and discriminative video representations are all challenges in deep learning. Therefore, this study proposes a generic, dense computational model.

4.1 Transfer Learning with Deep Learning as a Feature Extractor

Researchers are struggling to design systems that can apply experience learned from previous tasks to improve performance on a new task. Transfer learning can reduce the architecture learning time and effort, resulting in more robust and useful activity recognition systems [25]. Muhammad et al. [23] reveal that transfer learning is a good approach to action recognition. Two popular strategies for deep transfer learning are off-the-shelf pre-trained models and off-the-shelf pre-trained models as feature extractors.

This study uses Off-the-shelf models as feature extractor strategies, as shown in Table 3. Deep learning models have layered architectures with distinct layers learning different characteristics. Finally, for the final forecast, these layers are linked to the last layer. Using a pre-trained network without its final layer as a fixed feature extractor is made possible by the layered design. The important concept is to use the pre-trained model's weighted layers to extract features and not to change the layers' weights while the model is trained with fresh data for the current job.

Since the classes of the image have no bearing on the information recovered from pixels, the feature extraction is unsupervised. A new classifier is introduced and trained from scratch on top of the pre-trained model to repurpose previously learned feature maps for the UCF101 and HMDB51 datasets. Features often helpful for categorization are already present in the base network.

4.1.1 Spatial Feature Extraction

Spatial features exploit location/spatial data. CNN performs non-linear transformations in different locations of the image in the form of convolutions. VGG16, InceptionV3, ResNet101V2, InceptionResNet2 and NASNetMobile were used to extract spatial features from UCF101 and HMDB-51. Spatial features are extracted from RGB images. For RGB data, each video extracts a single frame per second. The network accepts 224 × 224 × 3 images, as shown in Fig. 5.

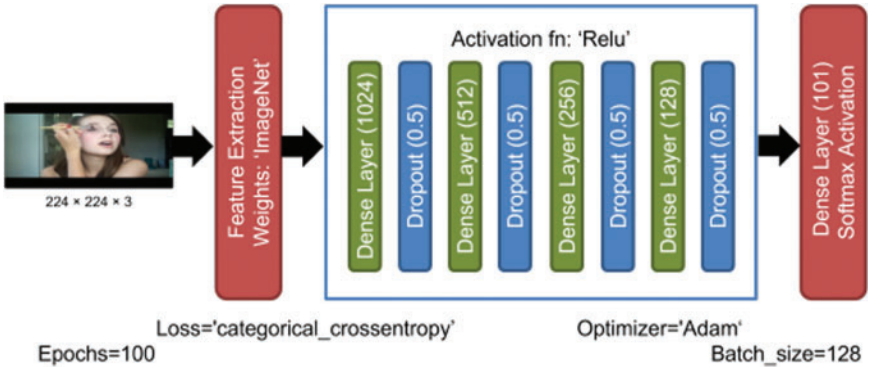

Figure 5: Network configuration for the pre-trained model to extract features

After extracting features, the classifier has four dense layers with ReLU activation and a dropout of 0.5. SoftMax does the final classification. Each model was compiled with categorical cross-entropy loss and an Adam optimizer with 128 batch sizes. The same network configuration has been used for temporal feature extraction. Tables 4 and 5 show the experimental results obtained by applying selected pre-trained models on RGB images of HMDB51 and UCF101 datasets. Moreover, it shows that ResNet101V2 achieved high accuracy compared to other state-of-the-art pre-trained models on RGB images.

Table 4: Results achieved by applying selected pre-trained to UCF101 dataset RGB images

Model	Training accuracy	Validation accuracy	Test accuracy
VGG16	78.59	85.44	28.89
InceptionV3	96.99	96.45	69.18
ResNet101V2	**98.92**	**98.52**	**76.46**
InceptionResNetV2	97.72	97.58	71.72
DenseNet121	97.79	98.05	70.61
NASNetMobile	98.48	97.5	67.62

Table 5: Results achieved by applying selected pre-trained to HMDB-51 dataset RGB Images

Model	Training accuracy	Validation accuracy	Test accuracy
VGG16	69.03	31.08	17.39
InceptionV3	82.31	37.07	31.83
ResNet101V2	**99.0**	**44.04**	**36.86**
InceptionResNetV2	85.02	39.43	35.1
DenseNet121	92.05	40.81	32.85
NASNetMobile	53.25	36.47	20.20

4.1.2 Temporal Feature Extraction

Temporal characterization occurs when a series of images are taken at different times. Correlations between the images are often used to monitor the dynamic changes of the object. We extracted a fixed number of frames per second from each clip, as per-second frame extraction did not yield good results. Therefore, we extract eight frames per second from each video clip.

The pattern of apparent mobility of picture objects between two successive frames brought on by the movement of the subject or camera is known as optical flow. Each vector in the 2D vector field represents the displacement of a point from the first to the second frame. Dense optical flow may be slow but produces more accurate results since it computes the optical flow vector for each pixel in the frame. This study made dense optical flow calculations using the Franeback technique [36]. In addition, pre-trained ResNet50V2 has been exploited to extract temporal features due to its classification accuracy compared to the state-of-the-art [37]. Results were obtained by extracting eight video frames from each clip of the UCF101 and HMDB51 datasets, as shown in Tables 6 and 7.

Table 6: Results of ResNet50V2 on the UCF101 dataset by extracting eight optical flow images per clip

Model	Training accuracy	Validation accuracy	Test accuracy
ResNet50V2	87	40.41	79.78

Table 7: Results of ResNet50V2 on the HMDB51 dataset by extracting eight optical flow images per clip

Model	Training accuracy	Validation accuracy	Test accuracy
ResNet50V2	92.1	24.30	40.07

4.2 Optimization

This study utilized seven state-of-the-art optimizers from the Keras library to fine-tune the architecture parameters, i.e., SGD [38], Adagrad [39], Adadelta [40], RMSprop [41], Adam [42], Adamax [43], Nadam [44]. After extracting features through ResNet, all these optimizers are used to compile the model with two dense layers for final prediction. AdaMax achieves the highest test accuracy among all, as shown in Fig. 6.

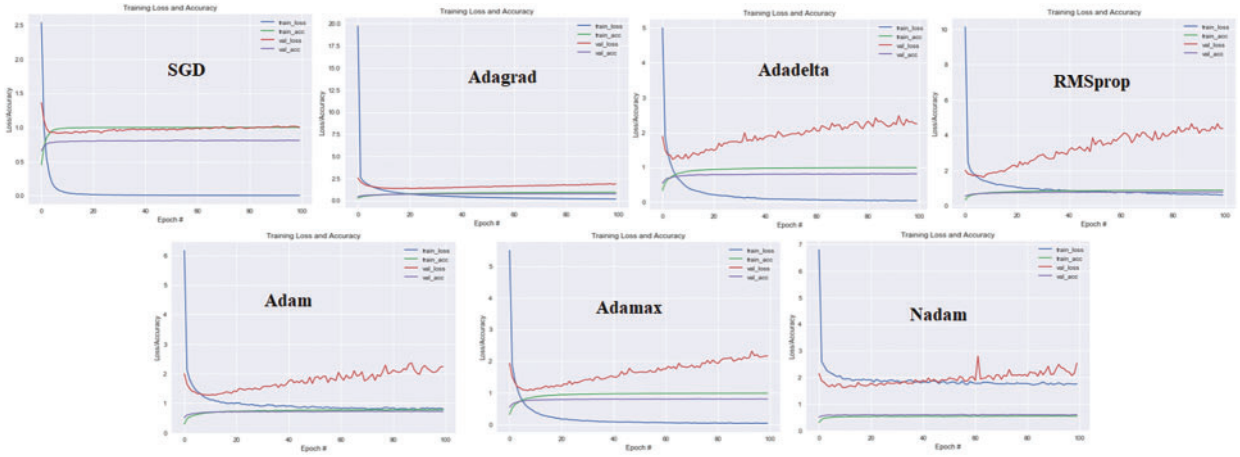

Figure 6: Line graph representing training loss, validation loss, and accuracy across 100 epochs

This study uses the AdaMax optimizer to accelerate the optimization process. AdaMax is a modification to the Adam version of gradient descent that generalizes the technique and leads to a more effective optimization for certain applications [45]. In its formulation, Adam, like RMSProp, uses an exponentially decaying weighted average estimate of the gradient's variance. The update rule for individual weights in Adam is to scale their gradients inversely proportionate to the (scaled) L2 norm of their current and previous gradients. In addition, AdaMax automatically adjusts each optimization problem parameter's step size (learning rate). A comparison of training accuracy and loss per epoch is shown in Fig. 6. AdaMax achieves 86.14% accuracy on HMDB51 and 92.57% on UCF101 datasets.

4.3 Activity Recognition Using BiLSTM

This section discusses the results obtained using the BiLSTM model for activity recognition. BiLSTM processes the input data in two directions. Once from the beginning to the end and once from the end to the beginning. BiLSTMs are an extension of LSTM models designed to collect input information in the past and the future of a given time stamp [46]. To apply BiLSTM, we extract eight frames from each video clip, extract features using ResNet50V2 and apply a single layer of BiLSTM with 256 units, followed by a dropout and a dense layer with a Softmax activation function. Finally, the BiLSTM layer is applied to both RGB and Optical Flow (OF) data, and the results are shown in Tables 8 and 9.

Table 8: Results obtained after applying BiLSTM on features extracted with ResNet50V2 on UCF-101

Model	Training accuracy	Validation accuracy	Test accuracy
BiLSTM (RGB)	95.67	85.80	92.12
BiLSTM (OF)	86.02	38.14	72.85

Table 9: Results obtained after applying BiLSTM on features extracted with ResNet50V2 on HMDB-51

Model	Training accuracy	Validation accuracy	Test accuracy
BiLSTM (RGB)	87.86	58.20	76.47
BiLSTM (OF)	73.30	21.32	61.63

4.4 Activity Recognition Using Ensemble

Deep learning models are non-linear learning processes that use a stochastic training algorithm to learn. Given adequate resources, they approximate any mapping function and are adaptable, learning intricate correlations between variables [4]. The models have significant variations because of this flexibility. By creating many models for the issues at hand and aggregating their predictions, the high volatility of the method can be reduced. The procedure of prediction aggregation belongs to a family of ensemble learning techniques.

This study uses ensemble learning to aggregate spatial and temporal information. The ensemble is applied to a four-stream CNN, where two streams use RGB data while the other uses optical flow images. Optical flow images are extracted to capture the movement or change in an action sequence. The first stream extracts one RGB image per second for varied-length clips and extracts features through ResNet101. The second stream extracts 8 RGB frames per video, extracts features through ResNet-50V2, and applies the BiLSTM layer on top. The third stream consists of extracting eight optical flow images per video, extracting features using ResNet50V2, and classifying them accordingly.

The fourth stream applies BiLSTM on top and then classifies the extracted features of eight optical flow images. Finally, the ensemble is applied using hard majority voting on all four streams. Ensemble applied using hard max voting works best if all models have improved performance, as shown in Fig. 7. The results obtained from the ensemble of those four streams on the UCF-101 and HMDB-51 datasets are shown in Tables 10 and 11. Experimental results show the supremacy of the proposed ensemble compared to other models.

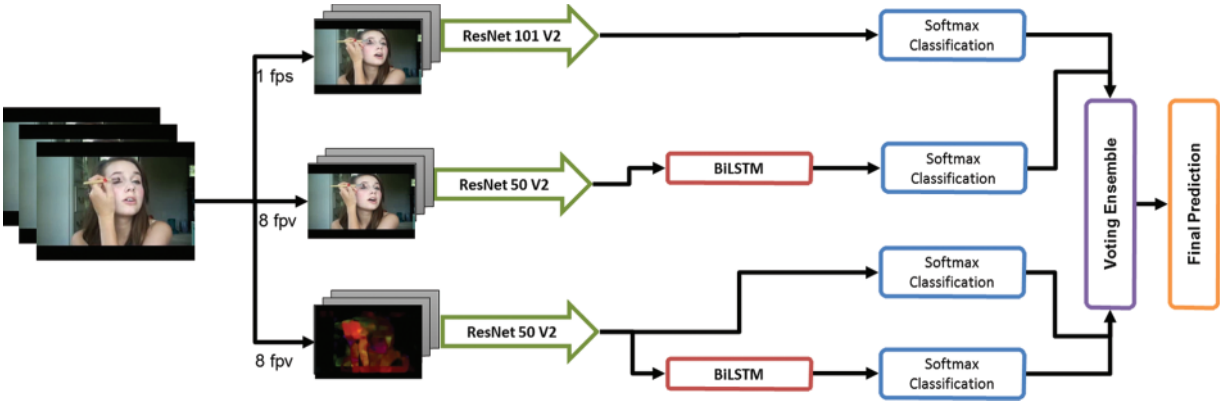

Figure 7: Activity recognition using a four-stream ensemble

Table 10: Results obtained on UCF-101 for all four streams and their ensemble

Model	Training accuracy	Validation accuracy	Test accuracy
ResNet101 (RGB)	94.18	92.07	92.57
ResNet50 + BiLSTM (RGB)	95.67	85.80	92.12
ResNet50 (OF)	87	40.41	79.78
ResNet50 + BiLSTM (OF)	86.02	38.14	72.85
Ensemble (RGB + OF)	**94.20**	**93.32**	**96.30**

Table 11: Results obtained on HMDB-51 for all four streams and their ensemble

Model	Training accuracy	Validation accuracy	Test accuracy
ResNet101 (RGB)	88.64	71.17	86.14
ResNet50 + BiLSTM (RGB)	87.86	58.20	76.47
ResNet50 (OF)	92.1	24.30	60.07
ResNet50 + BiLSTM (OF)	73.03	21.30	61.63
Ensemble (RGB + OF)	**89.41**	**83.32**	**90.07**

Two-stream ConvNets extract spatial information from RGB, and temporal information is extracted using dense optical flow images calculated on equidistant frames. Moreover, a transfer learning paradigm is adopted to extract features while reusing object identification knowledge for activity recognition. We evaluated six state-of-the-art networks: VGGNet, Inception, ResNet, InceptionResNet, DenseNet, and NASNet. These models are tested on the UCF101 and HMDB51 datasets. ResNet101V2 achieved 72.46% and 36.86% accuracy on the UCF101 and HMDB51 datasets, respectively. In addition, seven state-of-the-art optimizers are utilized for network parameter optimization. The AdaMax optimizer enhanced the accuracy by achieving 92.57% and 86.14% on the UCF-101 and HMDB-51 datasets.

This study uses BiLSTM and ensemble methods for activity recognition. BiLSTM captures long-term sequential information to explore temporal information. BiLSTM achieves 92.12% and 76.47%

accuracy on RGB, while 72.85% and 61.63% accuracy on optical flow. Finally, ensemble learning is used to fuse all streams using max hard voting. After applying the ensemble, the accuracy on UCF101 is increased to 96.3% and 90.07% on HMDB51. The proposed ensemble-based approach shows supremacy compared to state-of-the-art methods, as shown in Table 12.

Table 12: Comparison of the proposed approach with the state-of-the-art on UCF101 and HMDB51

Comparison on UCF101	Accuracy	Comparison on HMDB51	Accuracy
3DConvNets ensemble + iDT [47]	92.70	Four-stream I3D [50]	99.1
Two-stream [48]	94.80	CSTA-I3D [51]	98.2
CatNet [49]	96.0	WMHI [52]	97.1
ResNet (RGB)	92.57	ResNet (RGB)	86.14
ResNet + BiLSTM (RGB)	92.12	ResNet + BiLSTM (RGB)	76.47
ResNet (OF)	79.78	ResNet (OF)	60.07
ResNet + BiLSTM (OF)	72.85	ResNet + BiLSTM (OF)	61.63
Proposed method (Ensemble (RGB + OF))	**96.30**	**Proposed method (Ensemble (RGB + OF))**	**90.07**

5 Conclusion and Future Work

HAR from video sequences is challenging due to the dynamic backgrounds and complexity of real-world applications such as video surveillance, HCI, and human behavior characterization. However, deep learning methods such as convolutional and recurrent neural networks have recently achieved state-of-the-art results by automatically learning features from raw sensor data.

This study focused on grouped and individual human activity in video sequences acquired with an RGB camera because of its vast range of real-world applications. It uses two-stream ConvNet to extract spatial and temporal information and proposes a fine-tuned deep neural network. Primarily, transfer learning is applied to extract features while reusing object identification information. First, six state-of-the-art pre-trained models are exploited to find the best model for spatial feature extraction. Second, this study uses dense optical flow following the two-stream ConvNets and BiLSTM to capture long-term dependencies for temporal feature extraction.

In addition, seven state-of-the-art optimizers are used to fine-tune the proposed network parameters. Furthermore, two state-of-the-art datasets, UCF101 and HMDB51, are used for evaluation purposes. Finally, the proposed ensemble approach using max hard voting outperforms state-of-the-art methods with 96.30% and 90.07% accuracies on the UCF101 and HMDB51 datasets. In the future, we aim to detect human activities from a multimodal dataset using 3D-ConvNets by fusing temporal and spatial features. Moreover, transfer learning will be adopted with self-distillation to enhance HAR accuracy.

Data Availability Statement: This study uses UCF101 and HDMB51 datasets, which are publicly available.

Funding Statement: This work was supported by financial support from Universiti Sains Malaysia (USM) under FRGS grant number FRGS/1/2020/TK03/USM/02/1 and the School of Computer Sciences USM for their support.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] V. Choutas, P. Weinzaepfel, J. Revaud and C. Schmid, "Potion: Pose motion representation for action recognition," in *Proc. of CVPR*, Salt Lake City, Utah, USA, pp. 7024–7033, 2018.

[2] L. Onofri, P. Soda, M. Pechenizkiy and G. Iannello, "A survey on using domain and contextual knowledge for human activity recognition in video streams," *Expert Systems with Applications*, vol. 63, pp. 97–111, 2016.

[3] C. Jobanputra, J. Bavishi and N. Doshi, "Human activity recognition: A survey," *Procedia Computer Science*, vol. 155, pp. 698–703, 2019.

[4] J. Wang, Y. Chen, S. Hao, X. Peng and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019a.

[5] H. A. Ullah, S. Letchmunan, M. S. Zia, U. M. Butt and F. H. Hassan, "Analysis of deep neural networks for human activity recognition in videos–a systematic literature review," *IEEE Access*, vol. 9, pp. 126366–126387, 2021.

[6] B. Nguyen, Y. Coelho, T. Bastos and S. Krishnan, "Trends in human activity recognition with focus on machine learning and power requirements," *Machine Learning with Applications*, vol. 5, pp. 100072, 2021.

[7] Q. -Q. Chen, F. Liu, X. Li, B. -D. Liu and Y. -J. Zhang, "Saliency-context two-stream convnets for action recognition," in *IEEE ICIP*, Phoenix, Arizona, USA, pp. 3076–3080, 2016.

[8] Y. Liu, Q. Wu and L. Tang, "Frame-skip convolutional neural networks for action recognition," in *503 IEEE ICMEW*, London, U.K., vol. 14/16, pp. 573–578, 2017.

[9] J. Charmi, J. Bavishi and N. Doshi, "Human activity recognition: A survey," *Procedia Computer Science*, vol. 155, pp. 698–703, 2019.

[10] S. K. Yadav, K. Tiwari, H. M. Pandey and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, pp. 106970, 2021.

[11] H. Gammulle, S. Denman, S. Sridharan and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *IEEE WACV*, Santa Rosa, CA, pp. 177–186, 2017.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, vol. 27. Montreal, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), pp. 568–576, 2014a.

[13] J. D. Marshall, T. Li, J. H. Wu and T. W. Dunn, "Leaving flatland: Advances in 3D behavioral measurement," *Current Opinion in Neurobiology*, vol. 73, pp. 102522, 2022.

[14] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra and A. Kumar, "A review of deep learning-based human activity recognition on benchmark video datasets," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 2093705, 2022.

[15] M. Bock, A. Holzemann, M. Moeller and K. Van Laerhoven, "Improving deep learning for har with shallow lstms," in *Int. Symp. on Wearable Computers*, Virtually, USA, pp. 7–12, 2021.

[16] A. Z. M. Faridee, S. R. Ramamurthy, H. S. Hossain and N. Roy, "Happyfeet: Recognizing and assessing dances on the floor," in *Proc. of HotMobile*, Tempe Arizona, USA, pp. 49–54, 2018.

[17] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia *et al.,* "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, no. 4, pp. 1476, 2022.

[18] B. SravyaPranati, D. Suma, C. ManjuLatha and S. Putheti, "Large-scale video classification with convolutional neural networks," in *Int. Conf. on Information and Communication Technology for Intelligent Systems*, Ahmedabad, India, pp. 689–695, 2020.
[19] A. Khelalef, F. Ababsa and N. Benoudjit, "An efficient human activity recognition technique based on deep learning," *Pattern Recognition and Image Analysis*, vol. 29, no. 4, pp. 702–715, 2019.
[20] H. D. Mehr and H. Polat, "Human activity recognition in smart home with deep learning approach," in *7th ICSG*, Istanbul, Turkey, pp. 149–153, 2019.
[21] Z. Zheng, G. An, D. Wu and Q. Ruan, "Spatial-temporal pyramid based convolutional neural network for action recognition," *Neurocomputing*, vol. 358, pp. 446–455, 2019.
[22] B. Fernando and S. Gould, "Discriminatively learned hierarchical rank pooling networks," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 335–355, 2017.
[23] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran *et al.,* "Human action recognition using attention based LSTM network with dilated CNN features," *Future Generation Computer Systems*, vol. 125, pp. 820–830, 2021.
[24] W. Zhu, J. Hu, G. Sun, X. Cao and Y. Qiao, "A key volume mining deep framework for action recognition," in *Proc. of CVPR*, Las Vegas, NV, USA, vol. 591, pp. 1991–1999, 2016.
[25] M. Bilal, M. Maqsood, S. Yasmin, N. U. Hasan and S. Rho, "A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes," *The Journal of Supercomputing*, vol. 78, no. 2, pp. 2873–2908, 2022.
[26] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan *et al.,* "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. of CVPR*, Boston, MA, USA, pp. 2625–2634, 2015.
[27] Z. Shou, D. Wang and S. -F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proc. of CVPR*, Las Vegas, NV, USA, pp. 1049–1058, 2016.
[28] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," Ph.D. Dissertation, KSA, KAUST, Saudi Arabia, 2019.
[29] A. Ullah, K. Muhammad, J. D. Ser, S. W. Baik and V. H. C. Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer lstm," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.
[30] M. A. A. H. Khan, R. Kukkapalli, P. Waradpande, S. Kulandaivel, N. Banerjee *et al.,* "Ram: Radar-based activity monitor," in *IEEE 35th INFOCOM*, San Francisco, CA, USA, pp. 1–9, 2016.
[31] N. Crasto, P. Weinzaepfel, K. Alahari and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in *Proc. of the IEEE CVPR*, Utah, USA, pp. 7882–7891, 2019.
[32] K. He, X. Zhang, S. Ren and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, Amsterdam, Netherlands, pp. 630–645, 2016.
[33] Y. Fin, Z. Lan, S. Newsam and A. Hauptmann, "Hidden two-stream convolutional networks for 592 action recognition," in *ACCV*, Kyoto, Japan, pp. 363–378, 2018.
[34] K. Soomro, A. R. Zamir and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, Harvard, US, 2012a.
[35] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 ICCV*, Washington, DC, United States, pp. 2556–2563, 2011.
[36] G. Farneback, "Fast and accurate motion estimation using orientation tensors and parametric motion models," in *Proc. 15th ICPR-2000*, Barcelona, Spain, vol. 1, pp. 135–139, 2000.
[37] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting covid19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2," *Informatics in Medicine Unlocked*, vol. 19, pp. 100–360, 2020.
[38] I. Sutskever, J. Martens, G. Dahl and G. Hinton, "On the importance of initialization and momentums in deep learning," in *ICML*, USA, pp. 1139–1147, 2013.
[39] J. Duchi, E. Hazan and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimizations," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 2121–2159, 2011.

[40] M. D. Zeiler, "Adadelta: An adaptive learning rate method," arXiv preprint arXiv, pp. 1212–5701, 2012.
[41] G., Hinton, N. Srivastava and K. Swersky, "Rmsprop: A mini-batch version of rprop," *Coursera475 course lecture 6-Neural Networks for Machine Learning*. 2012. https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv, pp. 1412–6980, 2014.
[43] R. Llugsi, S. El Yacoubi, A. Fontaine and P. Lupera, "Comparison between adam, adamax and adam w optimizers to implement a weather forecast based on neural networks for the andean city of Quito," in *IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, NYC, USA, pp. 1–6, 2021.
[44] T. Nguyen, R. Baraniuk, A. Bertozzi, S. Osher and B. Wang, "Momentumrnn: Integrating momentum into recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1924–1936, 2020.
[45] S. Vani and T. M. Rao, "An experimental approach towards the performance assessment of various optimizers on convolutional neural network," in *3rd ICOEI*, Tirunelveli, India, pp. 331–336, 2019.
[46] Y. Li and L. Wang, "Human activity recognition based on residual network and bilstm," *Sensors*, vol. 22, no. 2, pp. 635, 2022.
[47] Y. Huang, Y. Guo and C. Gao, "Efficient parallel inflated 3D convolution architecture for action recognition," *IEEE Access*, vol. 8, pp. 45753–45765, 2020.
[48] Y. Wan, Z. Yu, Y. Wang and X. Li, "Action recognition based on twostream convolutional networks with long-short-term spatiotemporal features," *IEEE Access*, vol. 8, pp. 85284–85293, 2020.
[49] J. Wang, X. Peng and Y. Qiao, "Cascade multi-head attention networks for action recognition," *Computer Vision and Image Understanding*, vol. 192, pp. 102–898, 2020.
[50] S. A. Khowaja and S. L. Lee, "Semantic image networks for human action recognition," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 393–419, 2020.
[51] Y. Zhu and G. Liu, "Fine-grained action recognition using multi-view attentions," *The Visual Computer*, vol. 36, no. 9, pp. 1771–1781, 2020.
[52] S. Chaudhary and S. Murala, "Deep network for human action recognition using weber motion," *Neurocomputing*, vol. 367, pp. 207–216, 2019.