Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.035969
Article

Tech Science Press

Adaptive Weighted Flow Net Algorithm for Human Activity Recognition Using Depth Learned Features

G. Augusta Kani* **and P. Geetha**

Department of Information Science and Technology, Anna University, Chennai, 600025, Tamilnadu, India
*Corresponding Author: G. Augusta Kani. Email: augus.jesus@gmail.com

Received: 12 September 2022; Accepted: 25 November 2022

Abstract: Human Activity Recognition (HAR) from video data collections is the core application in vision tasks and has a variety of utilizations including object detection applications, video-based behavior monitoring, video classification, and indexing, patient monitoring, robotics, and behavior analysis. Although many techniques are available for HAR in video analysis tasks, most of them are not focusing on behavioral analysis. Hence, a new HAR system analysis the behavioral activity of a person based on the deep learning approach proposed in this work. The most essential aim of this work is to recognize the complex activities that are useful in many tasks that are based on object detection, modelling of individual frame characteristics, and communication among them. Moreover, this work focuses on finding out the human actions from various video resolutions, invariant human poses, and nearness of multi objects. First, we identify the key and essential frames of each activity using histogram differences. Secondly, Discrete Wavelet Transform (DWT) is used in this system to extract coefficients from the sequence of key-frames where the activity is localized in space. Finally, an Adaptive Weighted Flow Net (AWFN) algorithm is proposed in this work for effective video activity recognition. Moreover, the proposed algorithm has been evaluated by comparing it with the existing Visual Geometry Group (VGG-16) convolution neural networks for making performance comparisons. This work focuses on competent deep learning-based feature extraction to discriminate the activities for performing the classification accuracy. The proposed model has been evaluated with VGG-16 using a combination of regular UCF-101 activity datasets and also in very challenging Low-quality videos such as HMDB51. From these investigations, it is proved that the proposed AWFN approach gives higher detection accuracy of 96%. It is approximately 0.3% to 7.88% of higher accuracy than state-of-art methods.

Keywords: Activity classification; discrete wavelet; object detection; AWFN

1 Introduction

Human Activity Recognition (HAR) from videos is an important and challenging activity in the contemporary period the reason of the difficulty of intricate procedures used. Numerous investigations have been conceded out in research repositories on automated HAR. In general, object-based techniques are useful for the effective identification of human activities. Moreover, physical activities are performed based on movements and complex actions. Here the activities are also varying based on both the temporal-spatial dimensions. Different scenarios such as sensor-based, image-based and video-based techniques are widely used for the detection of human activity. In this work, a video processing-based scenario is performed in which HAR can be performed for distinguishing human actions. The methods used for HAR from the video features are classified into furnish and Deep Learning (DL) techniques. Among them, the DL algorithm namely the Convolution Neural Network (CNN) is mainly employed in image recognition problems as they use deep knowledge features. The number of layers in the design of DL-based classifiers is to decide using the application requirements. The main motivation of this activity detection is to detect human-object interaction, human-human interaction, body motion recognition, and sports activity recognition. It can be used in areas like sports, motion recognition and to know about the activities performed in videos.

The CNN draws out the deep-off features from every level using filters. Moreover, the CNN consists of several hidden layers and a normalized output layer. Since DL algorithms are efficient in image recognition and video analysis, they can be used in object-tracking applications in video analytics. In this approach, tracing is focusing on finding and creating the associations using the analysis of corrections among the frames. Conventional Machine Learning (ML) algorithms identify the important and most contributing features based on low-level correspondences. The analysis of the low-level correspondences is an important component in Feature Selection (FS) based on machine learning algorithms. In addition, DL algorithms are provided with many hidden layers that are specialized in the extraction and selection of many features. Here, the DL algorithms can identify the most contributing features from input video through their hidden layers [1–3]. One problem with the current classifiers is that they do not perform semantic analysis to get the semantically correct features. Therefore, existing DL algorithms must be enhanced to make them more suitable for HAR.

A system developed for HAR using time and location constraints is a challenging task and hence, it is a type of time series classification problem. Therefore, the work proposed in this paper considers spatial and temporal constraints in the classification process. For accomplishing the analysis of the video data about HAR more accurately, this article proposes a spatial-temporal constraint-based method for solving the classification problem by extending the DL algorithm namely, the Long Short-Term Memory (LSTM) which is used for performing the optical flow analysis based on spatial-temporal features applied to form a new classification algorithm called Adaptive Weighted Flow Net based Long Short-Term Memory (AWFN-LSTM) algorithm. Moreover, the proposed AWFN-LSTM is a DL algorithm and it is a type of Recurrent Neural Network (RNN) algorithm. As a result, it is well suited for handling temporal and geographical limitations. This proposed algorithm was implemented and tested using one regular UCF-101 activity data set and low-quality videos, namely HMDB51. Moreover, we extracted the key and important frames from the video data for every human activity using the difference values from the histogram. In addition, Discrete Wavelet Transform (DWT) is employed in this proposed model for extracting the coefficients more accurately from the sequence of key-frames that were extracted from the video files. The major advantages of the proposed system are that it improves the classification and prediction accuracy when it is related to the other important deep learning systems namely VGG-16 convolution neural networks and the CNN.

The proposed HAR system includes the following unique recommendations and contributions. The suggested method is a new contribution to this field because it uses a combination of deep learning and machine learning to perform spatial and temporal reasoning that accomplishes activity detection. In general, deep learning-based methods can handle uncertainty during detection. To tackle uncertainty during activity detection, the proposed wavelet-based VGG-16 with LSTM and AWFN uses the appropriate reasoning. The AWFN algorithm can detect human activities more precisely than traditional methods.

The rest of this present article is organized as follows: The second portion covers the existing and relevant work in the areas of human activity acknowledgement, feature assortment, and classification methods. Section 3 details the contributions of the proposed work. Section 4 depicts the results obtained from this work with suitable discussions. Section 5 derives the conclusions on this work and gives some possible future works.

2 Related Works

In recent times, many approaches were developed for HAR. In this scenario, multi-view HAR [4] has been projected as the important activity and it utilizes the geometrical features based on regions such as critical points, area, perimeter, orientation and hu-moments. It extracts a lot of features and they are used to find the activity.

Seo et al. [5] proposed a new approach to reduce the complexity of optical flow measurement by implementing the dynamic frame bouncing method and motion interjection for efficient optical flow approximation. For that Trajectory Shapes (TS), HOG, HOF, and Motion Boundary Histograms (MBH) features were extracted and PCA was used for performing the descriptors reduction. They encoded the extracted features utilizing Fisher Vectors (FV) and the Sparse Representation Classification (SRC) was used in their work to differentiate the human activities. Their model increased the overall computational efficiency and provided high accuracy in the categorization of activities in the Olympic sports dataset by providing 95.3% accuracy. But still, some errors are found in their model due to the skipping of frames.

Ullah et al. [6] found and used the texture features of activity for extracting the significant facts and to create recognition using it. Their system extracts interesting regions by integrating object proposals and motion cues. The shape features were extracted using histograms and gist features were used for motion cues. Extreme Learning Machine (ELM) was a classifier that was used by these authors to recognize the activity more accurately. Here, the mid-level features were represented well and hence it has increased the recognition accuracy to 95.6%. However, the object proposal method they developed was computationally expensive with high time complexity. Zhao et al. [7] resolved the issue of human activity acknowledgement in low-level recordings. In their model, layered elastic-motion tracking was suggested to capture both long-term trajectories of motion and mutual shape. Hybrid feature depiction was used for cording both the motion and the shape features. Action classification is carried out effectively by the region mixture model and in which there was no necessity for segmentation. The shortcoming includes the encryption of the spatial outlines of the features. Peng et al. [8] worked on boosting the activity recognition performance using descriptor fusion. They developed BoW visual model for HAR as it provides effective acknowledgement based on actions acknowledgement. They carried out Principal Component Analysis (PCA) as a feature reduction technique to ease the complexity. The k-means clustering algorithm and SV coding were applied to generate the codebook for feature encoding. The use of their model that descriptor fusion is used by them to increase the recognition accuracy to 88.12%. However, their model has a limitation while using the Bag of Words (BoW) method with visual features since it discards the temporal order information.

Elshourbagy et al. [9] prescribed to improve the BOW to get better exactness in human movement acknowledgement i.e., a bag of features. These features used are STIP detector, HOG, and HOF. With multilevel clustering and the SVM classifier, visual BOW created a code book to classify human actions. This method reduced not only the time but also the memory requirements and it achieved 95.6% of detection accuracy. However, it misidentified some of the basic activities by confusing the actions such as running and jogging.

A new HAR system using background subtraction was projected in [10] to perform silhouette-based view-invariant activity identification. It was carried out by frame differencing method. Contour-based present highlights from outlines and uniform revolution invariant Local Binary Patterns (LBP) were acquainted in that framework with extricating the essentials. The actions were categorized via multiclass SVMs. One limitation of their model is that the update of the background model is not considered but it is very critical.

Wang et al. [11] implemented a model for HAR in which the path of action-based areas and motion information were captured by the optical-flow method. Their work improved the detection accuracy using shape features, HOG features, HOF features, and MBH features. It provided a very high computational cost but it achieved only 60% classification accuracy with the HMD51 dataset. It extracted spatial, and temporal action-related areas [12] and performed by implementing Interest Patch (IP) detector. The activity localization problem using HOG, Visual Background Extractor (ViBe) background subtraction method, and canny contour detector technique. Optical flow, pixel values, and gradients along with spatial SVM, and temporal IP distribution histograms were established in their model and final decisions were determined by the scores of these two SVMs. Here, the dictionary was trained separately and the accuracy obtained was 88.81%. The weakness of that work is the high memory requirements.

In video analytics, the deep-learned features are novel to solve many recognition, detection, and grouping problems. Yu et al. [13] proposed a different variation of CNN for HAR. Here, global-level extraction is carried out by frame-level feature extraction. PCA is applied for feature selection. It needs only a less labelled dataset for the training process and the advantage of this approach is the requirement of minimum training time. The hierarchical video illustration method discussed in [14] is contending with a DL classifier for performing the video analysis. The authors used CNN on video data and they also used LSTM for performing the action identification. One reward of this method is its ability to perform effective Feature Selection (FS) and the accuracy provided by this system is 90.8%. The primary disadvantage of their CNN approach is high calculation intricacy. Zhu et al. [15] suggested a structural activity model that is integrating the context activity features namely intra, inter-activity and finally the motion features namely Spatial-Temporal Interest Points (STIP), Optical-Flow using Histogram (HOF) and Oriented Gradients using Histogram (HOG). In their model, a greedy search algorithm was used to label all the activities found in the video. Acknowledgement was finished utilizing the Bag of Words (BOW) examination with a Support Vector Machine (SVM) for arrangement. Though their model is considering various features for enhancing the accuracy of recognition, it fails to address the inter-activity motion regions. Uninterrupted emotion recognition [16] is achieved using a 3D model. But it is very complex to build a three-dimensional model for each action.

Chen et al. [17] presented a multi-view activity identification system in which descriptors from distinct views were combined to create a new enhanced feature that contained the transition between the different views [18,19]. Here, actions were identified utilizing facial emotion. Gorelick et al. [20] used a bag of features scheme combined with late fusion. The lack of intrinsic visual relationship estimates limits the recognition challenge, even though their system performs well in recognizing human contact. For recognizing human interactions in TV shows, multi-view HAR system was developed by mapping action information from numerous camera perspectives into one, and a non-linear-based knowledge transfer approach using DL was presented. Dalal et al. [21] were able to distinguish between human activities by applying the histograms and SVM was deployed for classification i.e., action grouping. Using the density of all features [22] observed, Guha et al. [23] performed HAR by linking the context between interest locations. It motivated on three major topics i.e., the intention of a classifier, data modelling, and FS for activity recognition [24].

CNN has exposed countless achievements in recent years and attained high-tech act on numerous jobs (e.g., [25–28], etc.). In the literature, there have been numerous proposals for HAR using deep learning techniques [29]. Recently DL [30] has approaches that have had a significant impact on an extensive range of training fields, together with video and image comprehension, speech acknowledgement, and biomedical image analysis. It learns the optical-flow features [31,32] in different approaches in CNN and it achieves good results in recognition. Initially, CNN was used to perform image classification tasks. Later they were also used to perform video analytics [33,34]. The CNN was able to handle dislocations effectively. Recent works on CNN include relocations, deep-flow [35], and epic-flow [36] handling. DL-based algorithms such as flow net and its variants [37] were employed for analyzing optical flow [38,39]. However, the work on deep features flow was considered in performing the analysis. Lite flow net [40] is made up of two small sub-networks that specialize in extracting pyramidal features and estimating optical flow. When compared with related methods, this technique can fetch the characteristics faster. Flow net [41], flow net 2.0, lite flow net, capsule networks [42], and spy net have investigated optical flow estimation Deep Neural Networks (DNN). Likewise, with a small network, spy net utilizes spatial pyramids, to ensure that the spatial features are useful in data analysis. Nguyen et al. [43] used capsule structures as a classifier, and the results were noticeably better than CNN-based GANs. Moreover, capsule networks are being expanded into the analysis of videos. A technique for collecting from capsule networks was proposed by Zhu et al. As a result, it is not necessary to employ a multi-scale strategy. For estimating optical flow, some methods employ extra tools such as outward edge detectors or images to bring up-to-date associations. Xu et al. [44] used this capsule in gait recognition. For estimating the performance of HAR via two datasets namely UCF101 [45] and HMDB-51 [46]. The convolutional capsule layers and neural networks [47,48] were elaborate during learning. Layers of compartment pooling are used to aid action recognition [49–53]. These models are vastly different from our approach [54–56]. Few other works are also available in the literature that discusses the use of deep learning algorithms in different applications [57–60]. The major works from the literature survey are compared in Table 1.

Table 1: Comparative analysis of the literature survey

Authors	Paper	Contributions	Limitations
Seo et al. [5]	Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection	A new approach to reduce the complexity of optical flow measurement by implementing the dynamic frame bouncing method and motion interjection for efficient optical flow approximation. Accuracy is 95.3%	Some errors are found in their model due to the skipping of frames
Ullah et al. [6]	Object and motion cues-based collaborative approach for human activity localization and recognition in unconstrained videos	Extreme Learning Machine (ELM) classifier was used by these authors to recognize the activity more accurately. Here, the mid-level features were represented well and hence it has increased the recognition accuracy to 95.6%.	The object proposal method they developed was computationally expensive with high time complexity.

(Continued)

Table 1 (continued)

Authors	Paper	Contributions	Limitations
Zhao et al. [7]	Region-based mixture models for human action recognition in low-resolution videos	A layered elastic-motion tracking was suggested to capture both long-term trajectories of motion and mutual shape. Hybrid feature depiction was used for cording both the motion and the shape features. Action classification is carried out effectively by the region mixture model and in which there was no necessity for segmentation.	The shortcoming includes the encryption of the spatial outlines of the features
Peng et al. [8]	Bag of visual words and fusion methods for action recognition: comprehensive study and good practice	They developed BoW visual model for HAR as it provides effective acknowledgement based on actions acknowledgement. They carried out Principal Component Analysis (PCA) as a feature reduction technique to ease the complexity. The k-means clustering algorithm and SV coding were applied to generate the codebook for feature encoding. The use of their model that descriptor fusion is used by them to increase the recognition accuracy to 88.12%.	Their model has a limitation while using the Bag of Words (BoW) method with visual features since it discards the temporal order information.
Elshourbagy et al. [9]	Enhanced bag of words using multilevel k-means for HAR	Proposed a method that improved the BOW to get better exactness in human movement acknowledgement i.e., a bag of features. These features used are STIP detector, HOG, and HOF. With multilevel clustering and the SVM classifier, visual BOW created a code book to classify human actions. This method reduced not only the time but also the memory requirements and it achieved 95.6% of detection accuracy.	It misidentified some of the basic activities by confusing the actions such as running and jogging.

Based on the literature survey carried out in this work, it is observed that an increased number of redundant features, low accuracies, and higher computational complexities are the major issues to be addressed. In addition to this, the computationally expensive algorithms are not able to recognize the similarity between the actions efficiently. To overcome the limitations of the systems present in the existing literature, computationally efficient systems are needed for identifying the key-frames and accurate activity recognition. The difficulty that occurred in the identification of frames is resolved in this work by introducing an efficient key frame detection technique.

The following are the primary contributions of this work:

- To apply the improved form of the VGG-16 algorithm for the precise identification of human activity.
- To develop a wavelet-based VGG-16 with LSTM for the effective sequential learning of features for the detection of activity.
- To propose an AWFN algorithm for making an optical flow prediction capability during HAR.

The performance measures such as accuracy, F1-score, precision, and recall have all been used to illustrate the proposed system's effectiveness. It has also been demonstrated that the AWFN provides more detection accuracy than the current systems. The premise is that the suggested wavelet-based VGG-16 with LSTM and AWFN algorithm is possible to accurately detect human behaviors. The initial hypothesis is achieved in this case by adding feature correlation to the softmax classifier to increase its accuracy in differentiating similar activities. The system was trained and tested using 70% of the data (frames) for training and 30% of the data (frames) for testing.

In deep learning, a large number of learnable parameters are included in VGG-16 to differentiate the actions. It learns all the features including shape, texture, boundaries, inter and intra-class similarity. The hybridization of LSTM with VGG-16 enhances the performance of activity detection with reduced complexity. LSTM has a larger range of parameters such as learning rate and input-output bias. Here, the fine adjustments of parameters are not needed. To effectively handle the redundant features, high-level feature correlation is added after the convolution layer in the proposed AWFN.

3 Framework for Human Activity Recognition

In Activity Recognition, locating the moving objects present in an active background is an inspiring task. The low-level detection was performed in earlier approaches. However, our contribution to this paper is the proposal of new technologies for object-level detection using DWT. It is performed by extracting the important frames using threshold computation, orientation, magnitude, and optical flow from the key-frames. All the essential key-frames are processed here to extract the wavelet coefficient based on edges. Here, the DWT coefficients are used as features of human objects. The wavelet coefficients represent the multi-scale and the directional information of motion pattern where A_p gives an approximation of human and horizontal, vertical, and diagonal coefficients are denoted by D_h, D_v, and D_d in turn.

For all key-frames, the approximation of $\left(A_p\right)$ is given to VGG-16 and VGG-16 with LSTM to learn Deep Features (DF) for each activity. The remaining coefficients H, V, and D are given to the Flow Net and also to our proposed AWFN for building the model for each activity, and the results obtained are fused for analysis. Fig. 1 shows the architecture of the human action identification system. This method was implemented using two challenging activity datasets which focus on both indoor, and outdoor activities. The pseudo-code used for developing the proposed HAR system is stated in Algorithm 1.

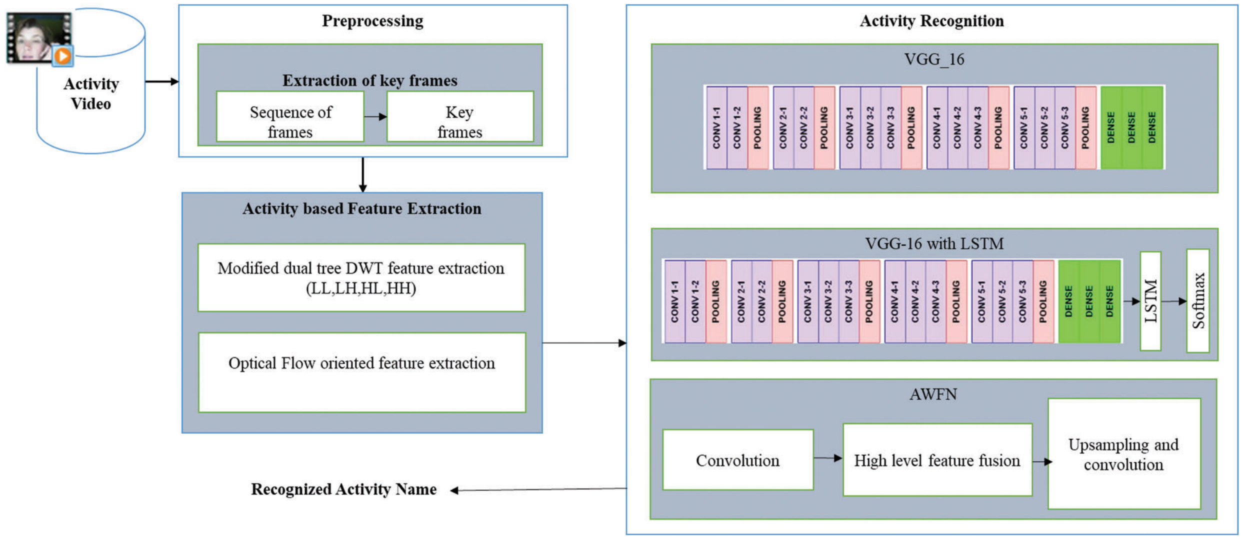


Figure 1: Architecture of human activity recognition system with the process of key-frames extraction, activity features extraction and activity recognition

3.1 Pre-Processing

During video pre-processing the extraction of frames is necessary for further processing. The larger set of frames extracted from a single video consumes more memory. To effectively handle the memory and simplify further processing, the extraction of key frames is required.

Extraction of the Key Frames (KF)

A keyframe (K_f) is a frame in a video sequence that is indicative of and capable of reflecting the abstract of the (V_m) video material [47]. It's critical to find individual shots in the video with many shots for K_f extraction. The KF is determined using visual and temporal information. To speed up processing, key-frames are extracted. Each video action results in a conversion. Each action video was turned into a frame sequence. It calculates the mean as well as the SD of Histogram Difference (HD) (H_d) between two successive frames. Now set a Threshold Value (TV) as 'T' and perform the comparison of all frames to the TV to yield key-frames. A frame that meets the threshold condition is called a keyframe. The threshold is obtained using Eq. (1).

$$T = H_d(mean) + H_d(standard) \quad (1)$$

The magnitude and orientation of all key-frames (K_f) are computed and stored in separate folders for finding a multi-view of the same activity. It is exposed in Fig. 2.

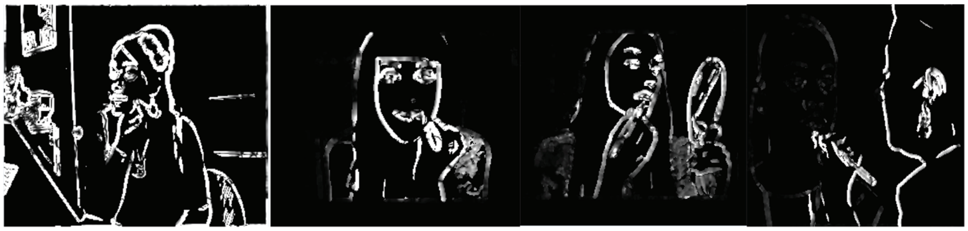


Figure 2: Apply makeup activity in different views

Algorithm 1: Pseudocode for HAR system

```
Input: Activity Videos
Output: Human Activities
// Preprocessing: Key Frame Identification
Extract frames 'F' from video $V_m$
Compute $H_d$, $H_d(\mu)$, $H_d(\sigma)$ and catch threshold (T) using Eq. (1)
for 1:Number_of_Frames
    Check If $H_d > T$ then
        Fixed the image as a keyframe
    Else
        Discard the frame
    End if
End for
// Magnitude and Orientation Feature Extraction
folder ← Compute Magnitude, Orientation of all key-frames ($K_f$)
// Modified Dual Tree DWT-based Feature Extraction
$K_{f_{(x,y)}}$ ← Read key-frames ($K_f$) as
For 1:end_of_$K_f$
    $R_1$ ← Apply G to $K_{f_{(x,y)}}$, $R_2$ ← Apply G to $K_{f_{(x,y)}}$
    $LL, HL$ ← Put on G to $R_1$, $R_2$ then downsample the result by 2 in 'y' direction
    $LL$, $HH$ ←Put on H to $R_1$, $R_2$ then downsample the result by 2 in the 'y'&'x' direction
Return features
End for
// Optical-flow-based Feature Extraction
for 1:Number_of_Frames
Calculate the displacement between consecutive key-frames ($K_f$) using Eq. (4)
End for
Return motion_features
// Learning using Vgg-16
Layer_1 to 2 ← Perform & Execute the Convolution operations based on 64 filters and the Max Pooling (MP) operations
Layer_3 to 4 ← Carryout Convolution operation using 128 filters + MP operations
Layer_5 to 7 ← Do Convolution by applying 256 filters with the MP operations
Layer 8 to 10 ← Execute the Convolutions using 512 filters with the MP operations
Layer_11 to 13 ← Do the Convolution by applying 512 filters and also Max pooling operations
Layer_14 to 15 ← Form a Fully Connected Network (FCN) with 4096 nodes
```

(Continued)

Algorithm 1: (continued)

```
28: Layer_16 ← Output layer using softmax function with nodes, which is identical to the number of Activities
29: Return the Activity
// Learning using Vgg-16 with LSTM
30: Layer_1 to 13 ← Execute Steps from 22 to 26
31: Layer_14 to 15 ← FCN formation using 4096 nodes
32: Layer_16 ←Flatten the previous layer features and Perform LSTM operations
33: Layer_16 ← Output layer with Softmax activation function trained with LSTM
34: Return the Sequential Recognized Activity
// Learning using AWFN
35: A ← Stack of two sequentially adjacent input images together ($K_f$)
36: a, b ← Yield arrangements of the two images separately
37: Resize input images and perform nine Convolution Layer (CL) with stride 2
38: Convolve 7 * 7 at layer_1 then Convolve 5 * 5 at layer_2
39: Convolve 3 * 3 at layer_3 to 6
40: Remove pooling layer outflow from dimension reduction
41: Apply motion_features obtained from 21 and resize to do association
42: Correlation instead of Kernel processing 41 * 41 * 1 at stride 2
43: Learn the higher representation together in the CLs of the network
44: Do four de-convolution and 4 upsampled predictions during refinement
45: Learn & Return the Activity
```

3.2 Activity-Based Feature Extraction

The accurate identification of features is responsible for the effective detection of activity. For that purpose two techniques such as modified dual-tree DWT and optical flow-oriented feature extraction methods are used in this work.

3.2.1 Modified Dual Tree DWT Feature Extraction

The spatial characteristics are extracted and the dimensionality is reduced using DWT. We employed the haar wavelet filter to decompose the dynamic picture in the suggested system and it is developed for feature detection. Decomposing a picture into four sub-images i.e., Lower-Lower (LL), Lower-Higher (LH), Higher-Lower (HL), and Higher-Higher (HH) using Low Pass Filter (LPF) given in Eq. (2) and High Pass Filter (HPF) mentioned in Eq. (3). It displays three details and an approximation of the image representation. Low frequency is included in the approximation components, whereas the details contain high-frequency components in the horizontal direction, vertical direction, and diagonal directions. This will decompose the image into many frequency bands that will enable the motion detector to employ only the useful frequencies and also ignore the irrelevant frequencies. We adopted the object detection method proposed by Cheng et al. [3] with DWT to identify the motion of objects. Consequently, the haar transform is expanded to compute the space of feature vectors of 180 attributes. Thus, it makes sense to consider the subset of scales and positions. Scales and positions are regarded as powers of two in the DWT. It's also called dyadic scales and positions.

$$\text{For Haar transform, } LPF\ (G)\ = \frac{1}{\sqrt{2}}\ [1, 1] \tag{2}$$

$$HPF\ (H) = \frac{1}{\sqrt{2}}\ [1, -1] \tag{3}$$

3.2.2 Optical Flow-Oriented Feature Extraction

An optical flow algorithm will calculate the displacement of the brightness patterns varying from one frame to another [33]. This algorithm uses the intensity values obtained from the neighbouring pixels for performing the analysis. Moreover, the proposed model performs displacement estimation on a few selected pixels. Now, it applies the spatial-temporal variations in brightness to extract the spatial effects. From spatial-temporal picture brightness fluctuations, this system can recover the motions of the image. It is used for extracting some special effects and estimating the dimensional and structural features of the scene. The equation of the optical flow when all the pixels are in a position is shown in Eq. (4). The velocity vector V_x and V_y of the local image flow will satisfy Eq. (4). I_x, I_y denotes the x and y directions.

$$I_x(q_1) \times V_x\ + I_y(q_1) \times\ V_y = -I_t(q_1) \tag{4}$$

The Optical-Flow (OF) technique determines the illumination gradient changes from one frame to the next as shown in Fig. 4. This is achieved by combining the intensity levels of adjacent pixels. Dense optical flow is achieved by determining the dislocation for a specific number of pixels that are present in the image, whereas dense optical-flow techniques estimate the movement for all input images. Here $q_1, q_2, q_3, q_4\ldots\ldots\ldots, q_n$ denote the pixels that are present inside the window (e.g., n = 10 for a 4 × 4 window), and $I_{x(q_i)}$, $I_{y(q_i)}$ and $I_{t(q_i)}$ denote the partial differentiation values of image 'I' based on variables (x, y) and the time 't', for pixel q_i at the current time.

From Fig. 3, it is known that among succeeding frames, the image intensity (I) is signified as a function of space (x, y) and time (t). For the leading image (Fig. 3), I (x, y, t) is the intensity of pixels obtained using $\left(d_x, d_y\right)$ over time as 't'. The obtained new image is $I\ (x + d_x, y + d_y, t + d_t)$. To begin, the pixel intensities of particular items are considered to be constant across two successive frames. Two successive frames will be divided by a small increment (d_t) on time such that the objects are not relocated considerably. A frame represents a "natural" scene having textured objects displays shades of grey which change smoothly.

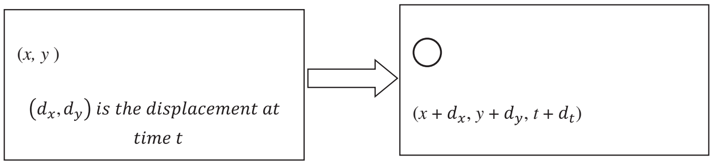
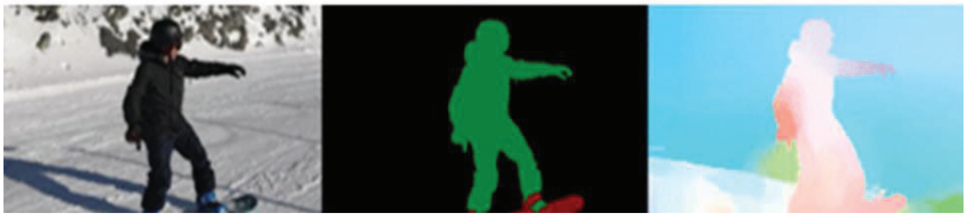

$(x, y\ )$
$\left(d_x, d_y\right)$ is the displacement at time t
$(x + d_x, y + d_y, t + d_t)$

Figure 3: Pixel movement in frames

3.3 Activity Recognition

Activity recognition systems accurately identify the activity based on the features. For the effective processing of features, deep learning techniques such as VGG-16, VGG-16 with LSTM, and AWFN are used.



(a) Sample input image (b) Displacement movement in the frame (c) Optical flow

Figure 4: The process of optical flow estimation

3.3.1 VGG-16

Visual Geometry Group, or VGG, is a typical deep Convolutional Neural Network (CNN) architecture containing several layers. With VGG-16 comprising 16 convolutional layers, respectively, the "deep" states to the total number of layers. By specifying filter, stride, and padding, the features are extracted. In ReLU Layer (Rectified Linear Unit) a non-linear activation function is how convolutional layers are typically viewed. For each part, it creates a threshold procedure. Any less-than-zero input value is set to zero and does not affect the input size. The maximum pooling layer and average pooling layer are the two forms of the pooling layer. Overfitting is managed by the pooling layer. The number of parameters is decreased. Both the amount of processing time and the representation's spatial size are decreased. Eq. (5) is used to determine the value using filter size 2 for the pooling layer.

A layer or layers that are fully connected come after the convolutional layers. A layer that is completely connected to the layer before it has connections between all of its neurons. To identify the layer patterns, the fully connected layer incorporates all of the features that the preceding layer has learned about the image. The last fully linked layer integrates them to classify the image, with classification serving as its primary function.

$$\sigma\left(\overrightarrow{M}\right)_h = \frac{e^{M_h}}{\sum_k^c e^{M_k}} \tag{5}$$

The neurons are transferred from input to output channels via the network's softmax layer, which provided the identified activity based on input frame attributes. The computation's softmax function was provided in Eq. (5). The input vector used was $\{M_0, \ldots, M_c\}$, where c stands for the overall activity count. By applying the exponentiation function to the feature vector created at the final stage of convolution block-5, the numerator in Eq. (6) calculates the probability value. The same equation's denominator outputs the normalized value. This value, which is always between 0 and 1, is written as i.e., $0 \leq normalized\_value \leq 1$.

3.3.2 VGG-16 with LSTM

In the network DL, the DF has been extracted by way of specifying the filter size, stride, structure, and contents of padding. In this proposed model, the network was constructed on the VGG-16 model on the LSTM layer. The FCC's VGG-16 model is consisting of 16 layers. It has 13 convolutional layers, 3 fully connected layers, and one softmax layer. It has layers such as ReLU, normalization, max pooling, and dropout layers. Network-based learning eliminates the need for a hand-crafted feature extraction method. The DFs are dug out directly from the order of images. For that LSTM layer was added. It convolves the learned features' input data.

Image input size will define the dimensions of the input image of a VGG-16 and it contains the raw pixel value of the wavelet approximation image. The dimension of an image corresponds to its height, width, and the number of colour channels present in the image. The size of an image is 320 × 240. During training, the input of the network is fixed as 64 × 64 × 1. It is extracted from the DWT approximation wavelet sequence. Detailed information is present in this image for distinguishing the activity. It reduces the processing time and digs out more deep data.

The convolution layer performs the Key Frame (KF) extraction, feature extraction, and feature selection. The convolution layer consists of many hidden layers for performing the feature analysis. The CL output is calculated using Eq. (6) while the padding size is zero otherwise Eq. (7) is used.

$$Convolution\ layer\ output\ \left(cl_{p=0}\right) = \left(\frac{Input\ size - Filter}{Stride}\right) + 1 \tag{6}$$

$$Convolution\ layer\ output\ \left(cl_{p\neq 0}\right) = \frac{Filter - Stride}{2} \tag{7}$$

The convolution and max pooling layers of CNN are working continuously and then provide the necessary features after performing the feature extraction. The convolution layer is more important due to its ability to work with FCL.

3.3.3 Adaptive Weighted Flow Net (AWFN)

In this proposed work we train a network through the KF's (K_f) obtained using Eq. (1). To find the activities in the frames the first approach is to filter the object flow from one frame to another. For that, we proposed AWFN flow in Fig. 1. In that, we convolve the extracted motion data to the image (Approximation image and optical flow of action) for obtaining the object flow i.e., action trajectories. Pooling in this CNN network reduces the dimensionality of the feature for that to make it a compressed feature we use narrowing and intensifying measures using a backpropagation algorithm.

A solution to the problem is used for extracting the motion information using discrete wavelet transform and the features (f_1) are convolved with $(w^2.h^2)$ the simple activity flow (f_2) using CNN. To find the gesture between the activity Flow Features (FF) and the Motion Feature (MF) we announce a motion flow activity correlation feature map on the way to identify the flow of action called AWFN.

Motion feature as C and the activity flow CNN feature as F_{f2} then the gesture between the two feature maps is found using Eq. (8). It is a single progression in the simple CNN convolution layer. The features in M_{f1} are aligned at A_1 and F_{f2} is aligned at A_2. Here the dimension of the dense feature is $j := 2j + 1$.

$$G(A_1, A_2) = \sum\nolimits_{0\in[-j,j]\times[-j,j]} \langle M_{f1}(A_1 + 0),\ M_{f2}(A_2 + 0)\rangle \tag{8}$$

AWFN is the path to build with a set of incremental but significant changes that aren't related to the problems tangentially that were identified. We begin by assessing the impact of dataset scheduling. When more advanced training data is provided by lower results. A learning schedule includes numerous datasets and hence considerably enhances performance. We also informed the flow net model using one explicit correlation layer outperformed the model without such a layer in this context. Secondly, we describe a warping technique and demonstrate how stacking several networks employing this operation will improve outcomes dramatically. We may create a variety of network versions with varied sizes and runtimes by altering the depth of the stack and the measures of the individual components. This helps to manage the accuracy-to-computational-resources trade-off. It provides networks with frame rates ranging from 8 to 144 frames per second. This method has complete F1-scores of 87.64%, 90.30%, 93.64%, 95.37%, 92.44%, and 89.48% for the UCF-101 actions datasets respectively, showing the efficiency of the proposed technique compared to state-of-the-art techniques. A multi-level network is

- D() is the downsampling function that decrease an m * n image into (K_f) to m/2 * n/2.

- u() is the resampling function used to perform the supportive services that resample optical-flow field i.e., motion.
- $(A_1,\ and\ A_2)$ were used for warping image (K_f), affording to the optical flow field.
- $\{G_0,\ldots, G_k\}$ is the set of trained CNNs.
- V_k is residual flow and it is computed by convolutional net G_k at the K^{th} pyramid level.

4 Experimental Results

4.1 The Dataset Description

For estimating the performance of HAR via two datasets namely UCF101 [45] and HMDB-51 [46] were selected. The HMDB51 dataset consists of accurate videos from numerous sources, such as movies and websites. It contains 6,766 video clips taken from 51 action groups, with each category covering a minimum of 100 clips. Around 2 GB for a total of 7,000 clips scattered in 51 action classes. Our tryouts followed the original valuation scheme, but it is only adopting the first training/testing split. Moreover, each of the action classes of this split contains 70 clips, and they are used in training and 30 clips are allotted for testing as shown in Table 2a. The UCF101 dataset contains 13,320 video clips in 101 action classes and 100 video clips for every class. We tested this proposed system based on training and testing of the model. The input dataset is separated into a training set enclosing 9.5K videos and a testing set holding 3.8K videos for training and testing more effectively.

Table 2a: The training and testing split for two different datasets used in the experiments

Dataset	Training	Testing
UCF-101	9537 (Split#1)	3783 (Split#1)
HMDB-51	1578 (Split#1)	676 (Split#2)

4.2 Results and Discussion

The actions from the Human Motion Database (HMDB-51) dataset and the University of Central Florida (UCF-101) were tested and the results are evaluated. Due to the poor quality of videos on HMDB-51, the accuracy for every action class on the HMDB-51 dataset was less when compared with UCF-101 dataset. The UCF-101 dataset generated excellent results for every action type. At hand, a total of 153 files in this folder, three action tests are split and reported in the dataset. The format of each file is [video-name] [id]. In this model, videos were used for both training and testing. Moreover, we related the performance of this proposed AWFN model with the existing VGG-16+LSTM model based on comparative analysis.

Moreover, the feature maps of the most recent FCL of FC-4096 of VGG-16 and the last MPL of FCNs-16 are extracted to represent the frame-level trajectory feature, respectively. The motion features are applied instead of pool frame-level features to get the final video feature. We combined two-stream features by fusing concatenation. The results of the AWFN model on UCF101 are mentioned in Fig. 5.

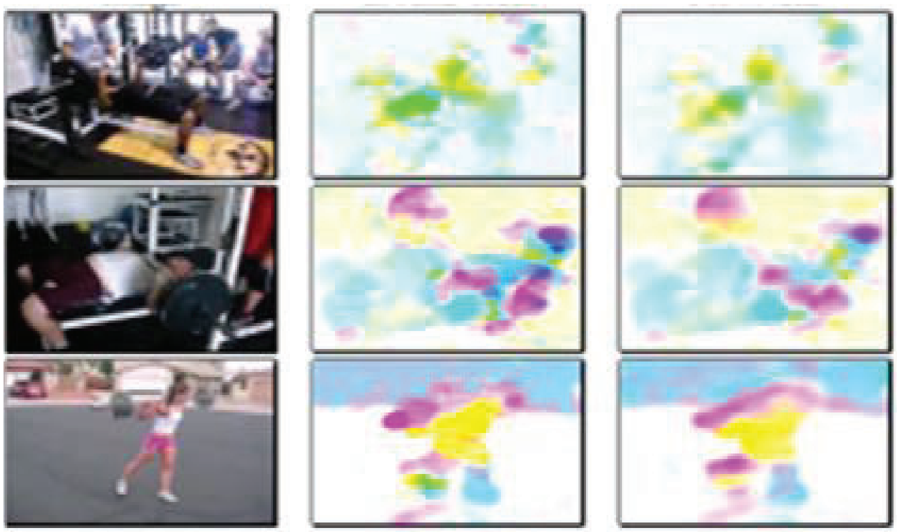

(a) Input image (b) Optical flow estimation (c) AWFN estimation

Figure 5: AWFN estimation of the UCF-101 dataset

The training and testing are split for the two distinct datasets such as UCF-101 and HMDB-51 are specified in Table 2a. Table 2b shows the motion segmentation and recognition comparison of extracted objects. F-measure, accuracy, precision, and recall is described in Table 3b, and for a few activity class. The comparison with other modules is examined in Table 4.

Table 2b: The number of object extraction along with the performance of AR

Model	Motion trajectory	Action recognition	
	Extracted objects	F-measure	Accuracy
Flow nets [37]	30/62	56.87%	65.27%
Optical flow [38]	30/65	89.41%	78.91%
Deep flow [39]	27/65	80.18%	80.79%
Lite flow [40]	31/65	79.70%	78.48%
AWFN (Proposed)	38/65	82.92%	89.51%

The extraction of objects from the KFs and their performance during activity detection for the literature and proposed methods are tabulated in Table 2b. In this, the proposed AWFN extracted 38 objects out of 65 objects in frames leading to offering higher action recognition power of 89.51%. This is approximately 8.72% to 24.24% higher recognition power as compared to the techniques used in related work for object extraction during activity recognition.

The precision, recall, f1-score, and accuracy of the proposed technique during distinct activity recognition are specified in Table 3b. In this, the archery action offered higher precision of 95.05%, and applying eye makeup offered a lower precision value of 84.66% as compared with other actions. This precision value offered how likely this proposed method detects the correct action. In the same manner, the recall value is also computed to identify the number of correctly identified actions divided by the total number of action inputs mentioned in sensitivity. Here, the baby crawling action has a higher recall value of 95.89% and applying eye makeup secured a lower recall value of 90.85%. The confusion matrix for the results is shown in Table 3a.

Table 3a: The confusion matrix during activity recognition

Ground_truth//Predicted	Positive	Negative
Positive	911	39
Negative	18	32

Table 3b: Precision, recall, f1-score, and accuracy of proposed technique during distinct activity recognition

Action recognition	Precision	Recall	F1-score	Accuracy
Apply eye makeup	84.66	90.85	87.64	88%
Apply lipstick	88	92.73	90.30	86.57%
Archery	95.05	92.28	93.64	89.04%
Baby crawling	94.87	95.89	95.37	91.71%
Balance beam	90.39	94.59	92.44	92.5%
Band marching	87.58	91.47	89.48	92.59%

Table 4: Comparison of proposed activity recognition technique with state-of-art methods

Survey paper	Method	Accuracy
Seo et al. [5]	TS+HOG+HOF+MBH+PCA+FV+SRC	95.3%
Ullah et al. [6]	BoW+STIP+HOG+HOF+code book+SVM	95.6%
Peng et al. [8]	BoVW+STIP+HOG+HOF+MBH+SVM	88.12%
Elshourbagy et al. [9]	pHog and Gist+HooF+ELM	95.67%
Xu et al. [12]	HOG+ViBe+canny contour detector+ IP descriptors(pixel values, gradients, and optical flow)+SVM	88.81%
Yu et al. [13]	SP-CNN+PCA+SVM	88%
Li et al. [14]	2D CNN+3D CNN+LSTM+MSD	90.8%
Fischer et al. [37]	Flow Net+Flow Net corr with CNN-optical flow	95.7%
Li et al. [49]	LSTM	90.3%
Vrskova et al. [50]	3D Convolutional Neural Network	85.2%
Aldahoul et al. [51]	EfficientDetD7	92.9%
Yildirim et al. [52]	MA-Net	91.34%
Zebhi et al. [53]	VGG-16	92.4%
Proposed method	AWFN	96%

In Table 3b, for different actions, f1-score is computed by identifying the geometric mean value between positive predictions and recall. Based on the f1-score value, the activity of baby crawling offered higher recognition power of 95.37% whereas the action applied eye makeup received a lower recognition rate of 87.64%. Likewise, the accuracy of AR is higher for band marching and lower for applying lipstick. Anyway, the mean accuracy of the proposed technique is improved significantly. The accuracy of the proposed wavelet-based VGG-16 with LSTM and using AWFN detected key frames offered 96% of accuracy. This is significantly higher accuracy as compared with previous literature. The comparison of already existing AR-based literature with the proposed technique is given in Table 4.

From the work carried out on classification seen that on temporal stream, AWFN outperforms the VGG-16+LSTM method by 1.0% on the UCF101 (88.8% *vs*. 89.8%). On the spatial stream, AWFN outpaces wavelet-based VGG-16 with the LSTM network by 1.5%. The results reveal that wavelet-based VGG-16 with LSTM+AWFN architecture can achieve higher recognition accuracy with fewer parameters (1.01 to Million). Fig. 6 presents the accuracy analysis performed between the CNN and the proposed CNN with spatio-temporal features.

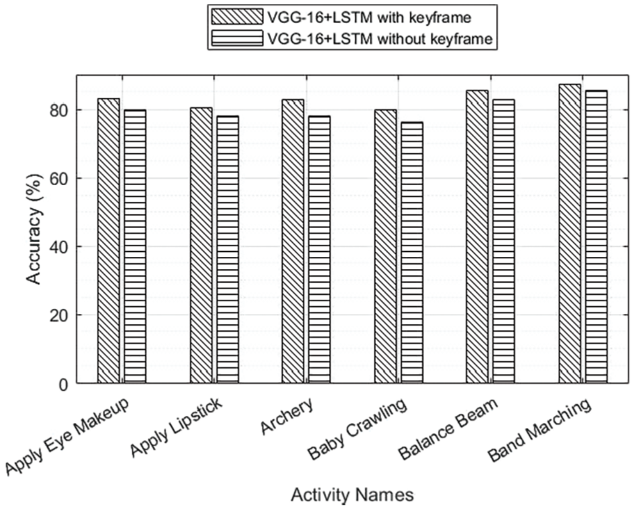


Figure 6: LSTM added VGG-16 activity with and without keyframe comparison

From Fig. 6, it is seen that software testing with VGG+LSTM was better by considering the standard datasets available. Fig. 7 depicts the accuracy analysis of wavelet-based VGG-16+LSTM with the related algorithms on single activity classes.

From this analysis, it is seen that the proposed CNN-based DL algorithm provides better classification accuracy than the related work. Fig. 8, depicts the accuracy comparison performance of AWFN with and without a keyframe. In this, the AWFN with keyframe provided better accuracy for all activity classes. The analysis of various activity detection using four distinct algorithms is depicted in Fig. 9. From this, this proposed wavelet-based VGG-16 with LSTM for AWFN key frames offered peak accuracies for all of the activity classes.

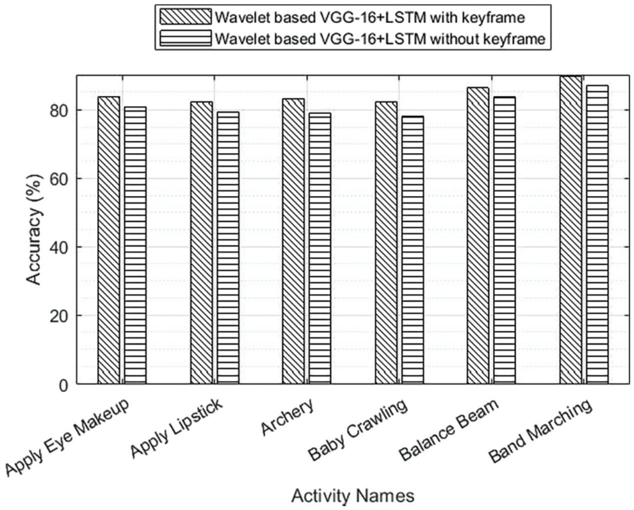


Figure 7: Wavelet-based VGG-16+LSTM activity comparison using keyframe

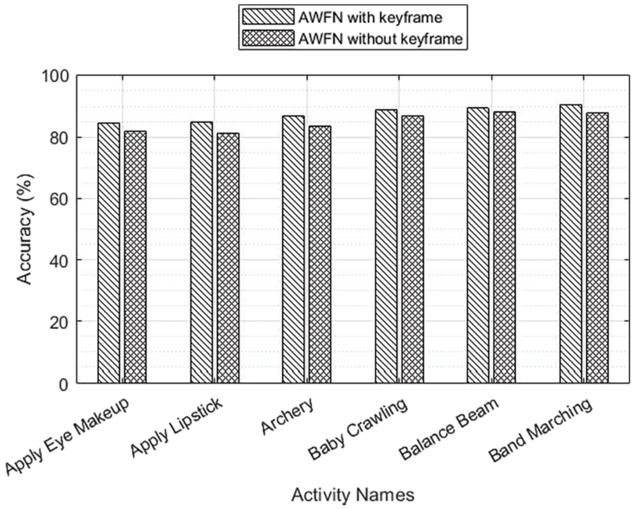


Figure 8: AWFN activity comparison using keyframe

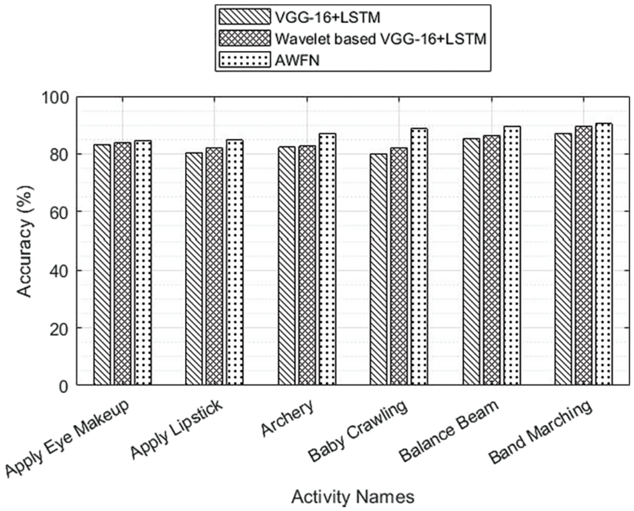


Figure 9: Wavelet-based LSTM, AWFN added activity comparison using keyframe

In the proposed system, the performance has been analyzed by testing the test divided into images. Moreover, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values are evaluated by analyzing the actual and predicted class label which in terms contributes to the evaluation of accuracy in Eq. (9). MATLAB was used here to extract KF's from the video. With python, we found the motion trajectories using the haar wavelet. Approximation of object-based frame digs out from the KF by applying a low-frequency filter. Flow-based trajectory path expanded using VGG-16, VGG-16+LSTM and AWFN are depicted in Fig. 1.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \tag{9}$$

$$\text{F}_{\beta} = \left(1 + \beta^2\right) . \frac{\text{precision} * \text{recall}}{\beta \cdot \text{precision} + \text{recall}} \tag{10}$$

On certain test splits, this approach was used. By looking at our technique, we can see that the standard f-score is the only metric that is meant to use accuracy and recall in the calculation Eq. (9). We calculated pixel-wise accuracy and recall values to estimate the area overlap. TP's and FP's concepts are used to determine precision and recall (PR) levels. The f-score range from 0 to 1, where the f-score number for an exact approach should be high. It is a subset of the universal f_measure based on Eq. (10). The suggested CNN-based model outperforms the previous works on several occasions, as shown in these figures.

The comparison with state of art methods analyzed from 2016 to 2022 is tabulated in Table 4. In this, the methods used for analyzing human activity along with their accuracies are mentioned. Machine learning (ML) algorithms such as SVM, ELM, and SRC are widely used for activity detection. These algorithms are a lower ability to handle the non-linear features extracted from activity-based images. Therefore, all the ML algorithms are less stable for activity recognition. Now a day, deep learning-based algorithms such as 2D CNN, 3D CNN, LSTM, MSD, EfficientDetD7, MA-Net, and VGG-16 are used for activity recognition. The usage of a larger number of layers in deep learning algorithms performed well during activity recognition. As compared with the state-of-art methods the proposed AWFN methods offered a higher activity recognition rate of 96% due to its accurate activity features pattern recognition ability.

5 Conclusion

The proposed HAR technique integrated the complex activities which usually require tasks namely object detection, and modelling of individual frame features. Initially, key frames are identified that offered higher recognition power as compared without the use of keyframe selection during HAR recognition. To compute the features, we proposed a wavelet-based optical flow technique that minimizes the action recognition error. Finally, learning algorithms such as the softmax layer of VGG-16, LSTM, flow-net, and AWFN are used in which AWFN offered higher prediction power during HAR. Based on performance results AWFN offered 96% accuracy as compared with existing literature. It is approximately 0.3% to 7.88% higher accuracy than existing methods. However, the time complexity of the proposed algorithm is the same as that of the base algorithm present in the literature namely VGG-16 algorithm. In future work, multiple-view information will be gathered for the detection of a group of activities. In the future, our goal is to test several pre-trained CNNs with rules for better functioning on the UCF-101 and HMDB-51 datasets. To offer a better classifier to improve activity recognition, multi-activity can be seen in many orientations.

Acknowledgement: We would also like to show our gratitude to professor Dr A. Kannan for sharing their pearls of wisdom with us during this research.

Funding Statement: The authors received no specific funding for this study.

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

References

[1] M. N. Al-Berry, A. M. Mohammed, H. M. Ebied, A. S. Hussein, M. F. Tolba *et al.,* "Weighted directional 3D stationary wavelet-based action classification," *Egyptian Computer Science Journal*, vol. 39, no. 2, pp. 83–97, 2015.

[2] L. Shao and R. Gao, "A wavelet-based local descriptor for human action recognition," in *BMVC*, Aberystwyth, UK, pp. 1–10, 2010.

[3] F. H. Cheng and Y. L. Chen, "Real time multiple objects tracking and identification based on discrete wavelet transform," *Pattern Recognition*, vol. 39, no. 6, pp. 1126–1139, 2006.

[4] A. B. Sargano, P. Angelov and Z. Habib, "Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines," *Applied Sciences*, vol. 6, no. 10, pp. 309, 2006.

[5] J. J. Seo, H. I. Kim, W. De Neve and Y. M. Ro, "Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection," *Image and Vision Computing*, vol. 58, pp. 76–85, 2017.

[6] J. Ullah and M. A. Jaffar, "Object and motion cues based collaborative approach for human activity localization and recognition in unconstrained videos," *Cluster Computing*, vol. 21, no. 1, pp. 311–322, 2018.

[7] Y. Zhao, H. Di, J. Zhang, Y. Lu, F. Lv *et al.,* "Region-based mixture models for human action recognition in low-resolution videos," *Neurocomputing*, vol. 247, no. 1, pp. 1–15, 2017.

[8] X. Peng, L. Wang, X. Wang and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, no. 3, pp. 109–125, 2016.

[9] M. Elshourbagy, E. Hemayed and M. Fayek, "Enhanced bag of words using multilevel k-means for human activity recognition," *Egyptian Informatics Journal*, vol. 17, no. 2, pp. 227–237, 2016.

[10] A. Kushwaha, S. Srivastava and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," *Multimedia Systems*, vol. 23, no. 4, pp. 451–467, 2017.

[11] X. Wang, C. Qi and F. Lin, "Combined trajectories for action recognition based on saliency detection and motion boundary," *Signal Processing: Image Communication*, vol. 57, pp. 91–102, 2017.

[12] K. Xu, X. Jiang and T. Sun, "Two-stream dictionary learning architecture for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 567–576, 2017.

[13] S. Yu, Y. Cheng, S. Su, G. Cai, S. Li *et al.,* "Stratified pooling based deep convolutional neural networks for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 11, pp. 13367–13382, 2017.

[14] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui *et al.,* "Learning hierarchical video representation for action recognition," *International Journal of Multimedia Information Retrieval*, vol. 6, no. 1, pp. 85–98, 2017.

[15] Y. Zhu, N. M. Nayak and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 91–101, 2012.

[16] C. Y. Chen and K. Grauman, "Efficient activity detection with max-subgraph search," in *2012 IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, pp. 1274–1281, 2012.

[17] H. Chen, J. Li, F. Zhang, Y. Li, H. Wang *et al.,* "3D model-based continuous emotion recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, pp. 1836–1845, 2015.

[18] L. Chen, L. Duan and D. Xu, "Event recognition in videos by learning from heterogeneous web sources," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, OR, pp. 2666–2673, 2013.

[19] D. M. Gavrila, “The visual analysis of human movement: A survey,” *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[20] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri *et al.,* “Actions as space-time shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[21] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR’05)*, San Diego, CA, USA, vol. 1, pp. 886–893, 2005.

[22] C. Schuldt, I. Laptev and B. Caputo, “Recognizing human actions: A local SVM approach,” in *Proc. of the 17th Int. Conf. on Pattern Recognition ICPR*, Cambridge, UK, vol. 3, pp. 32–36, 2004.

[23] T. Guha and R. K. Ward, “Learning sparse representations for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2011.

[24] G. Guo and A. Lai, “A survey on still image based human action recognition,” *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[25] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy *et al.,* “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2462–2470, 2017.

[26] X. Zhu, Y. Xiong, J. Dai, L. Yuan, Y. Wei *et al.,* “Deep feature flow for video recognition,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2349–2358, 2017.

[27] S. Zweig and L. Wolf, “Interponet, a brain inspired neural network for optical flow dense interpolation,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 4563–4572, 2017.

[28] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in Neural Information Processing Systems*, Montreal Canada, vol. 1, pp. 568–576, 2014.

[29] K. He, X. Zhang, S. Ren and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[30] D. Erhan, C. Szegedy, A. Toshev and D. Anguelov, “Scalable object detection using deep neural networks,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 2147–2154, 2014.

[31] T. Brox, A. Bruhn, N. Papenberg and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *European Conf. on Computer Vision*, Berlin, Heidelberg, Springer, pp. 25–36, 2004.

[32] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, 1981.

[33] J. Weickert, A. Bruhn, T. Brox and N. Papenberg, “A survey on variational optic flow methods for small displacements,” in *Mathematical Models for Registration and Applications to Medical Imaging*, Germany: Springer, Berlin, Heidelberg, vol. 10, pp. 103–136, 2006. https://dx.doi.org/10.1007/978-3-540-34767-5_5.

[34] T. Brox and J. Malik, “Large displacement optical flow: Descriptor matching in variational motion estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, 2010.

[35] P. Weinzaepfel, J. Revaud, Z. Harchaoui and C. Schmid, “Deepflow: Large displacement optical flow with deep matching,” in *Proc. of the IEEE Int. Conf. on Computer Vision*, Sydney, NSW, Australia, pp. 1385–1392, 2013.

[36] J. Revaud, P. Weinzaepfel, Z. Harchaoui and C. Schmid, “Epicflow: Edge-preserving interpolation of correspondences for optical flow,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 1164–1172, 2015.

[37] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazirbas *et al.,* “Flownet: Learning optical flow with convolutional networks,” arXiv preprint arXiv: 1504.06852, 2015.

[38] A. Ranjan and M. J. Black, “Optical flow estimation using a spatial pyramid network,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 4161–4170, 2017.

[39] X. Zhu, Y. Xiong, J. Dai, L. Yuan, Y. Wei *et al.,* “Deep feature flow for video recognition,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2349–2358, 2017.

[40] T. W. Hui, X. Tang and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8981–8989, 2018.

[41] Z. Sun and H. Wang, "Deeper spatial pyramid network with refined up-sampling for optical flow estimation," in *Pacific Rim Conf. on Multimedia*, Cham, Springer, pp. 492–501, 2018.

[42] Z. Zhu, G. Peng, Y. Chen and H. Gao, "A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis," *Neurocomputing*, vol. 323, no. 4, pp. 62–75, 2019.

[43] H. H. Nguyen, J. Yamagishi and I. Echizen, "Use of a capsule network to detect fake images and videos," arXiv preprint arXiv: 1910.12467, 2019.

[44] Z. Xu, W. Lu, Q. Zhang, Y. Yeung, X. Chen *et al.,* "Gait recognition based on capsule network," *Journal of Visual Communication and Image Representation*, vol. 59, no. 2, pp. 159–167, 2019.

[45] 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, 2012.

[46] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre *et al.,* "HMDB: A large video database for human motion recognition," in *2011 Int. Conf. on Computer Vision*, IEEE, Barcelona, Spain, pp. 2556–2563, 2011. https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#dataset.

[47] G. Augusta Kani, P. Geetha and A. Gomathi, "Human activity recognition using deep with gradient fused handcrafted features and categorization based on machine learning technique," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 7, pp. 1–7, 2018.

[48] G. Augusta Kani and P. Geetha, "An unsupervised approach for moving object detection using distinction features," in *Integrated Intelligent Computing, Communication and Security*, Singapore: Springer, pp. 631–641, 2019.

[49] X. Li, Y. He, Y. Yang, Y. Hong, X. Jing *et al.,* "LSTM based human activity classification on radar range profile," in *IEEE Int. Conf. on Computational Electromagnetics (ICCEM)*, Shanghai, China, pp. 1–2, 2019.

[50] R. Vrskova, R. Hudec, P. Kamencay and P. Sykora, "Human activity classification using the 3DCNN architecture," *Applied Sciences*, vol. 12, no. 2, pp. 931, 2022.

[51] N. Aldahoul, H. A. Karim, A. Q. M. Sabri, M. J. T. Tan, M. A. Momo *et al.,* "A comparison between various human detectors and CNN-based feature extractors for human activity recognition via aerial captured video sequences," *IEEE Access*, vol. 10, pp. 63532–63553, 2022.

[52] M. Yildirim and A. Çinar, "A new model for classification of human movements on videos using convolutional neural networks MA-Net," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 9, pp. 651–659, 2021.

[53] S. Zebhi, S. M. T. Almodarresi and V. Abootalebi, "Human activity recognition by using MHIs of frame sequences," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 3, no. 28, pp. 1716–1730, 2020.

[54] G. Augusta Kani and P. Geetha, "Comparative analysis on human action recognition using spatio temporal feature," *International Journal for Research in Engineering Application and Management (IJREAM)*, vol. 5, no. 3, pp. 608–613, 2019.

[55] N. Dilshad, J. Hwang, J. Song and N. Sung, "Applications and challenges in video surveillance via drone: A brief survey," in *Proc. ICTC*, Jeju Island, SK, pp. 728–732, 2020.

[56] N. Dilshad and J. Song, "Dual-stream siamese network for vehicle re-identification via dilated convolutional layers," in *Proc. SmartIoT*, Jeju Island, SK, pp. 350–352, 2021.

[57] K. Wang, C. M. Chen, M. S. Hossain, G. Muhammade, S. Kumar *et al.,* "Computer networks, transfer reinforcement learning-based road object detection in next-generation IoT domain," *Computer Networks*, vol. 193, no. 3, pp. 108078, 2021.

[58] S. Rajasoundaran, S. V. N. S. Kumar, M. Selvi, S. Ganapathy and A. Kannan, "Multi-tier block truncation coding model using genetic autoencoderss for gray scale images," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 42621–42647, 2022. https://doi.org/10.1007/s11042-022-13475-x.

[59] N. Dilshad, A. Ullah, J. Kim and J. Seo, "LocateUAV: Unmanned aerial vehicle location estimation via contextual analysis in an IoT environment," *Internet of Things Journal*, pp. 1, 2022. https://dx.doi.org/10.1109/JIOT.2022.3162300.

[60] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, V. H. C. De Albuquerque *et al.,* "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," *IEEE Transactions on Industrial Electronics*, vol. 12, no. 66, pp. 9692–9702, 2018.