INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Multi-dimensional CNN Based Feature Extraction with Feature Fusion and SVM for Human Activity Recognition in Surveillance Videos

Hetal Shah[1]*, **Mehfuza S Holia**[2]

1 Department of Computer Engineering, Gujarat Technological University, Ahmedabad, 382424, Gujarat, India
2 Assistant Professor, Department of Electronics, BVM Engineering College, Vallabh Vidyanagar, Gujarat, India

OPEN ACCESS

Received: 21-12-2023
Accepted: 24-04-2024
Published: 21-05-2024

Citation: Shah H, Holia MS (2024) Multi-dimensional CNN Based Feature Extraction with Feature Fusion and SVM for Human Activity Recognition in Surveillance Videos. Indian Journal of Science and Technology 17(21): 2177-2198. https://doi.org/10.17485/IJST/v17i21.3203

* **Corresponding author**.

hetal189@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background/Objectives: The accurate recognition of human activities from video sequences is very challenging due to low resolution, cluttered background, partial occlusion, and different viewpoints. Machine learning (ML) based automated HAR from surveillance videos is required with the fusion of various feature extraction techniques. **Methods:** In this paper, SVM with feature fusion is utilized for automatic recognition from surveillance videos. A Histogram of Oriented Gradient (HOG) is used to segment the frame to differentiate humans from other objects or background noise in the input video frames. The multi-feature extraction can be accomplished in terms of Gabor Wavelet Transform (GWT), Autocorrelogram, Gray-Level Co-Occurrence Matrix (GLCM), HSV histogram, and Multi-dimensional CNN. The proposed approach is implemented in MATLAB software and compared with existing approaches like Space-Time Interest Point (STIP) and Histogram of Optical Flow (HOF). **Findings:** The proposed approach outperforms the existing approaches in terms of reduced time consumption and high accuracy, 99.886% when using the UCF101 dataset and 99.538% when using the UTKinect dataset. **Novelty:** The most discriminative feature information is obtained with the feature-level fusion technique. From the feature information, various human actions are recognized with the classification algorithm.

Keywords: Human activity recognition; Machine Learning; Surveillance Videos; Human detection algorithm; Feature extraction; SVM classifier

1 Introduction

In computer vision technology, Human Activity Recognition (HAR) is essential for tracking movements captured in surveillance videos[(1)]. The video surveillance system contains infinite video cameras, monitors, recorders, and display units on a network to deliver the captured activities of humans as information to the central location[(2,3)]. The concept of HAR from video is used to extract the activities from each video frame and

analyze the features of HAR. In video surveillance, human activity detection is used to capture moving objects in images to prevent theft and to detect security attacks and fraud to manage unwanted incidents and crowd movements [4,5]. In order to record the actions in response to movement, human activity detection in video surveillance is intended to function around the clock.

The detection of human activities in surveillance video is complicated. Various levels are used to categorize human activities based on crowd movement, individual action, and group activity [6,7]. Alert messages are produced via an alarm or some other technique to detect human suspicious activities. Nowadays, detecting the suspicious activities of humans in video surveillance plays a vital role in computer vision technology, like observing people's actions in auditoriums, shopping malls, colleges, health centres [8,9], eldercare, prisons, monitoring vehicles, home nursing, military purposes, etc. Recognizing human activities in video surveillance is used to detect human activity in day-to-day life to mitigate the suspicious activity of humans [10].

In computer vision and image processing technology, suspicious human activity is classified into normal human activities and abnormal human activities [11–13]. Human Normal activities are also called usual activities. This is done in public places by humans, such as jogging, running, walking, clapping, boxing, etc. Humans perform rare activities like theft, fights, running crowds, attacks, crossing borders, and leaving luggage for explosive attacks at public places, which are called suspicious or unusual activities [14,15]. Pre-processing, feature extraction, segmentation, action detection, and classification are some steps in identifying suspicious human activity in surveillance videos. In recent times, many methodologies have been proposed for HAR in video surveillance.

HAR in video surveillance is still complicated due to background noise, the influence of the environment, object occlusion, viewpoint, variations in intra-class viewpoint, non-rigidness of the human body, loss of information, scaling of an object, and variation in illuminations [16,17]. Numerous existing approaches proposed to detect human activities in video surveillance may not ensure the accuracy of detecting activities [18,19]. However, the challenge is to automatically detect human activities in surveillance video and predict human abnormal activities [20]. Hence, the proposed approach introduces an ML-based HAR to improve accuracy in recognizing human activities and reduce the time for recognition.

Thus, a better understanding and in-depth analysis of these videos are essential to alert the security system. However, it is not trivial to recognize the human action for all applications with high accuracy due to unconstrained environments in real-time applications. Factors such as the complexity of activities, distinct backgrounds, dynamic recording, and action speed with different application areas make the human action identification process a challenging task. Thus, the proposed technique is based on the objective of achieving fast and accurate action recognition from input video. The proposed approach introduces an ML-based HAR in surveillance video that depends on feature extraction techniques to improve HAR with high accuracy and minimal consumption time.

The proposed HAR technique depends on the following contributions.

- To introduce a new automatic HAR model by fusing diverse feature extraction techniques and ML classifiers to recognize the accurate human activities from surveillance videos.
- The presented multi-feature extraction involves handcrafted feature extraction, and the multi-dimensional CNN model allows the extraction of deep hidden information for exact human action detection.
- The fusion of handcrafted and multi-dimensional CNN-based deep feature extraction offers high internal information by learning image features and improving recognition accuracy by correlating a compact set of features.
- The presented fusion model reduces the dimension and minimizes the execution time. Also, the input spatial structure is preserved by varying the dimension of the CNN learning procedure.
- The classification algorithm SVM handles high-dimensional data with non-linear decision boundaries, and the performance is evaluated using different metrics. The proposed HAR performance is compared with existing classifiers.

2 Related Works

The recent research related to HAR from surveillance video is mentioned below:

Kushwaha et al. [21] recognized HAR in video sequences by integrating multiple features. A proposed method integrates the Discrete Wavelet Transform (DWT), HOG, and Multi-scale Local Binary Pattern (LBP) to extract the frame sequences, structural information, and directional information of the frame. The proposed feature descriptor provides efficient and effective activity recognition in realistic scenarios. However, the performance is affected by unconstrained environments.

Deotale et al. [22] suggested an approach for HAR in untrimmed video in the sports domain based on the Deep Learning (DL) technique. In this method, HAR by the DL approach includes Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM). CNN was used to classify and recognize activity information like gate detection, emotion detection, gesture detection, etc. LSTM has extracted the sub-activity information. This approach provides better accuracy in recognizing

sub-activity information, but no correct activity sequence was detected.

Girdhar et al.[23] developed a method for HAR in automatic surveillance. In this work, DL-based inception-LSTM was proposed to recognize human activities in video surveillance. The suggested DL approach was used to recognize the suspicious activities of humans, which captures the temporal information first to increase the accuracy rate. However, there was still a need to identify suspicious human activity before it occurred.

Kushwaha et al.[24] suggested an algorithm for HAR in video by combining the orientation and magnitude details of optical flow (OF) and video motion. Support Vector Machine (SVM) was used to classify and validate the results for activity detection. Hence, the presented method was not suitable for multi-view and real-time environments.

Alawneh et al.[25] enhanced the HAR with respect to time series augmented data and the DL approach. The proposed approach evaluates the time series augmentation to reduce the overfitting problems and recognize the human activity using vanilla Recurrent Neural Network (RNN), Gated Recurrent Units (GRU), and LSTM. By using these methods, the rate of accuracy has improved. Still, it failed to detect human activities in an efficient manner, which increased the time required and made detection less accurate.

It is complicated to obtain discriminated features from human actions due to variation in the human body. The model must be trained to provide intelligent solutions for human recognition. Santos et al.[26] recognized human action based on DL approaches on a small dataset. It is based on the objective of getting better results in bigger datasets with higher performance. The architectures such as C2D-Resnet50, Inflated 3D ConvNet (I3D), and SlowFast were compared with baseline parameters. Marcondes et al.[27] analyzed surveillance systems for violence detection. A reliable surveillance system with environmental variation was suggested by adapting technologies. The comparison of traditional methods is mentioned in Table 1.

Table 1. Analysis of various techniques used in human activity detection

Author Name	Approaches used	Objectives	Advantages	Disadvantages
Kushwaha et al.[21]	DWT, LBP	HAR in surveillance video is carried out using a feature descriptor.	Provide efficient and effective recognition of activity in real-time scenarios.	Poor Outdoor activity detection in the environment.
Deotale et al.[22]	CNN, LSTM	Human activities are recognized through sub-activity information.	Provide better accuracy in recognizing sub-activity information.	Detecting accuracy in the correct sequence is not efficient.
Girdhar et al.[23]	Inception-LSTM	Automatically recognize the suspicious activities of humans in surveillance video.	Capture the temporal information first to increase the rate of accuracy.	The accuracy of detection is not precise.
Kushwaha et al.[24]	HOG, SVM	Enhancing the HAR by integrating HOG SVM techniques.	Provide efficient and effective feature vectors for human activity detection.	It is not suitable for all real-time and multi-view scenarios.
Alawneh et al.[25]	RNN, LSTM, and GRU	Enhance human activity detection.	Increase the rate of accuracy in human activity detection.	The detection rate is less, and time consumption is high.
Santos et al.[26]	C2D-Resnet50, I3D, and SlowFast	Better performance was achieved.	Consistent performance is attained even with the small dataset	The late models are required to model for audio and video models.
Marcondes et al.[27]	Adaptive theory	Improving the vehicle surveillance system.	Reliable and secure surveillance system.	Surveillance patterns are not explored.

2.1 Problem Statement

Human activity detection in video surveillance is a challenging one. Numerous approaches are proposed to detect the activity of humans in video surveillance, but they do not provide an efficient accuracy for recognizing the activity of humans. The major issues with the HAR are noisy background, environmental impacts, object occlusion, variation in illuminations, etc. Therefore, an efficient and effective approach to recognizing human activity is required. Thus, the Machine Learning HAR is proposed to automatically predict human activities in video surveillance and provide higher accuracy of activity detection with

less computation time.

3 Proposed Methodology

Human activity detection technology is used to represent the presence of a person in a particular place or environment. The basic process of object detection from input video is captured from background images; then, it categorizes the objects according to their class. In surveillance video, the detection of human activities plays a vital role due to crowd movement, unique person identification, suspicious human activity, unusual action recognition, etc.

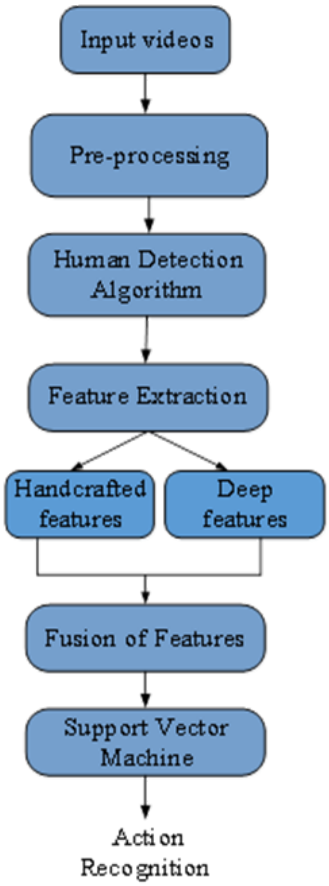

Fig 1. Flow diagram for Proposed Approach

Figure 1 shows the entire framework of the proposed HAR. The proposed HAR is carried out by four main steps: human detection algorithm, feature extraction, SVM classifier, and action recognition. Initially, the videos are converted into frames, and then the frame gets segmented to sort out the humans from other objects or background noise in the input video frames. After the identification of the human, the human detection algorithm utilizes the HOG feature to recognize human activity based on ML techniques.

After the detection of human activity using HOG, the tracked objects are tracked through diverse feature extraction techniques. Feature extraction is characterized by conventional methods and multi-dimensional CNN. Then, the fusion of features is used to integrate the feature extraction techniques to improve the accuracy rate in analyzing human actions. For classification purposes, the proposed approach uses SVM to classify human activities, resulting in humans' required action in surveillance video.

3.1 Input Video Pre-processing

Pre-processing is one of the most important techniques and plays a significant role in image classification. Extensive computations and inaccurate recognition performance may occur without performing pre-processing in the initial stage. Here,

the input is the video data, which includes the grouping of frame sequences at 30 frames per second (fps). The pre-processing method used to sequence the frames of human action videos is the Mean Subtraction (MS) strategy. The given sequence of frames can be characterized as $Fr_{\text{seq}} = \{Fr_1, Fr_2, \ldots Fr_i, \ldots Fr_N\}$. Each frame is represented as $Fr_i \in R^{h \times w \times c}$, where w indicates width, h means height, and C represents the RGB color channel. The mean value of the frame is computed using the below formula.

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

The mean subtraction obtained by RGB frame pre-processing is represented as

$$Fr_{pre} = \mu ImN_i - \mu Fr_i \tag{2}$$

Where, μFr_i illustrates the mean value of the frame, Fr_{pre} indicates the frame pre-processing, and μImN_i represents the ImageNet mean from the ImageNet dataset [28], and the values used are $[0.485, 0.456, 0.406]$.

3.2 Human Detection Algorithm

A human detection technique is used to detect human activities in video surveillance by creating bounding boxes like square or rectangular boxes to determine the person's exact position in the video frame. The main steps in the human detection algorithm for recognizing human activities are HOG and SVM classifier, as shown in Figure 2.

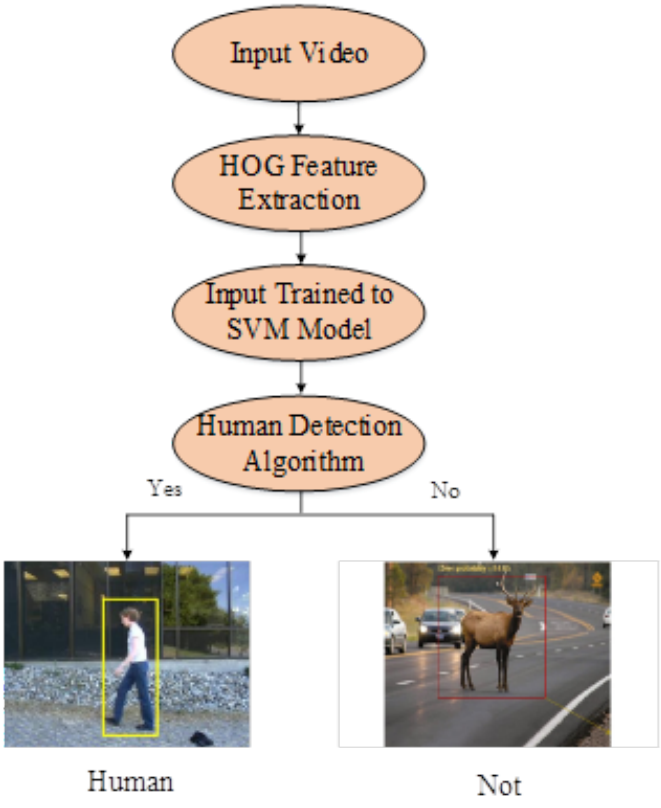

Fig 2. Human Detection Algorithm Flow

HOG feature is used in the human detection algorithm to enhance the HAR from given input video frames. SVM classifier is used to classify the presence of a person in detected objects from the input frame or not, which returns the detected person by representing the rectangular bounding box with width and height values. HOG [29] is utilized to extract and recognize features from video frames. HOG is also called a feature descriptor. HOG feature has the ability to detect objects even in multiple occlusions. Hence, the proposed approach uses an HOG feature to detect the objects in the input video frame. HOG feature

is the most effective feature used in human detection algorithms to extract object shape information from the input frame sequence. The image frames are initially divided into small spatial regions called cells to construct the efficient HOG descriptor. Local one-dimensional (1D) edge orientations or HOGs are computed for each pixel of the cell.

The HOG descriptor computes the intensity gradient distribution to deliver the information of the local shape object. The HOG feature is used to analyze image patches at different scales in multiple image locations. In HOG, horizontal and vertical gradients are evaluated to get the absolute value of x and y gradients. At last, the magnitude of the gradient is measured to enhance the features. Without analyzing the previous gradient knowledge, HOG estimates the object's shape using a local intensity gradient. In a localized manner, the HOG feature captures the edges and the gradient to obtain the object shape information.

Calculations involved in HOG features are:

- Initially, the pre-processing method is used in the HOG feature to eliminate background noise during the collection of frame samples. Pre-processing the images in HOG is divided into size of N x N pixels. In M-bins HOG, all pixels are collected and computed to build a final features vector. HOG features utilize the pre-processing method to improve the gradients for each cell size. Resize the image by having some fixed size or ratio of an image to extract the feature because other images are divided as
- Derivatives of the images are placed in X and Y directions.
- Gradients are calculated for every pixel in the image. The gradient is computed by taking the small patch from the input frame. Changes in X and Y directions are considered as gradients. Let's calculate the gradient magnitude and gradient direction α by following the equation.

$$\alpha = tan^{-1}\left(\frac{Y}{X}\right) \tag{3}$$

- The next step is to calculate the histogram orientation for each cell where the image is divided into 8 x 8 pixel cells. Calculating the histogram among small local regions is called spatial binning. The best result for human detection requires an unsigned 9-bin histogram between 0 and 180°.
- Block normalization is used to normalize the histogram of each cell using the overlapping blocks of cells. Attaching the histograms of all cells in a block produces a vector. A metric is run on the vector to normalize the histogram and reduce the effects of lighting variations. Hence 16 x 16 block contains four histograms to produce 36 x 1 element vector. Block normalization in HOG reduces the illumination contrast difference and improves the detector performance.

3.3 Multi-Feature Extraction

In the proposed approach, some human actions are categorized as follows: human-object interaction, playing musical instruments, human-human interaction, and sports. Hence, some actions are also used for experiment purposes: baby crawling, body weight squats, brushing teeth, jumping rope, mopping the floor, and push-ups. The handcrafted feature extraction techniques are Gabor Wavelet Transform (GWT), Auto correlogram, Grey-Level Co-occurrence Matrix (GLCM), and HSV Histogram. The deep features are extracted using Multi-dimensional CNN. Figure 3 shows the feature extraction techniques.

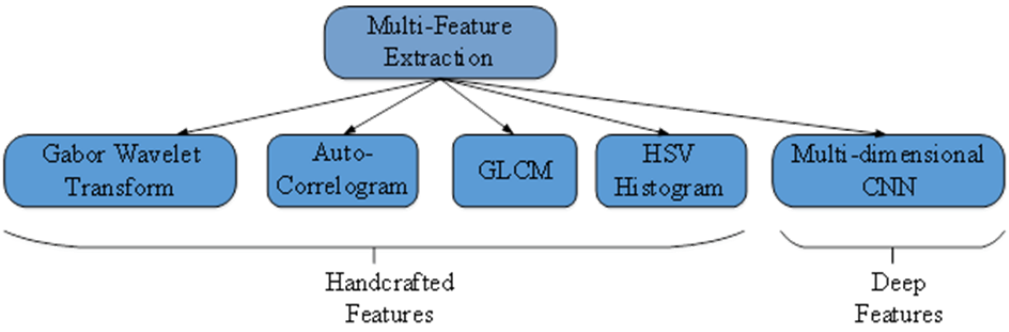

Fig 3. Feature Extraction

3.3.1 Gabor Wavelet Transform

Gabor Wavelet Transform (GWT)[30] is also called a linear filter. It is used to analyze the image to check whether it is in a localized region around the point with a specific frequency and direction. Gabor functions play a vital role in image processing by analyzing and extracting the features and textures from input video. Multiplying the plane wave with the Gaussian function produces the Gabor function. Using the GW filter in feature extraction enhances low-quality images because grey-scale character images directly carry out the extraction process in the GW filter. Some important features captured by the GW filter are orientation selectivity, spatial localization, quadrature phase relationship, and frequency selections. The 2D GWT filter is represented in Equation (4).

$$GF(X,Y) = g_{\alpha}(X,Y)exp(2\mu i\lambda(Xcos\theta + Ysin\theta)) \tag{4}$$

Where,

$$g_{\alpha}(X,Y) = \left(\frac{1}{\sqrt{2\mu\alpha_X}\alpha_Y}\right)exp\left(\frac{-1}{2}\left(\frac{X^2}{\alpha_X^2}+\frac{Y^2}{\alpha_Y^2}\right)\right) \tag{5}$$

Consider $g_{\alpha}(X,Y)$ is the function of the 2D Gaussian function used for the sinusoidal harmonic 2D signal. λ is the oscillating frequency used to determine the key parameters like α, θ in GWT.

$$GF_{uv}(X,Y) = \sum_i \sum_j A(X-i,Y-j)\psi^*_{nv}(i,j) \tag{6}$$

Here u, v denote the scale and orientation, (ψ) mother wavelet function, where the complex conjugate is ψ^*_{uv} and i, j is the mask size of the Gabor filter. The rotation and dilation of (ψ) in the GWT filter are represented in the equation below.

$$\psi(X,Y) = \left(\frac{1}{2\alpha_X\alpha_Y\pi}\right)exp\left(\frac{-1}{2}\left(\frac{X^2}{\alpha_X^2}+\frac{Y^2}{\alpha_Y^2}\right)\right)exp(2\pi i\lambda X) \tag{7}$$

With varying scales and orientations, mean and variance are estimated in the Gabor filter. The computed pair for energy in the filtered image (u,v) is represented by the following equation.

$$EC(u,v) = \sum_X \sum_Y |GF_{uv}(X,Y)| \tag{8}$$

In feature extraction techniques, GWT filters are used to extract the local image feature representation and texture information from the input video.

3.3.2 Auto-Correlogram

Auto-correlogram[31] is used for extracting the colour features from defective and non-defected images. Auto-correlogram is utilized to indicate spatial correlation distribution of identical colours. The auto-correlogram of the image captures the spatial correlation to check the strength similarities between the corresponding images. In an image, the colour relationship between the pixels is described by the colour correlogram method. This method shows the spatial relationship between various colour pairs and the level of each colour in a low-emissivity coating of a damaged image. Hence, the auto-correlogram is very strong in reducing the variations in shape appearance due to changes in camera positions, zooming objects, etc.

3.3.3 GLCM

GLCM[32] is also called a texture descriptor, which is used to measure the brightness value and extract the features of images based on the occurrence of pixel pairs. GLCM provides detailed information about the input images, such as adjacent interval, direction, variation in amplitude, etc. By using the spatial correlations of the grey level, GLCM describes the characteristics of the texture image. Scalars like correlation, homogeneity, and image contrast describe the GLCM. Construct the GLCM by calculating the values of grey-level intensity in a given image based on the linear relationship between two pixels. GLCM feature extraction focuses on determining the spatial composition of detected objects in the crowd's density. Also, it extracts the information from the input images.

3.3.4 HSV Histogram

Hue Saturation Value (HSV)[33] Histogram is used in the feature extraction stage to separate the input image into textured class and non-textured class. HSV represent the colour format to describe the colours by their shade, saturation, and brightness

value. HSV histogram is a very effective method to capture the multi-model patterns of colour information to recognize human activity. The respective histogram is updated by one for each pixel. In the image, each and every bin holds the percentage of pixels corresponding to the HSV colors. The Hue values are chosen between 0-350° ranges, and their corresponding value may differ from red. The saturation value may differ from 0-1. In HAR, the HSV histogram is used to enhance the edge information and position information and preserve the invariance in optical and geometric deformation. Initially, HSV color space is generated, and the color histogram is formed by counting the number of pixels in each quantization unit. By calculating the difference in HSV histogram distance, the input video frame difference is determined. Thus, the HSV histogram is used to extract the features of human activity to decrease measurement problems and represent colour and texture information.

3.3.5 Multi-dimensional CNN based deep feature extraction

In addition to the handcrafted features, the CNN features are directly learned from the video using DL-based approaches. The feature extraction efficiency is improved with the modelling of 3D-CNN, 2D-CNN, and 1D-CNN. In the proposed approach, automatic feature extraction can be accomplished through CNN. The feature extractions obtained through 3D-CNN[34], 2D-CNN[35], and 1D-CNN[36] have been utilized in recent research in which satisfactory performance was achieved. The extracted features from these algorithms are differentiated from each other. By combining these algorithms, more representative features are integrated to improve the proposed model. The proposed CNN approach requires three input layers to handle the sub-models. The similarities between 3D-CNN, 2D-CNN, and 1D-CNN are shared by convolution, flattening, and max-pooling operations. The architecture of feature extraction based on 3D-2D-1D-CNN is shown in Figure 4.

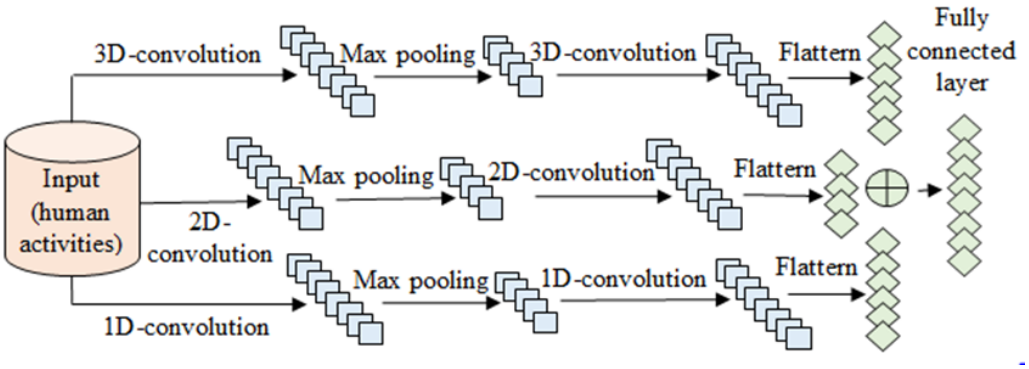

Fig 4. The framework of feature extraction with 3D-2D-1D-CNN technique

The convolution operation is invoked for output generation between the sub-models by representing the sample pattern change. The pooling layer minimizes the data size, model complexity, and model overfitting. The maximum values from the block are considered for extracting significant features with the maxpooling operation. The data is converted into a 1D matrix and fed to the fully connected layer. The features of each sub-model are integrated into the fully connected layers. It has the advantage of effective coordination, and it supports more representative feature extraction to enhance the overall performance of the model.

Here, the deep features are extracted by exploiting the pre-trained VGG16 model for learning hidden deep patterns from the sequence of human activity frames. In deep learning, there exist many CNN models, but the most popular VGG 16 framework is chosen for the following reasons. This model consists of a total of 16 weight layers for feature representation and helps to minimize the computational cost while extracting features from the frame sequence. Moreover, VGG 16 is deeper than other CNN architectures and has 138M (Million) learning parameters. The Multi-dimensional CNN is divided into several layers like five Conv (convolution) blocks, max-pooling, and 3 Fully-Connected (FC) layers. Each block comprises Conv layers accompanied by max-pooling layers. The non-linear activation function ReLU is regularly used in model training. The input frame is pre-processed into a size 224 x 224 x 3 before feeding it to the first Conv layer of the VGG16 model. The expression is given as,

$$F_{Conv} = K * Fr_{pre} \tag{9}$$

Where, K indicates the kernels (filters) of size 3×3, F_{conv} denotes the feature maps generated from the pre-processed frame Fr_{pre}. Now, the above equation is further expanded as,

$$F_{Com1} = \sum_m \sum_n K[m,n] Fr_{pre}[p,q] \tag{10}$$

Where, p and m signifies the rows, whereas q and n indicates the columns of Fr_{pre} frame, respectively. The second Conv layer is further used to shrink the input image spatial dimension, which is expressed as,

$$F_{com2} = \sum_m \sum_n K[m,n] F_{com1}\ [p,q] \tag{11}$$

Next to the second Conv layer, the max-pooling layer is used to minimize the feature vector size. The max-pooling layer output is given as,

$$Max^1_{pool} = \max_j^i \left(F^q_{p\,conv2} \right) \tag{12}$$

Where, j and i represents the max-pooling filters. The feature map extracted from the fifth block of the max-pooling layer is expressed as,

$$Max^5_{pool} = \max_j^i \left(F^q_{p\,Coms} \right) \tag{13}$$

After the pooling layers, three FC layers are added, which is useful for extracting intrinsic information. The expression of FC's first layer is mentioned as,

$$\begin{aligned} F_{HA} &= \left(Max^5_{pool}, d_{flatn} \right) \\ F_{HA} &= Fr^{pool\,5}_{pre} \end{aligned} \tag{14}$$

Where, the features of human actions are denoted as F_{HA}, the flattened layer is indicated as $d_{\text{flatt.}}$. The final layer responsible for the extraction of features can be modeled as,

$$F_{DP} = \left(F_{HA}, d_{flatn} \right) \tag{15}$$

3.4 Feature fusion

In the proposed approach, HAR is performed by fusion of features because feature-extracting techniques have difficulties in extracting shape, texture, and colour-based features. In this work, the HAR is presented using feature fusion and a classification-based human activity detection algorithm. Different features such as Gabor Wavelet transform, GLCM, Auto correlogram, and HSV Histogram are extracted and fused to perform effective HAR. Hence, this work introduces a multi-feature fusion approach to integrate details of different data features for performing feature extraction and activity recognition. It combines the feature dimension, texture, shape, scale-orientation, and colour features. Finally, the feature classification can be accomplished through the SVM classifier.

Classifying multiple patterns is challenging due to inter-class separability, the number of action classes, and inter-class separability. The classification with almost the same kind of feature is challenged. Hence, the problem can be resolved by combining different kinds of feature sets. It leads to the generation of a single feature vector. The fusion approach has three kinds of fusion: decision level, image level, and feature level. The feature level fusion is used with the proposed approach to deal with accuracy and execution time. The feature fusion is based on the objective of obtaining better recognition accuracy by matching the original features.

In the proposed feature fusion, the size of the maximum feature vector is estimated. The maximum probability features are used for padding. Maximum probability value with zero padding is utilized for generating equal-length vectors. The fusion process is described as follows. The features extracted with GWT, AC, GLCM, HSV, and M-CNN are represented with $\eta_{GWT}, \eta_{AC}, \eta_{GLCN}, \eta_{HSV}$ and η_{M-CNN}. The entire training samples are represented with N. Initially, the higher and lower feature vector length is computed using the following equation.

$$\eta_{max}(F) = Max\left(\eta_{GWT}(j), \eta_{AC}(j), \eta_{GLCM}(j), \eta_{HSV}(j), \eta_{M-CNN}(j)\right) \tag{16}$$

$$\eta_{min}(F) = Min\left(\eta_{GWT}(j), \eta_{AC}(j), \eta_{GLCM}(j), \eta_{HSV}(j), \eta_{M-CNN}(j)\right) \tag{17}$$

Afterward, the maximum feature vector length is followed by making the vector length of all features equal. The minimum feature vector probability is computed to select the feature with a higher occurrence of the entire vector.

$$Q = \frac{\sum_{j=1}^{P} q_j(y)}{P} \tag{18}$$

Where, the probability matrix is represented with Q, the number of favourable occurrences is denoted with $q_j(y)$, and M represents the number of minimum features $\eta_{min}(F)$. A higher probability value is used for padding equal-length vectors. After making the length of the feature vector equal, the features $\eta_{GWT}, \eta_{AC}, \eta_{GLCN}, \eta_{HSV}$ and η_{M-CNN} can be fused with a parallel approach.

$$\eta_F(FF) = max(\eta_k(j)), k \in \{\eta_{GWT}(j), \eta_{AC}(j), \eta_{GLCM}(j), \eta_{HSV}(j), \eta_{M-CNN}(j)\} \tag{19}$$

Where, $\eta_F(FF)$ represents the fused feature vector. After getting the fused feature, the dimension is reduced to minimize the execution time.

3.5 SVM Classifier

SVM [37] is an ML-based algorithm that plays a vital role in performing classification problems like activity recognition, object recognition, face recognition, speaker identification, etc. The SVM technique is used in human detection algorithms to enhance detection performance and classification. In the SVM classifier, the hyperplanes are designed to split all data points. Thus, SVM looks for the best hyperplane to separate two data classes to form a model used for classification. Hyperplane provides a large margin between two classes to enhance the classification process in SVM. The support vectors are considered as data points in SVM. These data points are placed in the boundary of the slab to separate the hyperplane. The main goal of the SVM method is to find a classifier with the least generalization error, which is attained by the boundary of maximum margin by the following equation.

$$Min\left(\frac{1}{2} \parallel WV \parallel^2 + \alpha \sum_{a=1}^{N} \beta_a\right) \tag{20}$$

$$Y_a(\langle WV, X_a \rangle + b) \geq 1 - \beta_a \forall_a \tag{21}$$

Here, WV is a weight vector, which is denoted inside the $\langle ., \rangle$ $product, \alpha > 0$ is a regularization coefficient, b is the bias, and the slack variable is $\beta_a \geq 0$. The human detection algorithm uses the SVM classifier to recognize and classify the actions independently. Even small human actions like walking, running, talking, etc., are observed to recognize the actions. In the human detection algorithm, the SVM classifier is mainly used to calculate the hyperplane with a high margin, which splits the feature vectors for every category of data. Then, the detected human actions are visualized as M-bounding boxes with the value of height and width in $X - Y$ origin. In the proposed approach, some actions are used for experiments: baby crawling, body weight squats, brushing teeth, jumping rope, mopping the floor, and push-ups. Based on the feature extraction techniques, SVM classifies the six activities for recognizing the actions of humans.

HAR finds several applications in diverse fields like sports, surveillance, medical diagnostics, video annotations, etc. However, based on trained features, the recognition system categorizes the spatio-temporal (ST) feature descriptor from the video sequence. However, most classifiers suffer from issues like large-sized feature vectors and extended training time. Therefore, SVM uses existing ST feature descriptors to resolve the problems associated with HAR. Here, SVM is utilized to recognize activities within the High-Dimensional (HD) feature space. In the proposed study, SVM was chosen and used as a classifier because this ML classifier offers superior results compared to other classifiers. Also, it is very effective when dealing with smaller and large-scale training datasets and shows improved performance in HD spaces. During the training process, hyperplanes are generated in HD space, which separates the training data into multiple classes. The kernel function can solve the non-linearly separable classes in SVM. Here, the RBF kernel is utilized as the kernel function, which makes use of γ (the regularisation factor) that adjusts the speed of the kernel function and identifies the flexibility of separating the SVM hyperplane. The RBF kernel processes the feature into a huge feature space. During the initial stages, SVM was developed to perform binary classification but can also perform multiclass classification accurately. The fused feature vector is given as input by the SVM classifier, which further accommodates each image to the subsequent label and accurately recognizes the activity being performed. Thus, the proposed SVM classifier performs better in labeling different activities at minimal time and with a low computation cost.

The proposed HAR is more efficient than the existing DL-based techniques in terms of HAR accuracy and less processing time. It can be attained with different feature extraction and feature fusion techniques. The feature fusion technique provides the pattern of action with the compact feature set. Thus, the proposed HAR model utilizes the relationship between the hybrid features and fuses with the feature level fusion in order to attain better performance.

4 Results and Discussions

The proposed technique is evaluated in MATLAB software and analyzed with traditional approaches to detect the activities of humans by using the UCF101 and UTKinect datasets. It is processed with an Intel(R) core (TM) i3-7100U processor and @2.40Hz 2.40 GHz Central Processing Unit (CPU), which are installed at 8.00GB RAM, and the system model is a 64-bit operating system, x64-based processor.

4.1 Dataset Description

4.1.1 UCF101 Dataset

The proposed approach utilizes the UCF101 dataset [38] for detecting HAR in video surveillance. The UCF101 is the large dataset for HAR from video surveillance. It has real user-uploaded videos with different camera motions and cluttered backgrounds. All videos are captured in several cluttered scenes, illumination conditions, various poses, partial occlusions, orientation of human objects, camera motion, various viewpoints, etc. UCF101 dataset has more number of action classes than others. Hence, the dataset is well-suited for HAR with the ML algorithm. Table 2 shows the experimental setup for the proposed approach.

Table 2. Experimental Setup for the proposed approach

Number of Actions	Training Videos and Frames	Testing Videos and frames
six	239/7170	60/1800

Table 3. Attributes of the UCF101 dataset [38]

Actions	101
Clips per Group	4-7
Min Clip Length	1.06 sec
Resolution	320240
Groups Per Action	25
Total Duration	1600 mins
Frame Rate	25 fps
Clips	13320
Mean Clip Length	7.21 sec
Max Clip Length	71.04 sec
Audio	Yes (51 actions)

The UCF101 dataset has a major role in HAR. Table 3 illustrates the detailed summary of the UCF101 dataset, consisting of 13k video clips and 27h of video data. Each video has an average length of 180 frames, consisting of 101 actions ranging from day-to-day life activities. Each group holds certain categories, such as Human-Object Interaction holds 20 categories, Body-Motion holds 16 categories, Human-Human Interaction holds 5 categories, playing musical instruments holds 10 categories, and sports holds 50 categories.

Therefore, it's very complementary to consider 101 activities for the segmentation process. Thus, the proposed approach only considers six experimental actions: baby crawling, body weight squats, brushing teeth, jump rope, mopping the floor, and push-ups. In the proposed work, 6 classes were selected from the UCF dataset. The classes are selected based on a large proportion of data. Due to the selection of majority class data, the result is highly impacted by avoiding overfitting or misclassification. It tends to increase classification performance and minimal processing time. In this study, one of the reliable ways to evaluate the proposed ML performance is to train the model with available data and assess its classification performance. Here, the Train/Test Split approach is used to separate a portion of data and to use that data for validation. The entire dataset is randomly divided into training and test sets. The data split ratio considered for training and testing is 80:20. Figure 5 represents the actions considered in the proposed method.

4.1.2 UTKinect Dataset

This UTKinect dataset [39] includes ten indoor daily activities, such as standing up, walking, picking up, sitting down, throwing, carrying, pulling, pushing, clapping hands, and waving hands. Each action is performed by ten people two times. This dataset is represented in three modalities, namely RGB, 3D skeleton, and depth. In total, 200 action samples are seen in this dataset.

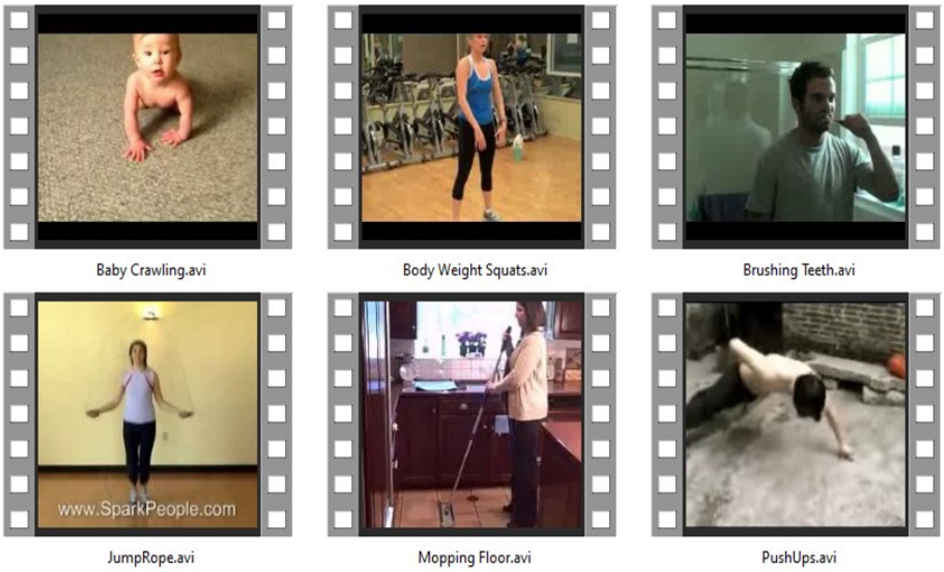

Fig 5. Actions used for the proposed experiment

The whole dataset is divided into training and test sets. The ratio of data split considered for training and testing is 80: 20. In the proposed work, 6 actions were selected from the UTKinect dataset, such as walk, sit down, stand up, pickup, throw, and clap hands. Figure 6 denotes the actions considered from the UTKinect dataset.

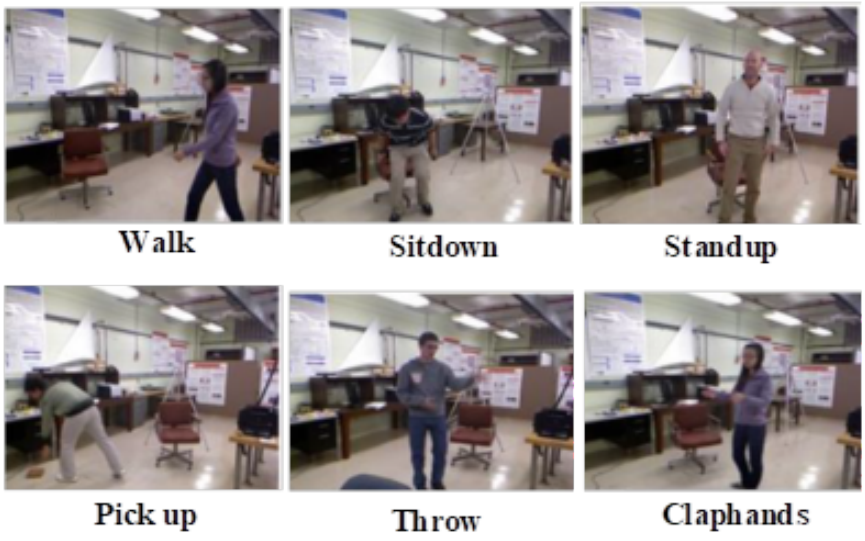

Fig 6. Actions used from UTKinect for the proposed experiment

4.2 Evaluation Metrics

The proposed approach obtained fast and accurate HAR in surveillance video based on some actions from the UCF101 and UTKinect datasets. Some popular metrics used for evaluation are accuracy, recall, precision, specificity, MCC, and F-score. Accuracy is determined by calculating the percentage of correct predictions from the total number of samples.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions\ made} \tag{22}$$

$$\text{Precision} = \frac{T^P}{T^P + F^P} \tag{23}$$

$$\text{Specificity} = \frac{T^N}{T^N + F^P} \tag{24}$$

$$\text{Recall} = \frac{T^P}{T^P + F^N} \tag{25}$$

$$F - \text{score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{26}$$

$$MCC = \frac{T^P \times T^N - F^P \times F^N}{\sqrt{(T^P + F^P)(T^N + F^N)(T^P + F^N)(T^N + F^N)}} \tag{27}$$

Where, True Positive is T^P, False Positive is F^P, True Negative is T^N, MCC illustrates Matthews Correlation Coefficient, and False Negative is F^N.

The loss function represents the inaccuracy of classification. It is estimated with the error value between output prediction and target value.

4.3 Time Analysis

The time consumed by the existing and proposed feature extraction methods is mentioned in Table 4. In the proposed study, the time analysis is conducted based on single-feature extraction and fused-feature extraction. Gabor Wavelet Transform (GWT) obtained (0.0286) time for extracting the local and texture features. Auto-correlogram consumes (0.0290) time for extracting the colour features from the defected and non-defected frames. The grey-level co-occurrence Matrix obtains (0.0293 sec) time for measuring the brightness values, and the HSV Histogram consumes (0.1304 sec) time for extracting the edge and position information to recognize the human activities in surveillance videos using proposed techniques. For recognizing the human activities in surveillance video, the existing Space-Time Interest Point (STIP) detection methods and HOG consume 0.048534 times, and HOF obtains 12.533267 times for recognition. The run time for the proposed fused feature extraction is 0.0124 sec.

Table 4. Time Analysis of Existing and Proposed Feature Extraction Methods

Method	Time (sec)
STIP + HOG	0.048534
HOF	12.533267
GWT	0.0286
Auto-correlogram	0.0290
GLCM	0.0293
HSV Histogram	0.1304
Proposed [Fused Feature Extraction]	**0.0124**

The time consumption of the proposed classification technique with existing classifiers is analyzed in Table 5. To prove the effectiveness of the proposed HAR model, different ML classifiers are compared with reference to the run time in seconds. The proposed and existing classifiers are executed in the MATLAB platform, and the run time analysis is based on the UCF101 dataset. The proposed model acquires lesser run time due to the fused feature extraction with SVM classification. However, the other classifiers, namely NB, DT, RF, and LR, are analyzed based on single feature extraction and take more time than the proposed classifier model.

Table 5. Proposed time analysis performance with various ML classifiers

Method	Time (sec)
Proposed [Fused Feature Extraction + SVM]	**0.254**
Decision Tree (DT)	0.472
Naïve Bayes (NB)	0.343
Random Forest (RF)	0.71
Logistic Regression (LR)	0.68

The runtime of the proposed method (fusion of features + SVM) consumes 0.254 sec, whereas the existing classifiers obtained higher run times as NB (0.343 sec), DT (0.472 sec), RF (0.71 sec), and LR (0.68 sec). The proposed approach consumes less time because of the fused feature extraction procedure combined with the classification process. However, the runtime for existing ML classifiers with single-feature extraction processes is higher compared to the proposed human activity detection algorithm. This proves the efficacy of the ML-based proposed HAR model.

4.4 Performance analysis

The accuracy performance of the proposed HAR is given in Figure 7. The accuracy values are higher for the proposed approach, and they are lower for the existing approaches. The DWT+LBP+HOG and proposed approach performance is higher than that of other approaches. The lower performance is obtained with DWT based approach. The accuracy performance of DWT and HOG-SVM is lower than 80%, and the performance is higher for the remaining approaches. The accuracy results are lower than 80% for MHI-HOG, HOG-SVM, and Multiclass SVM approaches. From the entire number of input samples, the number of true predictions is estimated to measure the model's accuracy. Hence, the accuracy result is mainly based on the amount of true predictions. Accuracy evaluates the classification model with the correctly classified human actions and the total number of human actions.

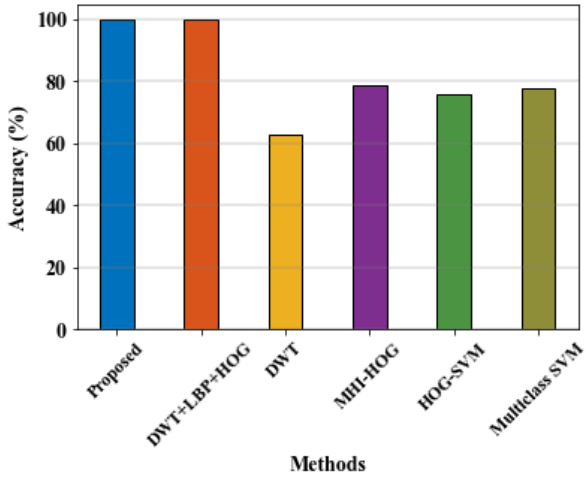

Fig 7. Accuracy performance comparison with existing approaches

The precision result of the proposed HAR is shown in Figure 8. The precision values of human action recognition are higher for the proposed technique and lower for the DWT technique. The precision value is lower than 0.2 for DWT based HAR technique. The recall value of DWT+LBP+HOG was obtained to be higher than 0.9, and the Multiclass SVM performance was lower than 0.8. But all the HAR-based techniques have achieved a performance higher than 0.7 except DWT. From the entire positive identification, the true positives are measured at a precision rate. The relationship between the measurements is determined with precision metrics.

Recall analysis with existing approaches is shown in Figure 9. The recall value is compared with existing HAR-based approaches. The higher recall value is obtained with the proposed approach, and the lower is obtained with the DWT-based approach. The recall value of DWT+LBP+HOG is higher than 0.9, but it is lower than 0.8 for the remaining MHI-HOG, HOG-SVM, and Multiclass SVM approaches. The specificity performance of the proposed approach is shown in Figure 10. The specificity results produce lower results than the proposed one. The performance is higher than 0.96 for DWT+LBP+HOG,

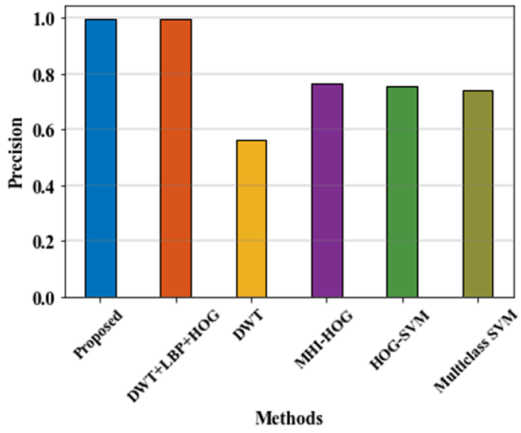

Fig 8. Precision comparison with existing approaches

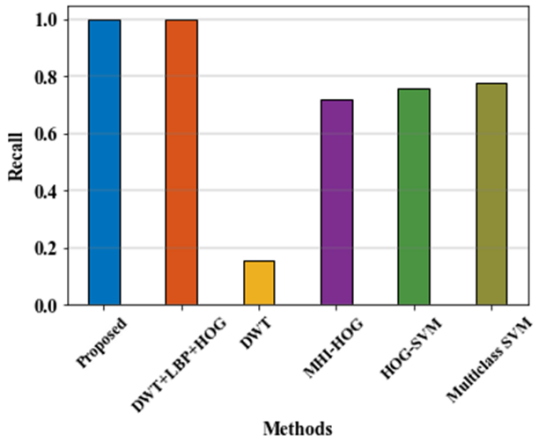

Fig 9. Recall performance with traditional techniques

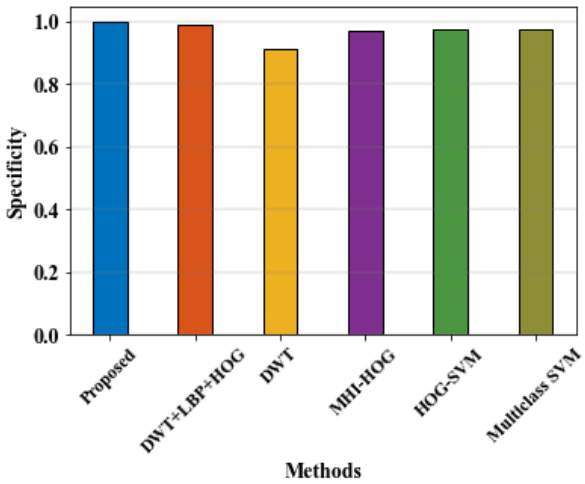

Fig 10. Specificity comparison with existing techniques

MHI-HOG, HOG-SVM, and Multiclass SVM. The performance of DWT is lower than 0.92.

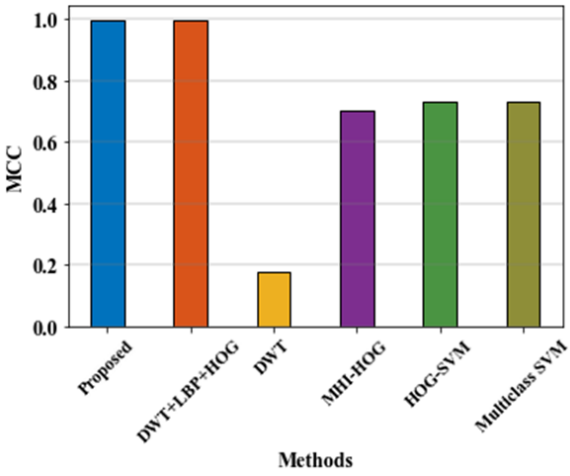

Fig 11. MCC performance comparison with existing approaches

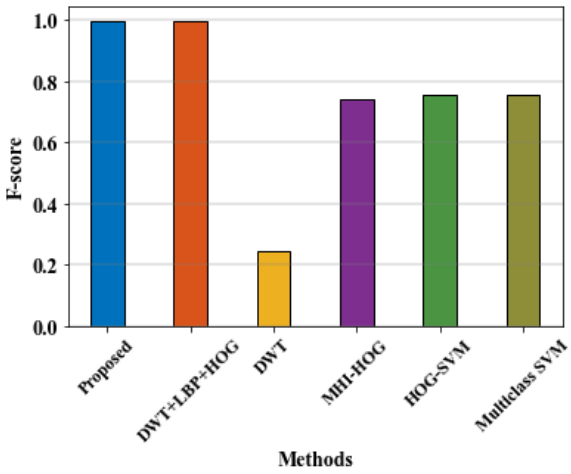

Fig 12. F-score comparison with existing approaches

The MCC performance of the proposed HAR is given in Figure 11. The performance is lower for the existing DWT-based HAR. The performance of the remaining approaches is higher than 0.6. The F1-score result analysis is shown in Figure 12. The f1-score of HAR measures the accuracy of the ML model. It can be obtained by combining the score of precision and recall. For the proposed HAR and existing DWT+LBP+HOG approaches, the performance is higher than 0.9. The optimal level of performance is obtained with the proposed approach. When considering the existing deep approaches, the f1-score value of HAR is achieved at a lower than 0.8. The lowest performance is obtained with the DWT architecture, and the f1-score values are obtained lower than 0.7. The lower performance is due to inefficient parameters; it complicates the training process and leads to misclassification.

Figure 13 illustrates the Confusion Matrix of the proposed approach from the UCF101 dataset. Figure 14 indicates the UTKinect dataset confusion matrix.

Table 6 shows the analysis of the proposed HAR on the UCF101 dataset. The proposed performance isanalyzed using different metrics: accuracy, recall, precision, specificity, F-score, and MCC. Here, the proposed comparison is done with various existing classifiers [21] to prove the effectiveness of the presented human activity detection algorithm. The proposed approach offers better results and can be suitable for real-world environments. The efficiency of the proposed approach is due to the fused feature extraction and classification performance.

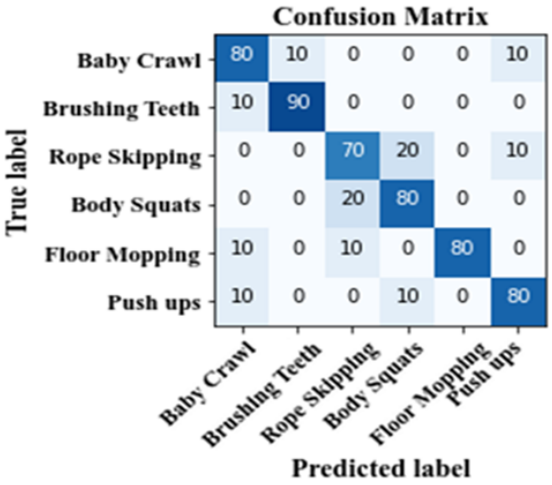

Fig 13. Confusion Matrix of the proposed approach (UCF 101 dataset) in various actions

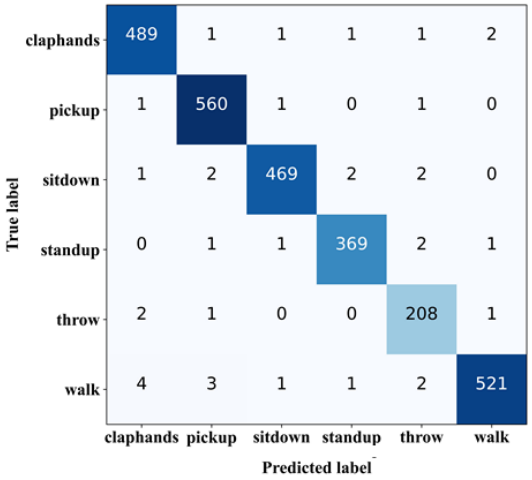

Fig 14. Confusion Matrix of the proposed approach (UTKinect dataset) in various actions

Table 6. Proposed HAR performance with UCF101 dataset [21]

Techniques	Accuracy (%)	Recall	Precision	Specificity	MCC	F-score
Proposed	**99.886**	**0.999**	**0.998**	**0.999**	**0.998**	**0.998**
DWT+LBP+HOG	99.76	0.997	0.997	0.99	0.997	0.997
DWT	62.57	0.156	0.562	0.910	0.180	0.244
MHI-HOG	78.80	0.720	0.767	0.968	0.705	0.743
HOG-SVM	75.80	0.758	0.759	0.973	0.731	0.759
Multiclass SVM	77.66	0.777	0.742	0.975	0.733	0.759

Table 7. Comparison of Proposed HAR performance with different DL algorithms

Method	Dataset	Accuracy (%)
Proposed	**UCF101 dataset**	**99.886**
Two-stream ConvNet [40]	UCF101 dataset	93.30
Implicit CNN [41]	UCF101 dataset	89.8
GMM-KF-GRNN [42]	UCF101 dataset	89.30
AR3D [43]	UCF101 dataset	89.28
S3D-ConvNet [44]	UCF101 dataset	86.6
Encoding RNNs [45]	UCF101 dataset	81.9
P-RRNNs [46]	UCF101 dataset	91.4
Deep Bi-LSTM [47]	UCF101 dataset	91.21
CNN_Bi-GRU [48]	UCF101 dataset	91.79

Table 7 indicates the comparative analysis of the proposed HAR model with different DL algorithms. The proposed accuracy outcome is compared with other DL models like Two-stream ConvNet [40], Implicit CNN [41], Gaussian Mixture Model-Kalman Filter-Gated RNN (GMM-KF-GRNN) [42], Attention Residual 3D (AR3D) Network [43], Segments based 3D ConvNet (S3D-ConvNet) [44], Encoding RNNs [45], P-RRNNs [46], Deep Bi-LSTM [47], and CNN_Bi-GRU [48]. The same UCF101 dataset is used by the existing DL models for accuracy performance comparison. The proposed HAR model obtained superior accuracy (99.886%) outcomes compared to the existing DL approaches.

Table 8. Proposed HAR performance comparison on UTKinect dataset with different DL algorithms

Method	Dataset	Accuracy (%)
Proposed	**UTKinect dataset**	**99.538**
Normalized JT [49]	UTKinect dataset	96.8
SGR [50]	UTKinect dataset	98.5
Deep Networks [51]	UTKinect dataset	96.68
3D Action Recognition [52]	UTKinect dataset	98.9
Multilayer LSTM [53]	UTKinect dataset	95.9
Cuboid CNN [54]	UTKinect dataset	96.1
Graph Model [55]	UTKinect dataset	96.0
EFC [56]	UTKinect dataset	94.9
1D CNN [57]	UTKinect dataset	96.9

Table 8 describes the comparative performance of the proposed HAR model with other networks executed with the UTKinecet dataset. The accuracy of the proposed network is compared with other different architectures to define the efficacy of the proposed HAR network. The existing networks executed with UTKinect dataset used for comparison are Normalized Joint Trajectories (JT) [49], Sparsified Graph Regression (SGR) [50], Deep Networks [51], 3D action recognition [52], Multilayer LSTM [53], Cuboid CNN [54], Graph Model [55], Elastic Functional Coding (EFC) [56], and 1D CNN [57]. The proposed HAR model executed with the UTKinect dataset achieved higher accuracy (99.538%) output when compared with other approaches.

4.5 Analysis of train and test split

From Figure 15, the accuracy and loss curves are measured by changing the epoch size from 0 to 100. In the proposed approach, Multi-dimensional CNN is utilized for feature extraction. The number of epochs varies to analyze the overall performance of the proposed model. With 50 epochs, the accuracy lies between 92% to 95%. If the size of the epoch is 100, the accuracy is increased to 99%. The loss curve decreases with the increase in the size of the epoch. With 50 epochs, the loss value is varied between 0.2-0.1. When it reaches 100, the values are lower than 0.1. A lower error rate is obtained due to the efficient training process of fusion techniques, which accomplishes maximum accuracy. The proposed approach recognizes human activity with high accuracy and less time consumption than existing approaches.

In the present comparison, the second-best classifier is varied from the proposed classifier, with a performance deviation of 0.1%. Also, the time required to compute the proposed HAR is reduced. The traditional approaches require 0.343 sec for HAR.

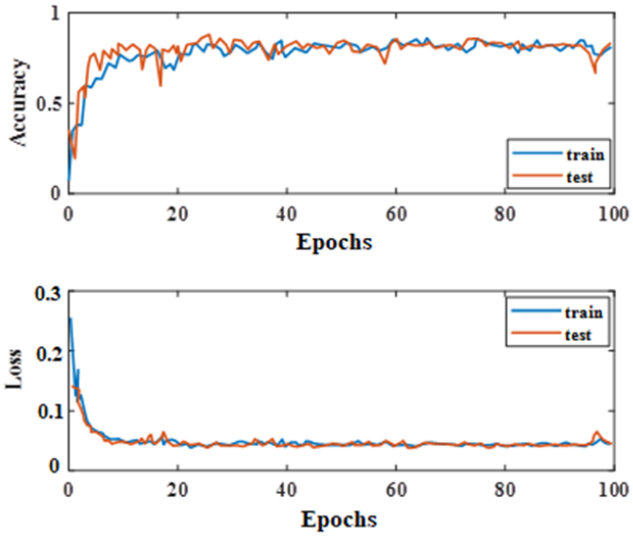

Fig 15. The accuracy and loss result of the proposed approach with varying numbers of epochs

However, in the case of the proposed HAR, the processing time is minimized to 0.254 sec.

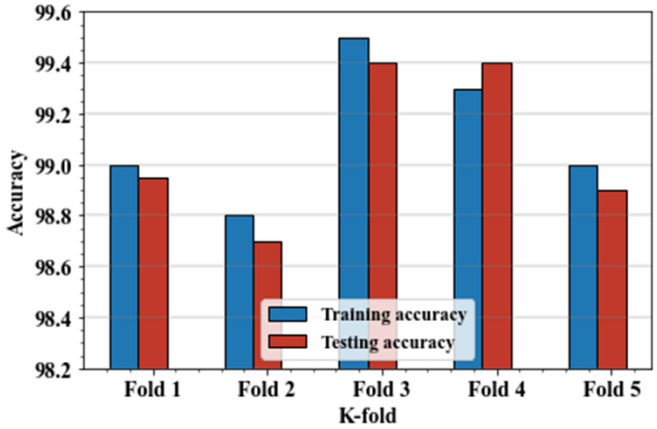

Fig 16. K-fold validation performance for training and testing accuracy

The k-fold cross-validation result for the training and testing performance is given in Figure 16. The number of folds is varied from 1 to 5. In most folds, the training accuracy is higher than the testing accuracy. The training accuracy obtained with fold 1, fold 2, fold 3, fold 4, and fold 5 are 99%, 98.8%, 99.5%, 99.3% and 99%. The testing accuracy obtained with fold 1, fold 2, fold 3, fold 4, and fold 5 are 98.95%, 98.7%, 99.4%, 99.4%, and 98.9%. While increasing the number of folds, the performance is slightly degraded, and the performance is higher with fold 3 and fold 4.

4.6 Discussion

Recently, HAR in computer vision has continued to be an active area of research due to the evolution of numerous intelligent systems like surveillance, analysis, and control. In this work, a new multi-feature extraction and fusion strategy is introduced for the accurate detection of human actions from video surveillance. The SVM classifier categorizes human actions and minimizes the cost of computation in a high-dimensional environment. The proposed HAR method is better than other methods because the results indicate that the multi-feature extraction has provided consistent and robust features for action detection in humans. Also, the fusion strategy is found to be very beneficial in reducing the dimension of the feature subset. The proposed human

detection algorithm eliminates the computation complexity and acquires a more compact feature representation using the set of fused features. Finally, six different human actions from the UCF 101 dataset, such as baby crawling, brushing teeth, rope skipping, body squats, floor mopping, and push-ups, whereas six actions from the UTKinect dataset, such as walking, sit down, stand-up, pickup, throw, clap hands are classified by the SVM classifier, showing multiclass classification. Thus, it's necessary to integrate and evaluate the information to achieve better system accuracy for recognition. Hence, in the proposed approach, the feature extraction techniques are integrated as a fusion of features to improve the accuracy of HAR in surveillance video.

5 Conclusion

Recognition of human activity in video surveillance plays a vital role in computer vision technology and various application fields. This work introduces a fusion model with ML-based human activity detection. In this proposed approach, HAR in video surveillance is discussed in detail. Tracking and recognizing human activity is becoming an essential role in day-to-day life. Many approaches are proposed for recognizing human activities in surveillance videos, but the accuracy of detecting human activities is very low. This work introduces ML-based human activity detection in surveillance videos based on the fusion of handcrafted and deep feature extraction techniques. The important findings of the proposed fused feature extraction and ML-based HAR are characterized as:

- A new human detection algorithm (HDA) is introduced, which utilizes the HOG feature descriptor to sort out the humans from other objects or background noise in the input video frames.
- Processing multiple feature extraction techniques like GWT, Autocorrelogram, GLCM and, HSV histogram, Multi-dimensional CNN for extraction of features.
- A new fused feature extraction strategy is developed, which fuses all the features attained to enhance the accuracy rate of analyzing human actions. The proposed fused feature descriptor offers a discriminate, robust and discriminative feature vector, producing efficient results in unconstrained environments.
- The outcome of the fusion feature has been proven to be better than that of the single feature descriptor, which has higher activity recognition accuracy and low time consumption. The Prediction is invoked with SVM, which categorizes the required human actions in surveillance video.

Initially, the proposed approach introduces a human detection algorithm for recognizing the person in the required environment. After extracting the shape, texture, and colour-based features using feature extraction techniques, the fusion of features is used to integrate the extracted features to achieve better recognition. Finally, the SVM technique is utilized to classify the feature vector by each category of actions and returns the M-bounding box to recognize human actions in video surveillance. The proposed technique is evaluated with MATLAB software and analyzed with earlier methods to detect human activities in video surveillance. In the proposed approach, six actions are taken for experimental purposes from UCF101 and UTKinect datasets to prove the superiority of the proposed fusion based HAR model. The proposed approach gives more efficient results than traditional approaches in achieving accuracy (99.88%) (99.538%) and less time consumption for detecting HAR. The limitation of this research is only a minimal number of human actions are identified. In the future, this work can be developed with advanced DL models for recognizing human activities. Also, it is planned to recognize more human activities and to analyze the difficulties and challenges faced. In addition, it is considering real world data and processing with multiple persons that interact in a similar scene. Moreover, the performance can be analyzed using real-time datasets that consider high-level, complex activities.

References

1) Singh T, Vishwakarma DK. A deeply coupled ConvNet for human activity recognition using dynamic and RGB images. *Neural Computing and Applications*. 2021;33(1):469–485. Available from: https://dx.doi.org/10.1007/s00521-020-05018-y.
2) Snoun A, Jlidi N, Bouchrika T, Jemai O, Zaied M. Towards a deep human activity recognition approach based on video to image transformation with skeleton data. *Multimedia Tools and Applications*. 2021;80(19):29675–29698. Available from: https://dx.doi.org/10.1007/s11042-021-11188-1.
3) Anuradha SG, Teja KD. Deep Learning based Human Activity Recognition System with Open Datasets. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12(13):3143–3147. Available from: https://doi.org/10.17762/turcomat.v12i13.9093.
4) Ullah A, Muhammad K, Ding W, Palade V, Haq IU, Baik SW. Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. *Applied Soft Computing*. 2021;103:1–13. Available from: https://dx.doi.org/10.1016/j.asoc.2021.107102.
5) Al-Saedi HH. Recognition of normal and abnormal human actions. Весенние дни науки. In: International Conference of Students and Young Scientists "Spring Days of Science". 2021;p. 333–335. Available from: http://elar.urfu.ru/handle/10995/99844.
6) Elharrouss O, Almaadeed N, Al-Maadeed S, Bouridane A, Beghdadi A. A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence*. 2021;51(2):690–712. Available from: https://dx.doi.org/10.1007/s10489-020-01823-z.
7) D DA, Minu RI. Edge computing based surveillance framework for real time activity recognition. *ICT Express*. 2021;7(2):182–186. Available from: https://dx.doi.org/10.1016/j.icte.2021.04.010.

8) Franco A, Magnani A, Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data. *Pattern Recognition Letters*. 2020;131:293–299. Available from: https://dx.doi.org/10.1016/j.patrec.2020.01.010.
9) Zhang Y, Po LM, Liu M, Rehman YAU, Ou W, Zhao Y. Data-level information enhancement: Motion-patch-based Siamese Convolutional Neural Networks for human activity recognition in videos. *Expert Systems with Applications*. 2020;147. Available from: https://dx.doi.org/10.1016/j.eswa.2020.113203.
10) Wan S, Qi L, Xu X, Tong C, Gu Z. Deep Learning Models for Real-time Human Activity Recognition with Smartphones. *Mobile Networks and Applications*. 2020;25(2):743–755. Available from: https://dx.doi.org/10.1007/s11036-019-01445-x.
11) Singh R, Sonawane A, Srivastava R. Recent evolution of modern datasets for human activity recognition: a deep survey. *Multimedia Systems*. 2020;26(2):83–106. Available from: https://dx.doi.org/10.1007/s00530-019-00635-7.
12) Shreyas DG, Raksha S, Prasad BG. Implementation of an Anomalous Human Activity Recognition System. *SN Computer Science*. 2020;1(3). Available from: https://dx.doi.org/10.1007/s42979-020-00169-0.
13) Dwivedi N, Singh DK, Kushwaha DS. Orientation Invariant Skeleton Feature (OISF): a new feature for Human Activity Recognition. *Multimedia Tools and Applications*. 2020;79(29-30):21037–21072. Available from: https://dx.doi.org/10.1007/s11042-020-08902-w.
14) Mukherjee S, Anvitha L, Lahari TM. Human activity recognition in RGB-D videos by dynamic images. *Multimedia Tools and Applications*. 2020;79(27-28):19787–19801. Available from: https://dx.doi.org/10.1007/s11042-020-08747-3.
15) Kwon H, Tong C, Haresamudram H, Gao Y, Abowd GD, Lane ND, et al. IMUTube: Automatic Extraction of Virtual on-body Accelerometry from Video for Human Activity Recognition. In: and others, editor. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies;vol. 4, Issue 3. Association for Computing Machinery (ACM). 2020;p. 1–29. Available from: https://dx.doi.org/10.1145/3411841. doi:10.1145/3411841.
16) Gul MA, Yousaf MH, Nawaz S, Rehman ZU, Kim H. Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture. *Electronics*. 2020;9(12):1–14. Available from: https://doi.org/10.3390/electronics9121993.
17) Ghazal S, Khan US, Saleem MM, Rashid N, Iqbal J. Human activity recognition using 2D skeleton data and supervised machine learning. *IET Image Processing*. 2019;13(13):2572–2578. Available from: https://dx.doi.org/10.1049/iet-ipr.2019.0030.
18) Naveed H, Khan G, Khan AU, Siddiqi A, Khan MUG. Human activity recognition using mixture of heterogeneous features and sequential minimal optimization. *International Journal of Machine Learning and Cybernetics*. 2019;10(9):2329–2340. Available from: https://dx.doi.org/10.1007/s13042-018-0870-1.
19) Zhou X, Liang W, Wang KIK, Wang H, Yang LT, Jin Q. Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things. *IEEE Internet of Things Journal*. 2020;7(7):6429–6438. Available from: https://dx.doi.org/10.1109/jiot.2020.2985082.
20) Ehatisham-Ul-Haq M, Javed A, Azam MA, Malik HMA, Irtaza A, Lee IH, et al. Robust Human Activity Recognition Using Multimodal Feature-Level Fusion. *IEEE Access*. 2019;7:60736–60751. Available from: https://dx.doi.org/10.1109/access.2019.2913393.
21) Kushwaha A, Khare A, Srivastava P. On integration of multiple features for human activity recognition in video sequences. *Multimedia Tools and Applications*. 2021;80(21-23):32511–32538. Available from: https://dx.doi.org/10.1007/s11042-021-11207-1.
22) Deotale D, Verma M, Suresh P. Human Activity Recognition in Untrimmed Video using Deep Learning for Sports Domain. *SSRN Electronic Journal*;p. 1–12. Available from: https://dx.doi.org/10.2139/ssrn.3769815.
23) Girdhar P, Johri P, Virmani D. Incept_LSTM : Accession for human activity concession in automatic surveillance. *Journal of Discrete Mathematical Sciences and Cryptography*. 2022;25(8):2259–2273. Available from: https://doi.org/10.1080/09720529.2020.1804132.
24) Kushwaha A, Khare A, Khare M. Human Activity Recognition Algorithm in Video Sequences Based on Integration of Magnitude and Orientation Information of Optical Flow. *International Journal of Image and Graphics*. 2022;22(01). Available from: https://dx.doi.org/10.1142/s0219467822500097.
25) Alawneh L, Alsarhan T, Al-Zinati M, Al-Ayyoub M, Jararweh Y, Lu H. Enhancing human activity recognition using deep learning and time series augmented data. *Journal of Ambient Intelligence and Humanized Computing*. 2021;12(12):10565–10580. Available from: https://dx.doi.org/10.1007/s12652-020-02865-4.
26) Santos F, Durães D, Marcondes F, Gomes M, Gonçalves F, Fonseca J, et al. Modelling a Deep Learning Framework for Recognition of Human Actions on Video. In: World Conference on Information Systems and Technologies;vol. 1365 of Advances in Intelligent Systems and Computing. Springer, Cham. 2021;p. 104–112. Available from: https://doi.org/10.1007/978-3-030-72657-7_10.
27) Marcondes FS, Durães D, Gonçalves F, Fonseca J, Machado J, Novais P. In-vehicle violence detection in carpooling: a brief survey towards a general surveillance system. InDistributed Computing and Artificial Intelligence. In: 17th International Symposium on Distributed Computing and Artificial Intelligence;vol. 1237 of Advances in Intelligent Systems and Computing. Springer, Cham. 2020;p. 211–220. Available from: https://doi.org/10.1007/978-3-030-53036-5_23.
28) Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2009;p. 248–255. Available from: https://doi.org/10.1109/CVPR.2009.5206848.
29) Hosotani D, Yoda I, Sakaue K. Wheelchair recognition by using stereo vision and histogram of oriented gradients (HOG) in real environments. In: 2009 Workshop on Applications of Computer Vision (WACV). IEEE. 2010;p. 1–6. Available from: https://doi.org/10.1109/WACV.2009.5403043.
30) Vishwakarma DK, Rawat P, Kapoor R. Human Activity Recognition Using Gabor Wavelet Transform and Ridgelet Transform. *Procedia Computer Science*. 2015;57:630–636. Available from: https://dx.doi.org/10.1016/j.procs.2015.07.425.
31) Hazra D. Retrieval of color image using color correlogram and wavelet filters. In: International Conference on advances in computer engineering. 2011;p. 151–154. Available from: https://www.scribd.com/document/336767426/123.
32) Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973;SMC-3(6):610–621. Available from: https://dx.doi.org/10.1109/tsmc.1973.4309314.
33) Kumari B, Kumar R, Singh VK, Pawar L, Pandey P, Sharma M. An Efficient System for Color Image Retrieval Representing Semantic Information to Enhance Performance by Optimizing Feature Extraction. *Procedia Computer Science*. 2019;152:102–110. Available from: https://dx.doi.org/10.1016/j.procs.2019.05.032.
34) Zhang H, Li Y, Jiang Y, Wang P, Shen Q, Shen C. Hyperspectral Classification Based on Lightweight 3-D-CNN With Transfer Learning. *IEEE Transactions on Geoscience and Remote Sensing*. 2019;57(8):5813–5828. Available from: https://dx.doi.org/10.1109/tgrs.2019.2902568.
35) Chui KT, Gupta BB, Chi HR, Arya V, Alhalabi W, Ruiz MT, et al. Transfer Learning-Based Multi-Scale Denoising Convolutional Neural Network for Prostate Cancer Detection. *Cancers*. 2022;14(15):1–13. Available from: https://dx.doi.org/10.3390/cancers14153687.
36) Hakim M, Omran AAB, Inayat-Hussain JI, Ahmed AN, Abdellatef H, Abdellatif A, et al. Bearing Fault Diagnosis Using Lightweight and Robust One-Dimensional Convolution Neural Network in the Frequency Domain. *Sensors*. 2022;22(15):1–24. Available from: https://dx.doi.org/10.3390/s22155793.
37) Chathuramali KGM, Rodrigo R. Faster human activity recognition with SVM. In: International conference on advances in ICT for emerging regions. IEEE. 2013;p. 197–203. Available from: https://doi.org/10.1109/ICTer.2012.6421415.

38) Soomro K, Zamir AR, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. 2012. Available from: https://doi.org/10.48550/arXiv.1212.0402.
39) Xia L, Chen CCC, Aggarwal JK. View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE. 2012;p. 20–27. Available from: https://doi.org/10.1109/CVPRW.2012.6239233.
40) Xie Y. Deep Learning Approaches for Human Action Recognition in Video Data. . Available from: https://doi.org/10.48550/arXiv.2403.06810.
41) Ning Z, Suk-Hwan L, Eung-Joo L. Human Activity Recognition Based on Loss-Net Fusion Domain Convolutional Neural Networks. In: 2019 IEEE International Conference on Computation, Communication and Engineering (ICCCE). IEEE. 2020;p. 146–149. Available from: https://doi.org/10.1109/ICCCE48422.2019.9010800.
42) Jaouedi N, Boujnah N, Bouhlel MS. A new hybrid deep learning model for human action recognition. *Journal of King Saud University - Computer and Information Sciences*. 2020;32(4):447–453. Available from: https://dx.doi.org/10.1016/j.jksuci.2019.09.004.
43) Dong M, Fang Z, Li Y, Bi S, Chen J. AR3D: Attention Residual 3D Network for Human Action Recognition. *Sensors*. 2021;21(5):1–14. Available from: https://dx.doi.org/10.3390/s21051656.
44) Li W, Xu N, Liu G, Zhao L, Fang X. Segments-Based 3D ConvNet for Action Recognition. In: 2020 International Conference on Computer Science and Communication Technology (ICCSCT) 2020 ;vol. 1621 of Journal of Physics: Conference Series. IOP Publishing. 2020;p. 1–7. Available from: https://dx.doi.org/10.1088/1742-6596/1621/1/012042.
45) Richard A, Gall J. A bag-of-words equivalent recurrent neural network for action recognition. *Computer Vision and Image Understanding*. 2017;156:79–91. Available from: https://dx.doi.org/10.1016/j.cviu.2016.10.014.
46) Yu S, Xie L, Liu L, Xia D. Learning Long-Term Temporal Features With Deep Neural Networks for Human Action Recognition. *IEEE Access*. 2019;8:1840–1850. Available from: https://dx.doi.org/10.1109/access.2019.2962284.
47) Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access*. 2017;6:1155–1166. Available from: https://dx.doi.org/10.1109/access.2017.2778011.
48) Ahmad T, Wu J, Alwageed HS, Khan F, Khan J, Lee Y. Human Activity Recognition Based on Deep-Temporal Learning Using Convolution Neural Networks Features and Bidirectional Gated Recurrent Unit With Features Selection. *IEEE Access*. 2023;11:33148–33159. Available from: https://dx.doi.org/10.1109/access.2023.3263155.
49) Ghodsi S, Mohammadzade H, Korki E. Simultaneous joint and object trajectory templates for human activity recognition from 3-D data. *Journal of Visual Communication and Image Representation*. 2018;55:729–741. Available from: https://dx.doi.org/10.1016/j.jvcir.2018.08.001.
50) Gao X, Hu W, Tang J, Liu J, Guo Z. Optimized Skeleton-based Action Recognition via Sparsified Graph Regression. In: Proceedings of the 27th ACM International Conference on Multimedia. ACM. 2019;p. 601–610. Available from: https://doi.org/10.1145/3343031.3351170.
51) Rhif M, Wannous H, Farah IR. Action Recognition from 3D Skeleton Sequences using Deep Networks on Lie Group Features. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE. 2018;p. 3427–3432. Available from: https://doi.org/10.1109/ICPR.2018.8546027.
52) Mohammadzade H, Hosseini S, Rezaei-Dastjerdehei MR, Tabejamaat M. Dynamic Time Warping-Based Features With Class-Specific Joint Importance Maps for Action Recognition Using Kinect Depth Sensor. *IEEE Sensors Journal*. 2021;21(7):9300–9313. Available from: https://dx.doi.org/10.1109/jsen.2021.3051497.
53) Zhang S, Liu X, Xiao J. On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE. 2017;p. 148–157. Available from: https://doi.org/10.1109/WACV.2017.24.
54) Zhu K, Wang R, Zhao Q, Cheng J, Tao D. A Cuboid CNN Model With an Attention Mechanism for Skeleton-Based Action Recognition. *IEEE Transactions on Multimedia*. 2020;22(11):2977–2989. Available from: https://dx.doi.org/10.1109/tmm.2019.2962304.
55) Kao JY, Ortega A, Tian D, Mansour H, Vetro A. Graph Based Skeleton Modeling for Human Activity Analysis. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE. 2019;p. 2025–2029. Available from: https://doi.org/10.1109/ICIP.2019.8803186.
56) Anirudh R, Turaga P, Su J, Srivastava A. Elastic functional coding of human actions: From vector-fields to latent variables. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2015;p. 3147–3155. Available from: https://doi.org/10.1109/CVPR.2015.7298934.
57) Liu G, Zhang Q, Cao Y, Tian G, Ji Z. Online human action recognition with spatial and temporal skeleton features using a distributed camera network. *International Journal of Intelligent Systems*. 2021;36(12):7389–7411. Available from: https://dx.doi.org/10.1002/int.22591.