Computers, Materials & Continua
DOI: 10.32604/cmc.2023.028743
Article

Tech Science Press

Two-Stream Deep Learning Architecture-Based Human Action Recognition

Faheem Shehzad[1], Muhammad Attique Khan[2], Muhammad Asfand E. Yar[3], Muhammad Sharif[1], Majed Alhaisoni[4], Usman Tariq[5], Arnab Majumdar[6] and Orawit Thinnukool[7,*]

[1]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Pakistan
[2]Department of Computer Science, HITEC University, Taxila, Pakistan
[3]Department of Computer Science, Bahria University, Islamabad, Pakistan
[4]Computer Sciences Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11671, Saudi Arabia
[5]College of Computer Engineering and Science, Prince Sattam Bin Abdulaziz University, Al-Kharaj, 11942, Saudi Arabia
[6]Faculty of Engineering, Imperial College London, London, SW7 2AZ, UK
[7]College of Arts, Media, and Technology, Chiang Mai University, Chiang Mai, 50200, Thailand
*Corresponding Author: Orawit Thinnukool. Email: orawit.t@cmu.ac.th

Received: 16 February 2022; Accepted: 06 May 2022

Abstract: Human action recognition (HAR) based on Artificial intelligence reasoning is the most important research area in computer vision. Big breakthroughs in this field have been observed in the last few years; additionally, the interest in research in this field is evolving, such as understanding of actions and scenes, studying human joints, and human posture recognition. Many HAR techniques are introduced in the literature. Nonetheless, the challenge of redundant and irrelevant features reduces recognition accuracy. They also faced a few other challenges, such as differing perspectives, environmental conditions, and temporal variations, among others. In this work, a deep learning and improved whale optimization algorithm based framework is proposed for HAR. The proposed framework consists of a few core stages i.e., frames initial preprocessing, fine-tuned pre-trained deep learning models through transfer learning (TL), features fusion using modified serial based approach, and improved whale optimization based best features selection for final classification. Two pre-trained deep learning models such as InceptionV3 and Resnet101 are fine-tuned and TL is employed to train on action recognition datasets. The fusion process increases the length of feature vectors; therefore, improved whale optimization algorithm is proposed and selects the best features. The best selected features are finally classified using machine learning (ML) classifiers. Four publicly accessible datasets such as Ut-interaction, Hollywood, **Free Viewpoint Action Recognition using Motion History Volumes (IXMAS)**, and centre of computer vision (UCF) Sports, are employed and achieved the testing accuracy of 100%, 99.9%, 99.1%, and 100% respectively. Comparison with **state of the art techniques** (SOTA), the proposed method showed the improved accuracy.

Keywords: Human action recognition; deep learning; transfer learning; fusion of multiple features; features optimization

1 Introduction

Human Action Recognition (HAR) is a critical research topic in machine learning and computer vision applications [1]. Because of its covariant properties, HAR has gained a lot of popularity in recent decades. Real-world applications of HAR include robotics, location estimation, sports analysis, pedestrian detection, human-computer interaction, video games, and video surveillance [2]. Several human actions such as pointing, running, pushing, boxing, kicking, hand waving, jogging, clapping, diving, and named a few more are recognized in the video sequences (a few samples shown in Fig. 1). These actions are recognized through computerized systems automatically and effectively [3]. Nowadays, many HAR methods are used like wireless network-based method, video-based method, and sensor-based method [4]. However, video-based HAR approaches are gaining popularity due to their high recognition rate and ease of usage. Furthermore, HAR is broadly used in different industrial applications [5]. During the past few years, the big breakthroughs have been witnessed in this field. Also, the research interest in this field is evolving like understanding of actions and scenes, studying the human joints, and human posture recognition. The precision of HAR has been raised due to the growing of learning-based **artificial intelligence** (AI) [6]. Even though various innovations are being witnessed in AI technology, but still there exist quite a few challenges also in this field. For their learning-based algorithms, this field requires large datasets and corresponding labels. Several videos are obtained from the YouTube platform and manually fine-tuned in terms of detail, actions, and comprehension. The manual recognition and understanding process is time-consuming and labor-intensive [7].

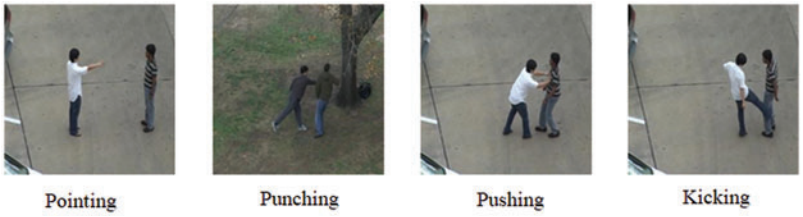


Figure 1: Sample human actions frames collected from UT-Interaction dataset [8]

Activity recognition in video sequences is a moving issue because of the comparability of visual substance, changes in the perspective for similar activities [9], camera movement with activity entertainer, posture and scale of an entertainer, and diverse enlightenment situations [10]. Human activities range from simple leg or arm movement to complex coordinated movement of consolidated legs, arms, and body. For example, kicking a football is a basic activity, whereas hopping for a top shoot is an aggregate movement of arms, legs, head, and entire body [11]. For many reasons, correctly recognition of human actions in video frames remains a difficult process, like having inter-class and intra-class variation, lightning, environmental and angle variation, etc. [12]. To deal with these issues handcrafted methods for feature extraction like histogram optical flow and histogram oriented gradient are used in previous research studies [13]. Because missing of a 3-dimensional (3D) structure in the video sequence, these methods are unable to recognize actions using 2D data [14].

Recently, the deep learning shows the much performance in the area of computer vision and machine learning for several applications such as medical [15], biometric [16], video surveillance, agriculture [17], and object classification [18]. From those, HAR is active research and many researchers improve the performance through deep learning techniques [19]. Convolutional neural network (CNN) is form of deep learning, being used to improve the rate of HAR [20]. Generally, HAR methods are based on two steps, i.e., features extraction and classification [21]. Different CNN pre-trained models like AlexNet [22], very very deep (VGG), and ResNet, and named a few others [23] are used with the transfer learning concept for HAR. These techniques give improved accuracy than traditional feature extraction techniques. But some time, due to complex nature of dataset, a single CNN model not performed well; therefore, information fusion of more than one model can be employed. The fusion process increases the computational time due to more number of predictors [24]. Hence, feature selection techniques are more suitable. The selection techniques minimize the volume of data to save the cost of modeling and, in some conditions, improve the functioning of the algorithm [9]. Finally, the final features are passed to the different classifiers for classification purposes. Different classifiers such as multiclass Support Vector Machine (M-SVM) [25], K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Complex tree (CT), and Ada-boost are used for action classification [26]. Recently, the researchers introduced many techniques for human action recognition (HAR) [27]. Those techniques are based on a few well-known steps such as preprocessing of original video frames, region of interest detection (ROI) for more accurate features extraction, and finally recognition using machine learning classifiers [28].

However, they continue to face a slew of issues that degrade overall accuracy and lengthen computational time. The major challenges are as follows: (i) imbalanced datasets increase the prediction probability of the maximum class; (ii) feature extraction from the last layer does not correctly visualize the original human action due to a lack of the required number of features; and (iii) redundant and irrelevant features reduce system recognition accuracy. Furthermore, the presence of these features increases the computational time of the system. In this work, we proposed a new framework for HAR based on deep learning features fusion and improve whale optimization algorithm. Our major contributions are listed as follows:

- Two pre-trained CNN models are fine-tuned, and new dense layers are added. The fine-tuned models were then trained on action datasets to extract features from a combination of layers (convolutional and fully connected) rather than a single target layer.
- Using a modified correlation extended serial approach, the extracted features of both fine-tuned models are fused.
- Based on the update criteria for the best feature selection, an improved whale optimization algorithm is introduced. Machine learning classifiers are used to classify the selected features.

The rest of the article is organized in the following order. Section 2 discussed the recent related work of HAR. Proposed HAR framework is discussed in Section 3. In this section the entire framework is described in the mathematical and visual manner. Results of the proposed HAR framework are presented in Section 4. The conclusion of this article is presented in Section 5.

2 Related Work

Many HAR techniques have been proposed in the literature based on deep learning and traditional features. Liu et al. [29] introduced a two-stream deep neural network for HAR. This network recognizes unusual behavior of human in the video sequences. **Volumetric motion history images** (VMHI) and original frames are the two main parts of this model and tested on **Royal Institute**

of Technology (KTH), Weizmann, and Ut-interaction datasets and showed improved accuracies. Chenarlogh et al. [30] introduced three different CNN architectures to optimize the performance of HAR in limited data. Three architectures of this model include 1-stream, 2-stream, and 4-stream. They tested their architecture on the IXMAS dataset and attained average accuracy of 88.05% on 4-stream architecture. Sharif et al. [31] suggested an approach to overcome the problem of the robust feature selection method. In HAR, extracting the prominent and salient features inside a video frame is a challenging job. The suggested method initially fuses three different feature categories and selects the most optimized features using strong correlation and Euclidean distance methods. Finally, classification is performed by a multi-class classifier. They used KTH, **large human motion database** (HMDB51), UCF YouTube, and Weizmann datasets and shows more than 94% classification accuracy. Jaouedi et al. [32] introduced a hybrid deep network for HAR to overcome the problem of detecting a moving person from a scene and detecting human motion from a background. The suggested method is tested by using KTH, UCF101, and UCF sports datasets and attained an average accuracy of 96.3%. Sharif et al. [33] suggested a novel HAR technique by using the combination of handcrafted and deep features. Initially, saliency-based method was employed for human silhouette extraction. Afterward, deep and handcrafted features are extracted and combined to make a final vector. The main purpose of features fusion is getting the maximum information of human actions for accurate classification. They tested their technique using UCF11 (YouTube), UT-interaction, IXMAS, Weizmann, and UCF sports datasets and attained better accuracy than SOTA.

Abdelbaky et al. [34] presented an architecture PCANet TOP for feature extraction and action classification based on SVM classifier. They tested their method using UCF Sports, KTH, YouTube action, and Weizmann datasets and attained an accuracy of 92.67%. Afza et al. [35] suggested a technique that fused traditional features and later selected the best of them for final classification. M-SVM classifier is used for action identification and achieved above 95% accuracy on four datasets-UCF YouTube, UCF Sports, Weizmann, and KTH. Abdelbaky et al. [36] presented a simple Neural network based on (PCA) network to minimize the issues related to real-time recognition systems and 3-dimensional signals in a video frame. This scheme uses an unsupervised learning approach instead of supervised learning approach. Sahoo et al. [37] suggested the HAR-Depth technique with shape learning and sequential learning streams combined with **depth history image** (DHI). The presented method is used to get maximum data from the action videos to overcome the error rate of correct recognition. Muhammad et al. [38] suggested a **Bi-Long shorter memory (BiLSTM)** based HAR approach using Dilated Convolutional Neural Network. This approach gives better performance in video surveillance for security needs. The HAR sequential process was followed by the aforementioned techniques. They used CNN architectures to extract features but skipped the preprocessing and optimization steps. The difference between the above studies is the long computational time and redundant features that can be addressed by these two steps.

3 Proposed Methodology

The proposed HAR architecture is presenting in Fig. 2. The proposed framework includes several steps such as: (i) frames initial preprocessing (ii) fine-tuned two pre-trained deep learning models such as Inceptionv3 and Resnet101 and extract deep features (iii) fusion of deep learning features using modified correlation extended serial approach (iv) best features selection using improved whale optimization algorithm, and (v) classification using machine learning algorithms and compute results. The detail of each step is given in below subsections.

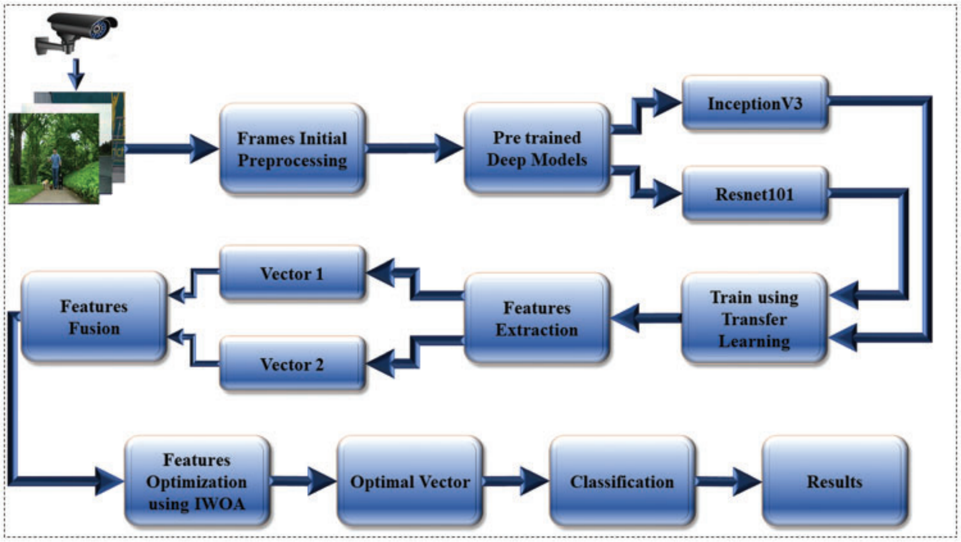


Figure 2: Proposed deep learning based framework for HAR

3.1 Video Frames Preprocessing

Pre-processing is the most important steps in image processing, with applications in different fields like agriculture, medicine, and surveillance, to mention a few [39]. Pre-processing is critical in surveillance to deal with light changes, complicated backgrounds, noise reduction, and other issues. In this work, the pre-processing step is employed to convert the action video sequences into frames. Originally, the each extracted video frame having dimension $512 \times 512 \times k$, where $k = 3$. The converted frames are later resized into a size $256 \times 256 \times 3$. These extracted video frames are converted into relevant classes and later utilized for the training of CNN models.

3.2 Convolutional Neural Network

CNN is a neural network with a convolution operation in at least one of its layers instead of matrix multiplications. CNN networks are now being used to improve the recognition rate of HAR. In a CNN, three basic layers are used: convolutional, pooling, and fully connected. In convolutional layer, different filters are applied to the image with different parameters for feature extraction. The basic parameters are size of kernel and the number of kernels. Mathematically, the convolution operation is formulated as follows:

$$H[a \times b] = (g * i)[a \times b] = \sum\nolimits_{k} \sum\nolimits_{j} i[k, j] g[a - k, b - j] \tag{1}$$

where input image is denoted by g, kernel by i and $a \times b$ shows the row and column of the resultant matrix. The $*$ represent the convolutional operator and H represent the output of convolution operation. After each convolutional layer, a ReLu activation layer is added to remove the negative features and place with zero.

$$ReLu = Max(0, x), x \in H \tag{2}$$

Pooling layer is used to decrease the size of the tensor to increase the calculation speed. In pooling layers, specific function is performed like max operation and average operation. Max pooling layer is used get a maximum value from each filter region and average pooling layer is utilized to get an AVG value in the each filter region. Another layer named fully connected layer, is employed to smooth the result before classification placed to output layer of a neural network. Mathematically, the FC layer is formulated as follows:

$$F_0^{out} = H\left[a \times b\right] \tag{3}$$

$$F_i^{in} = F_{i-1}^{out} * H_i + b_i \tag{4}$$

$$F_i^{out} = \Delta_i\left(F_i^{in}\right) \tag{5}$$

where, F_i^{out} is final FC layer, Δ represent activation function, and i is layer number. After the FC layer, the Softmax classification layer is added for features classification.

$$Softmax\left(F_i^{out}\right) = \frac{exp\left(F_i^{out}\right)}{\sum_j F_j^{out}} \tag{6}$$

3.3 Deep Learning Features

In this work, two pre-trained CNN models namely Inceptionv3 [40] and Resnet101 [41] are utilized for features extraction. Inception V3 CNN consists of 01 input layer, 94 convolutional layers, 01 fully connected layer, and 04 MaxPooling layer. Total number of layers in this network are 315. This network accepts input image of size 229 × 299 × 3. ResNet101 CNN model comprises of 01 input layer, 105 convolutional layers, 01 fully connected layer, and 01 MaxPooling layer. The total numbers of layers in this network are 347. This network accepts the input image of size 224 × 224 × 3. Initially both models were trained on ImageNet dataset which have 1000 object classes. Therefore, we fine-tuned both models and removed the last layers (fully connected layer-Classification layer) and added new dense layers and trained on action datasets using transfer learning. We considered the 70% video frames for training and rest 30% for testing purposes. The transfer learning (TL) concepts opted for training of fine-tuned models. Since the pre-trained nets are trained on selective classes (i.e., ImageNet dataset) but in our case, the target task is action recognition. Therefore, we need to train the network on selected action dataset. In the case of InceptionV3 CNN model, the last three layers such as 'predictions', 'predictions_softmax', and 'ClassificationLayer_predictions' are replaced with 'new_fc', 'predictions_softmax', and 'new_classoutput' layers. In the case of Resnet101 the last three layers such as 'fc1000', 'prob', and 'ClassificationLayer_predictions' are replaced with 'new_fc', 'prob', and 'new_classoutput' layers. The hyper parameters are initialized such as mini batch size of 16, initial learning rate is 0.05, epochs 200, and dropout factor is 0.5. Then, the newly fine-tuned models are trained through TL. Visually, the process of TL is illustrated in Fig. 3.

Inception V3 Features: We use the avg_pool layer of fine-tuned Inception V3 CNN model and applied activation for features extraction. On this layer, a feature vector is obtained of dimension $N \times 2048$ and represented with V_1.

ResNet101 Features: We employed pool5 layer of fine-tuned ResNet101 model and applied activation function for feature extraction. On this layer, a feature vector of dimension $N \times 2048$ is obtained and represented with V_2.

3.4 Deep Features Fusion

Features fusion is the process of combined multi-level information in one vector for better recognition accuracy. In this work, we fused two deep extracted feature vectors V_1 and V_2 using a new approach named modified correlation extended serial approach. Consider, $V_1 \in M_i$ and $V_2 \in N_j$, then the correlation is find out among i and j based on the following formula:

$$Cor = \frac{\sum \left(M_i - \overline{M}\right)\left(N_j - \overline{N}\right)}{\sqrt{\sum \left(M_i - \overline{M}\right)^2 \sum \left(N_j - \overline{N}\right)^2}} \tag{7}$$

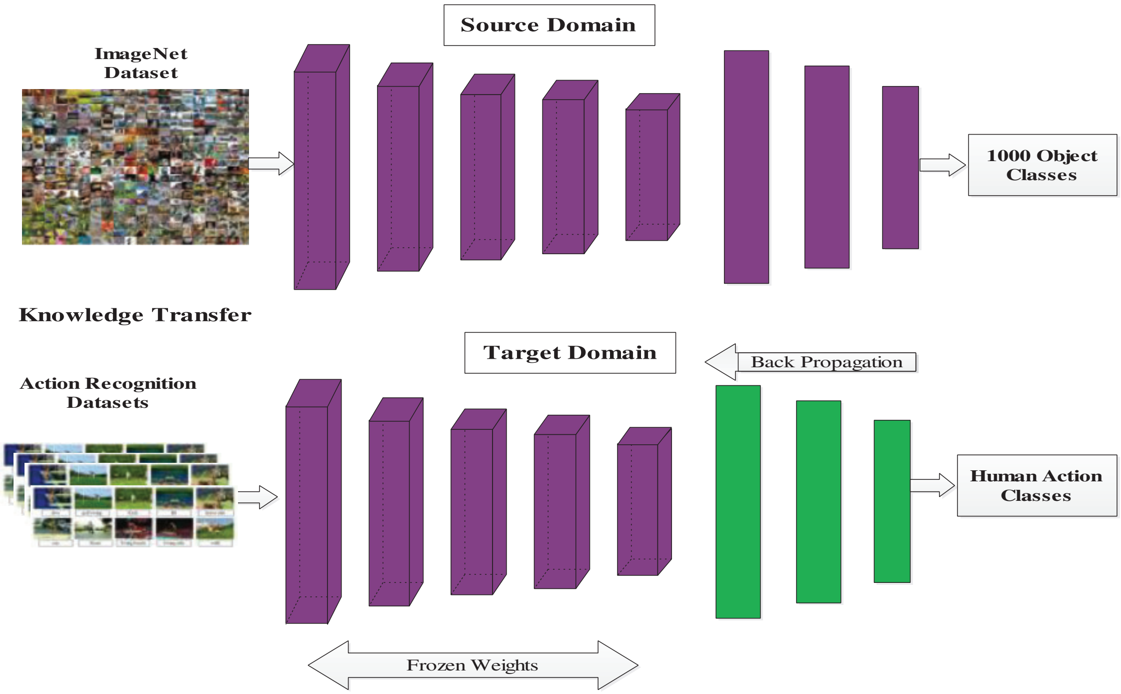


Figure 3: Process of TL for HAR

Based on this formula, the features that have positive correlation (+1) are selected in a new vector denoted by V_3 and features that have correlation value 0 or -1, are added in V_4. After that, the mean value is computed of V_4 and compared each feature with that as follows:

$$CT = \begin{cases} \tilde{V}_4 & for\ V_4 \geq \mu \\ Ignore, & Otherwise \end{cases} \tag{8}$$

The updated vector $\tilde{V}_4$ and V_3 are finally fused in one vector based on the following mathematical equation:

$$\mathrm{V}_5(\mathrm{k}) = \begin{pmatrix} V_3(\mathrm{k})_{\mathrm{m\times n}} \\ \tilde{V}_4(\mathrm{k})_{\mathrm{m\times n}} \end{pmatrix} \tag{9}$$

The resultant feature vector obtained of dimension $N \times \mathrm{V}_5$, where the seize of V_5 is 2205 in this work that further optimized using improved whale optimization algorithm.

3.5 Deep Features Optimization

For feature selection, we used an improved whale optimization algorithm (IWOA). The fused vector $\mathrm{V}_5(\mathrm{k})$ is given to IWOA as input and the algorithm returns the optimized feature vector as output. The working of optimization process is given below.

Whale optimization algorithm (WOA) is a metaheuristic algorithm, presented by first Mirjalili in 2016 [42]. There are three basic steps performed in this algorithm namely encircling prey, spiral updating position, and random search for prey.

Encircling prey: The humpback whale will encircle the prey once the location of the prey has been established. The encircling prey mechanism of whale is formulated by Eqs. (10) and (11).

$$E = |DY * (i) - Y(i)| \tag{10}$$

$$Y(i+1) = Y * (i) - BE \tag{11}$$

where i is the current number of iterations; $Y * (i)$ denotes the best whale position vector by a long shot; $Y(i)$ denotes the current whale position vector; B and D denote the vector coefficient and are calculated by Eqs. (12) and (13).

$$B = 2bs_1 - b \tag{12}$$

$$D = 2s_2 \tag{13}$$

where s_1 and s_2 denote the casual numbers $(0, 1)$; b denotes a convergent factor and linearly decreased from 2 to 0; b is calculated by Eq. (14).

$$b = 2 - 2\frac{I}{I_{max}} \tag{14}$$

where I denotes the current number of iterations and I_{max} denotes the maximum iterations.

Updating Spiral position: Because humpback whales swim in a circle toward their prey, therefore, the circular position updating is done through the following equation.

$$Y(i+1) = Y * (i) - E_p e^{bl} \cos(2\pi l) \tag{15}$$

where $E_p = |Y * (i) - Y(i)|$ shows the distance between the prey and whale; b represents the constant and l stands for a unintended number from $(0, 1)$. Noticing that, while the whale go swimming in a curved toward its food, it also has to contract to envelop it. Therefore, the encircling prey method is selected by the probability P_t and the circular model is chosen by $1 - P_t$. Eq. (16) illustrates the calculating procedure:

$$Y(i+1) = \begin{cases} Y * (i) - BE & p < P_t \\ Y * (i) - E_p e^{bl} \cos(2\pi l) & p \geq P_t \end{cases} \tag{16}$$

It is set on the statistical method to attack prey and become close to prey in order to minimize the value of b, so that $B's$ range likewise fell with b in the iteration progression. When the value of b falls from 2 to 0, B is said to be within a random value [−b, b]. Furthermore, when the value of A is [−1, 1], the whale's next place could be right now or anywhere else between its prey. The whale attacks its victim when B is smaller than 1. While swimming along with the spiral pattern, the humpback whale surrounds its prey. To replicate the whale's hunting behavior, the probability of the encircling prey mechanism and curl position revise is set to 0.3.

Arbitrary search for prey: When a whale goes on randomly searching for prey, it must vary its position by going on a random search. The positions are computed as follows:

$$E = |DY_{rand} - Y(i)| \tag{17}$$

$$Y(i+1) = Y_{rand} - BE \tag{18}$$

where Y_{rand} denotes the casually choosing the whale's position vector. When $B \geq 1$, a seeking agent will refresh the positions of all other whales to the searching whale, forcing it to flee the target in order to locate better feed. In this approach, the exploration ability of the algorithm may be improved, allowing WOA to be searched from all angles.

In this work, we update two primary parameters B and D. This algorithm can balance exploitation and exploration because to its B set. As a result, the likelihood of a locally optimal increases. The B and D parameters in WOA are set to 0.4 and 0.5, respectively, which is clearly unnecessary. Meanwhile, WOA's capacity to search for all-around optimization has to be improved. The revised definition is defined as follows:

$$P_{i+1} = \begin{cases} P_0 & i = 1 \\ P_i * b + P_{min} & i > 1 \end{cases} \tag{19}$$

$$P_i' = 1 - P_i \tag{20}$$

$$B = exp\left[-30x\left(\frac{i}{I_{max}}\right)^S\right] \tag{21}$$

where P_0 is the initial probability of adaptive search enclosing mechanism; P_i' and P_{i+1} are the probability of encircling prey mechanism of *ith* and $(i+1)th$ generation; P_{min} shows the probability of minimum enveloping; P_i' denotes the probability of updated helix position of ith generation; i shows the iterations and I_{max} denotes the maximum iteration and $S = 2$.

The jumping behavior is also changed when the whale attempts to divide the region the value of local optimal can drop into minimum value by randomly updating the whale's location. The jumping behavior is defined as follows:

$$Y_{t(i+1)} = Y_t(i) + g(1 - 2rand)(max(Y_{all}) - min(Y_{all}))/2 \tag{22}$$

where g denotes jumping coefficient and Y_{all} denotes all whales. The neural network is employed as a fitness function and the fitness is calculated based on the mean square error rate (MSER). The best selected features are passed to several machine learning classifiers for the final action recognition.

4 Experimental Results and Discussion

This section comprises a full discussion of the results and analyses. The proposed framework is tested on four different datasets namely, (i) Ut-interaction, (ii) UCF Sports, (iii) Hollywood, and (iv) IXMAS. Several hyperparameters are employed for the training of pre-trained models such as learning rate is 0.05, mini batch size is 16, epochs are 200, and optimizer is stochastic gradient descent. The 10 fold cross-validation is opted on all 4 datasets, where the training and testing ratio was 50:50. Eight different classifiers, including Fine **K-nearest neighbour** (KNN), Ensemble Subspace KNN, Cubic SVM, Weighted KNN, Linear SVM, Cosine KNN, Quadratic SVM, Medium KNN, and Ensemble Bagged Trees are utilized for the classification results. The proposed framework is implemented in MATLAB 2020a, using personal computer having specification, Core i7 with 16 GB of DDR4 RAM and 16GB graphics card.

4.1 Numerical Results

The proposed framework results are presenting here in the form of tabular and confusion matrixes. The results are presented here for each dataset separately.

UT-Interaction Dataset Results: The results of UT-Interaction dataset are presented in Tab. 1. The **Ensemble Subspace KNN** (ESKNN) classifier attained the highest accuracy of 100% and other parameters like precision, recall, and F1 score values are 1.0, 1.0, and 1.0, respectively. The rest of the classifiers also achieved better results of >99%. Fig. 4 illustrated the confusion matrix of ESKNN classifier. Through this figure, the computed performance measures can be verified. The computational time of each classifier is also noted and the minimum testing time is 58.508 (s) of Ensemble Baggage Tree.

Table 1: Classification results on UT-Interaction dataset using proposed framework

Classifiers	Parameters				
	Recall	Precision	F1 Score	Accuracy (%)	Time (s)
ES KNN	1.0	1.0	1.0	**100**	171
Fine KNN	1.0	1.0	1.0	99.9	121.9
Cubic SVM	1.0	1.0	1.0	99.9	100.08
Weighted KNN	0.9983	0.9983	0.9966	99.7	104.95
Cosine KNN	0.9937	0.99	0.9937	99.4	101.02
Quadratic SVM	1.0	1.0	1.0	99.9	92.417
Linear SVM	0.9983	0.9983	0.9966	99.7	80.563
Medium KNN	0.9916	0.9961	0.9916	99.2	124.71
EBT	0.9816	0.98	0.9783	98.3	**58.508**

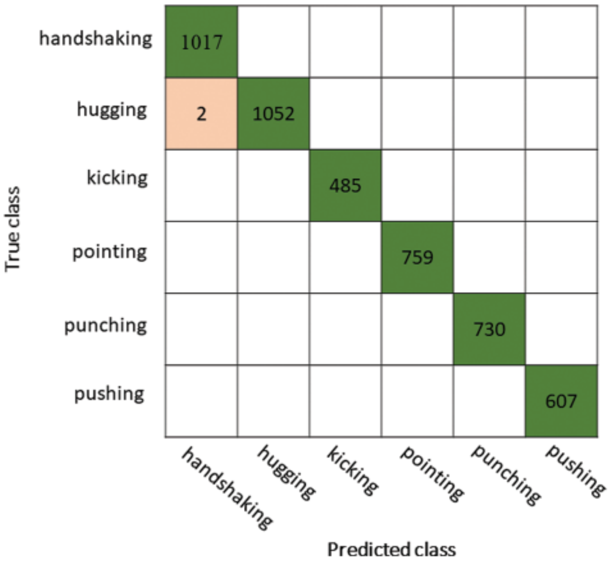


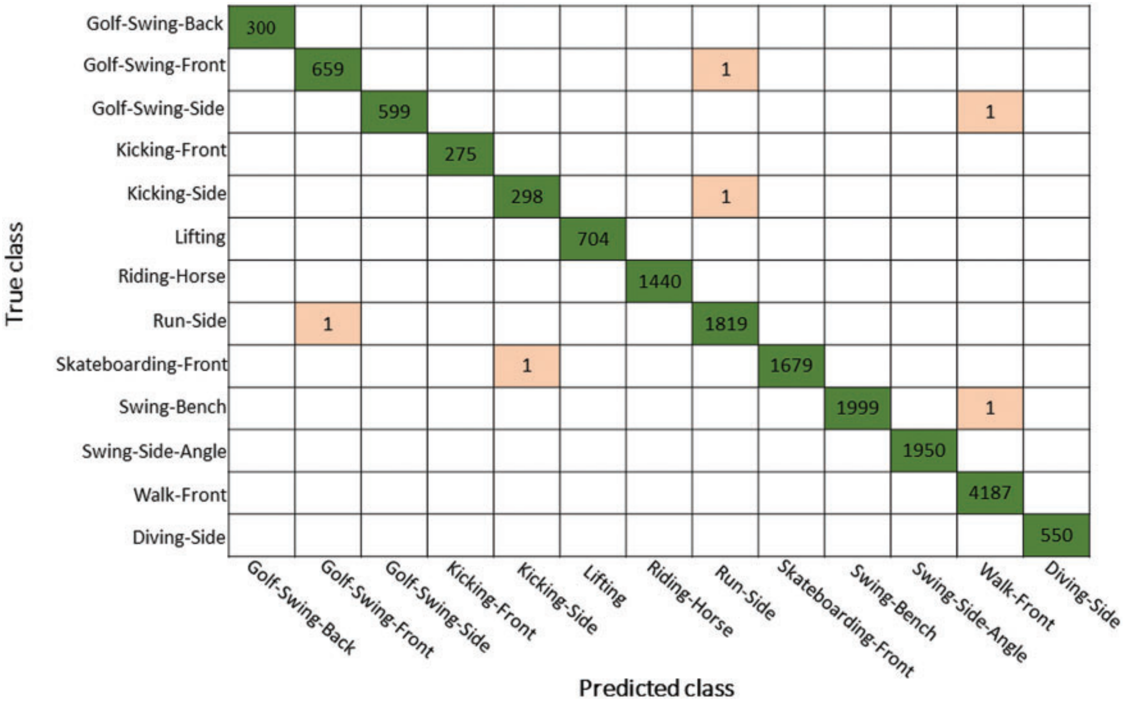
Figure 4: Subspace KNN classifier's confusion matrix on Ut-Interaction dataset using optimal features fusion

UCF Sports Dataset Results: Tab. 2 presents the recognition results of UCF Sports dataset using proposed framework. In this table, Quadratic SVM classifier attained the highest accuracy of 100% and other parameters like precision, recall, and F1 score values are 1.0, 1.0, and 1.0, respectively.

These values can be further verified through a confusion matrix given in Fig. 5. The other classifiers also give the better accuracy using proposed framework on selected dataset. The computational time is also noted for each classifier and minimum noted time is 107.02 (s) of Ensemble Baggage Tree (EBT). Similarly, the hollywood dataset results are presented in Tab. 3 and confusion matrix illustrated in Fig. 6.

Table 2: Classification results on UCF Sports dataset using proposed framework

Classifiers	Parameters				
	Recall	Precision	F1 Score	Accuracy (%)	Time (s)
ES KNN	1.0	1.0	1.0	100	241.5
Fine KNN	1.0	1.0	1.0	100	178.21
Cubic SVM	1.0	1.0	1.0	100	178.66
Weighted KNN	1.0	1.0	1.0	99.9	168.68
Cosine KNN	1.0	1.0	1.0	99.9	161.11
Quadratic SVM	**1.0**	**1.0**	**1.0**	**100**	174.86
Linear SVM	1.0	1.0	1.0	99.9	144.5
Medium KNN	1.0	1.0	1.0	99.9	171.76
EBT	0.9937	0.99	0.9925	99.3	**107.02**

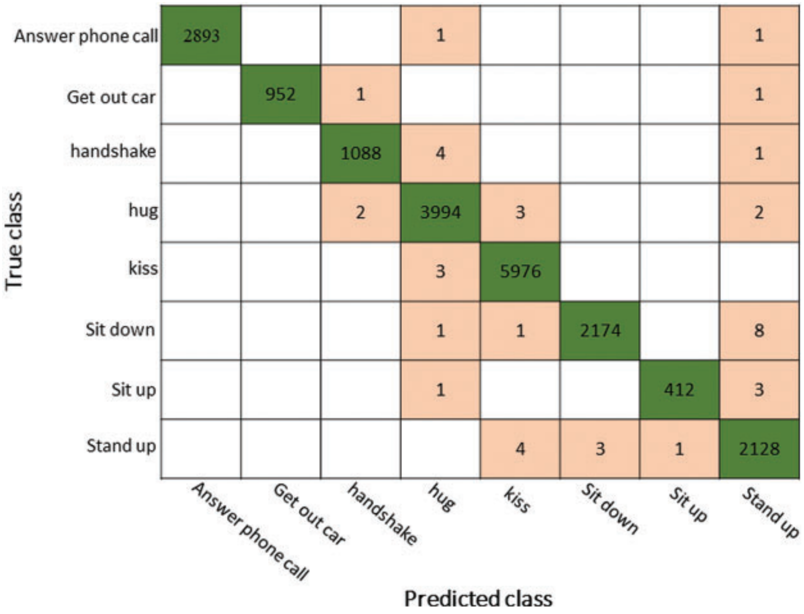


	Golf-Swing-Back	Golf-Swing-Front	Golf-Swing-Side	Kicking-Front	Kicking-Side	Lifting	Riding-Horse	Run-Side	Skateboarding-Front	Swing-Bench	Swing-Side-Angle	Walk-Front	Diving-Side
Golf-Swing-Back	300												
Golf-Swing-Front		659						1					
Golf-Swing-Side			599									1	
Kicking-Front				275									
Kicking-Side					298			1					
Lifting						704							
Riding-Horse							1440						
Run-Side		1						1819					
Skateboarding-Front					1				1679				
Swing-Bench										1999		1	
Swing-Side-Angle											1950		
Walk-Front												4187	
Diving-Side													550

Figure 5: Quadratic SVM classifier's confusion matrix on UCF Sports dataset using optimal features fusion

Table 3: Classification results on Hollywood dataset using proposed framework

Classifier	Parameters				
	Recall	Precision	F1 Score	Accuracy (%)	Time (s)
Ensemble Subspace KNN	1.0	1.0	1.0	**99.9**	139.4
Fine KNN	1.0	1.0	1.0	99.9	211.5
Cubic SVM	1.0	1.0	1.0	99.8	150.6
Weighted KNN	0.9983	0.9983	0.9966	99.7	210.5
Cosine KNN	0.9966	0.9958	0.9966	99.5	194.4
Quadratic SVM	0.9983	0.9983	0.9966	99.7	147.8
Linear SVM	0.9937	0.99	0.9925	99.3	278.3
Medium KNN	0.9937	0.99	0.9937	99.4	216.2
Ensemble Bagged Trees	0.9887	0.9812	0.9837	98.7	**120.8**



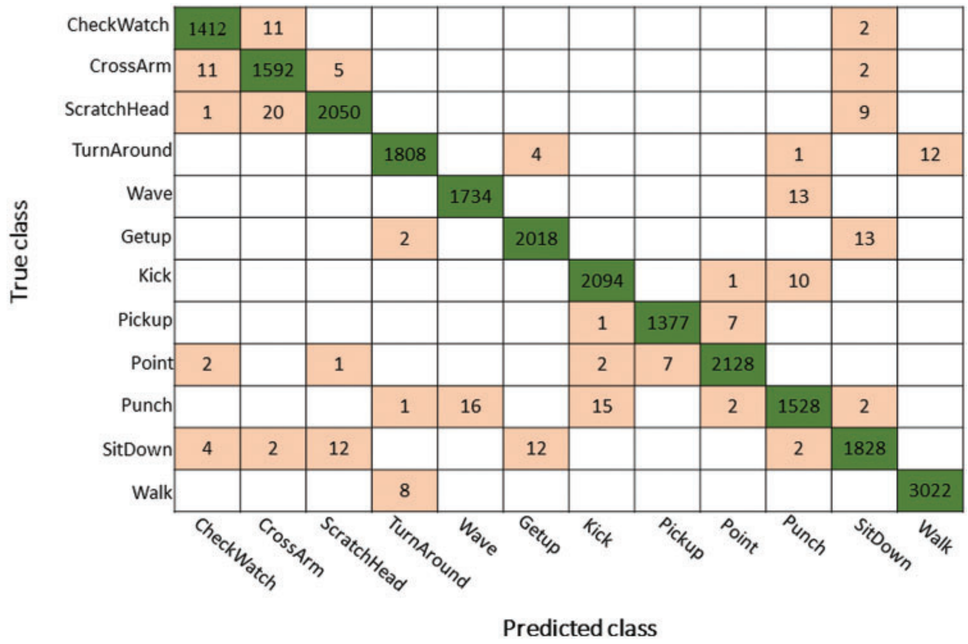
	Answer phone call	Get out car	handshake	hug	kiss	Sit down	Sit up	Stand up
Answer phone call	2893			1				1
Get out car		952	1					1
handshake			1088	4				1
hug			2	3994	3			2
kiss				3	5976			
Sit down				1	1	2174		8
Sit up				1			412	3
Stand up					4	3	1	2128

Figure 6: Fine KNN classifier's confusion matrix on Hollywood dataset using optimal features fusion

IXMAS Dataset Results: The results of IXMAS action dataset are given in Tab. 4 using proposed framework. This table shows the best accuracy is achieved by ESKNN classifier of 99.1% and other measures like precision, recall, and F1 score are 0.9916, 0.9916, and 0.99, respectively. These values can be further verified through a confusion matrix, illustrated in Fig. 7. Foe each classifier listed in this table, the computational time is also computed. The minimum noted time is 178.1 (s) for LSVM, whereas the EBT executed in 202.87 (s). Overall, the ESKNN classifier performs better than the rest on the classifier based on time and accuracy.

Table 4: Classification results on IXMAS dataset using proposed framework

Classifier	Parameters				
	Recall	Precision	F1 Score	Accuracy (%)	Time (s)
Ensemble Subspace KNN	0.9916	0.9916	0.99	**99.1**	492.8
Fine KNN	0.9909	0.9909	0.99	99.0	291.9
Cubic SVM	0.9866	0.985	0.985	98.7	339.8
Weighted KNN	0.9716	0.97	0.9716	97.5	295.5
Cosine KNN	0.9541	0.955	0.9558	95.8	242.2
Quadratic SVM	0.9825	0.9808	0.98	98.2	255.7
Linear SVM	0.9683	0.9633	0.965	96.8	**178.1**
Medium KNN	0.9612	0.966	0.9652	95.9	293.6
Ensemble Bagged Trees	0.9525	0.9533	0.9541	95.0	202.87



	CheckWatch	CrossArm	ScratchHead	TurnAround	Wave	Getup	Kick	Pickup	Point	Punch	SitDown	Walk
CheckWatch	1412	11									2	
CrossArm	11	1592	5								2	
ScratchHead	1	20	2050								9	
TurnAround				1808		4				1		12
Wave					1734					13		
Getup				2		2018					13	
Kick							2094		1	10		
Pickup							1	1377	7			
Point	2		1				2	7	2128			
Punch				1	16		15		2	1528	2	
SitDown	4	2	12			12				2	1828	
Walk				8								3022

Figure 7: KNN classifier's confusion matrix on IXMAS dataset using optimal features fusion

4.2 Discussion

A detailed discussion has been conducted in this section for the proposed framework based on the accuracy achieved by the selected datasets. The proposed framework is illustrated in Fig. 2 which consists of series of steps. The entire proposed framework results are given in Tabs. 1–4 and confusion matrixes in Figs. 4–7. Based on the tables and confusion matrixes, it is noted the proposed framework achieved maximum accuracy on selected datasets. However, it is essential to analyse the performance of middle steps such as original deep features extraction and best selected features for each CNN model.

The main purpose of employing classification results of middle steps is to analyse the importance of optimization algorithm. Another purpose of this analysis is to check the following question: if optimization algorithm is employed separately on deep extracted features then what will be the accuracy?

Tabs. 5–8 presents the accuracy of middle steps on selected action dataset. In these tables, it is noted that the accuracy is initially computed by using fine-tuned ResNet101 and Inception V3 models features. After that, the optimization algorithm is employed on original deep extracted features of ResNet101 (Best ResNet101) and Inception V3 (Best Inception V3). The proposed framework results are given in the last column for the sake of comparison. Based on accuracy values, given in these tables, it is noted that the optimization process improves the recognition accuracy but one the other end, proposed framework gives the better results.

Table 5: Comparison of overall proposed framework accuracy with middle steps on UT-Interaction dataset

Classifiers	ResNet101	Inception V3	Best ResNet101	Best Inception V3	Proposed
Ensemble Subspace KNN	95.2	96.5	98.8	96.6	**100**
Fine KNN	96.1	95.6	98.9	97.6	99.9
Cubic SVM	96.5	95.5	97.7	96.6	99.9
Weighted KNN	94.8	96.9	97.2	97.9	99.7
Cosine KNN	96	97.2	96.7	98.2	99.4
Quadratic SVM	97	96.2	98.6	98.5	99.9
Linear SVM	96.9	95.5	97.4	98.2	99.7
Medium KNN	97	94.1	96.5	98.1	99.2
Ensemble Bagged Trees	95.7	95.6	97.1	97.9	98.3

Table 6: Comparison of overall proposed framework accuracy with middle steps on UCF Sports dataset

Classifiers	ResNet101	Inception V3	Best ResNet101	Best Inception V3	Proposed
Ensemble Subspace KNN	94.6	94.9	95.8	95.5	100
Fine KNN	95.8	93.9	96.7	95.3	100
Cubic SVM	93.7	94.9	94.1	95.6	100
Weighted KNN	93.6	92.8	95.5	95.0	99.9
Cosine KNN	94.5	95.8	95.3	96.4	99.9

(Continued)

Table 6: Continued

Classifiers	ResNet101	Inception V3	Best ResNet101	Best Inception V3	Proposed
Quadratic SVM	94.8	94.9	95.9	96.0	100
Linear SVM	94.1	95.8	96.4	96.8	99.9
Medium KNN	93.5	94.1	97.8	95.7	99.9
Ensemble Bagged Trees	94.2	93.6	98.1	94.8	99.3

Table 7: Comparison of overall proposed framework accuracy with middle steps on Hollywood dataset

Classifiers	ResNet101	Inception V3	Best ResNet101	Best Inception V3	Proposed
Ensemble Subspace KNN	95.7	95.8	96.8	97.3	99.9
Fine KNN	95.9	97.1	97.9	98.0	99.9
Cubic SVM	95.8	96.5	96.5	97.4	99.8
Weighted KNN	95.5	94.4	97.1	96.3	99.7
Cosine KNN	96.1	96.7	96.5	97.7	99.5
Quadratic SVM	96.7	95.9	97.1	96.9	99.7
Linear SVM	95.3	96.2	95.4	97.5	99.3
Medium KNN	95.0	95.6	97.8	98.4	99.4
Ensemble Bagged Trees	93.8	94.2	97.3	96.0	98.7

Table 8: Comparison of overall proposed framework accuracy with middle steps on IXMAS dataset

Classifiers	ResNet101	Inception V3	Best ResNet101	Best Inception V3	Proposed
ES KNN	94.2	94.9	95.7	96.7	99.1
Fine KNN	93.3	93.4	94.8	95.4	99.0
Cubic SVM	92.5	93.4	93.6	95.4	98.7
Weighted KNN	93.5	94.7	94.6	95.6	97.5
Cosine KNN	92.9	93.6	93.5	94.9	95.8
Quadratic SVM	91.1	92.5	92.2	96.2	98.2
Linear SVM	92.6	91.3	93.7	93.7	96.8
Medium KNN	92.8	93.1	93.9	95.4	95.9
EBT	91.5	94.4	92.1	95.8	95.0

At the end, a comparison of proposed framework accuracy is conducted with state of the art (SOTA) techniques using different selected datasets, as given in Tab. 9. In this table, it is noted that authors of [34,35,43] used UCF sports dataset and achieved accuracies of 99.3%, 96.8%, and 92.67%, respectively. The proposed method attained 100% on UCF Sports dataset with minimum execution time. Similarly, authors used UT-Interaction dataset and achieved accuracies of 96.7%, 96.4%, and 99%. The proposed method achieved an accuracy of 100%. For IXMAS and Hollywood dataset, the proposed framework achieved an accuracy of 99.1% and 99.9% which is improved than the recent methods. Overall, values given in this table, it is clear that the proposed framework of HAR achieved improved accuracy than SOTA techniques.

Table 9: Comparison of proposed framework with SOTA techniques

Reference	Year	Dataset	Accuracy (%)
[43]	2021	UCF Sports	96.8
[35]	2021	UCF Sports	99.3
[34]	2021	UCF Sports	92.67
Proposed	**-**	**UCF Sports**	**100**
[44]	2021	UT-Interaction	96.7
[45]	2021	UT-Interaction	96.4
[29]	2021	UT-Interaction	99
Proposed	**-**	**UT-Interaction**	**100**
[21]	2020	IXMAS	95.2
[30]	2019	IXMAS	88.05
Proposed	**-**	**IXMAS**	**99.1**
[6]	2021	Hollywood	99.2
Proposed	**-**	**Hollywood**	**99.9**

5 Conclusion

Human action recognition (HAR) is rapidly gaining popularity in the field of pattern recognition and machine learning based on its important application-video surveillance. In this article, a new framework is proposed for HAR based deep learning and improved WOA. The experimental process is conducted on four publicly accessible datasets such as Ut-Interaction, Hollywood, IXMAS, and UCF Sports and attained an accuracy of 100%, 99.9%, 99.1%, and 100%, respectively. Comparison with SOTA techniques, it is observed that the proposed framework recognition accuracy is improved than the recent techniques. From the results, we conclude that the fusion based framework give the better accuracy than recognition performance on individual deep learning features and optimization algorithm. The optimization algorithm reduces the execution time during the testing process. The improved optimization algorithm reduced computational time without reducing classification accuracy, which is the work's strength. In the future, large datasets such as UCF101, Muhavi, and HMDB51 will be used for evaluation. Furthermore, for HAR, a single stream CNN framework will be considered.

Funding Statement: The authors would like to thank research network for collaboration. This research work is supported in part by Chiang Mai University and HITEC University.

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

References

[1] M. Sharif, T. Akram, M. Raza, T. Saba and A. Rehman, "Hand-crafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition," *Applied Soft Computing*, vol. 87, no. 2, pp. 105986, 2020.

[2] Y. D. Zhang, M. Allison, S. Kadry, S. H. Wang and T. Saba, "A fused heterogeneous deep neural network and robust feature selection framework for human actions recognition," *Arabian Journal for Science and Engineering*, vol. 11, no. 2, pp. 1–16, 2021.

[3] P. Zhang, C. Lan, J. Xing, W. Zeng and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 1963–1978, 2019.

[4] R. Zhao, W. Xu, H. Su and Q. Ji, "Bayesian hierarchical dynamic model for human action recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, NY, USA, pp. 7733–7742, 2019.

[5] A. Kamel, B. Sheng, P. Yang, P. Li and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 6, pp. 1806–1819, 2018.

[6] S. Khan, M. Alhaisoni, U. Tariq, H. S. Yong and A. Armghan, "Human action recognition: A paradigm of best deep learning features selection and serial based extended fusion," *Sensors*, vol. 21, no. 11, pp. 7941, 2021.

[7] Y. D. Zhang, S. A. Khan, M. Attique, A. Rehman and S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 35827–35849, 2021.

[8] H. Slimani, Y. Benezeth and F. Souami, "Learning bag of spatio-temporal features for human interaction recognition," in *Twelfth Int. Conf. on Machine Vision*, New Delhi, India, pp. 1143302, 2020.

[9] M. Sharif, F. Zahid, J. H. Shah and T. Akram, "Human action recognition: A framework of statistical weighted segmentation and rank correlation-based selection," *Pattern Analysis and Applications*, vol. 23, no. 8, pp. 281–294, 2020.

[10] M. Ahmed, M. Ramzan, H. U. Khan, S. Iqbal and J. I. Choi, "Real-time violent action recognition using key frames extraction and deep learning," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 1–15, 2021.

[11] I. M. Nasir, M. Raza, J. H. Shah and A. Rehman, "Human action recognition using machine learning in uncontrolled environment," in *2021 1st Int. Conf. on Artificial Intelligence and Data Analytics*, Riydah, Saudi Arabia, pp. 182–187, 2021.

[12] S. Kiran, M. Y. Javed, M. Alhaisoni, U. Tariq and Y. Nam, "Multi-layered deep learning features fusion for human action recognition," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 1–15, 2021.

[13] M. Alhaisoni, A. Armghan, F. Alenezi, U. Tariq and Y. Nam, "Video analytics framework for human action recognition," *Computers, Materials & Continua*, vol. 70, no. 4, pp. 1–15, 2021.

[14] S. A. Khan, S. Hussain, S. Xiaoming and S. Yang, "An effective framework for driver fatigue recognition based on intelligent facial expressions analysis," *IEEE Access*, vol. 6, no. 5, pp. 67459–67468, 2018.

[15] M. Nawaz, T. Nazir, A. Javed, U. Tariq and M. A. Khan, "An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization," *Sensors*, vol. 22, no. 1, pp. 434, 2022.

[16] F. Saleem, M. Alhaisoni, U. Tariq, A. Armghan and F. Alenezi, "Human gait recognition: A single stream optimal deep learning features fusion," *Sensors*, vol. 21, no. 5, pp. 7584, 2021.

[17] Z. U. Rehman, F. Ahmed, R. Damaševičius, S. R. Naqvi and W. Nisar, "Recognizing apple leaf diseases using a novel parallel real-time processing framework based on MASK RCNN and transfer learning: An application for smart agriculture," *IET Image Processing*, vol. 15, no. 9, pp. 2157–2168, 2021.

[18] M. Rashid, M. Alhaisoni, S. H. Wang, S. R. Naqvi and A. Rehman, "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," *Sustainability*, vol. 12, no. 4, pp. 5037, 2020.

[19] M. Koohzadi and N. M. Charkari, "Survey on deep learning methods in human action recognition," *IET Computer Vision*, vol. 11, no. 2, pp. 623–632, 2017.

[20] M. Zahid, F. Azam, M. Sharif, S. Kadry and J. R. Mohanty, "Pedestrian identification using motion-controlled deep neural network in real-time visual surveillance," *Soft Computing*, vol. 6, no. 2, pp. 1–17, 2021.

[21] K. Javed, S. A. Khan, T. Saba, U. Habib and J. A. Khan, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimedia Tools and Applications*, vol. 13, no. 2, pp. 1–27, 2020.

[22] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 5, pp. 1097–1105, 2012.

[23] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI Conf. on Artificial Intelligence*, NY, USA, pp. 1–6, 2017.

[24] T. Akram, M. Sharif, N. Muhammad, M. Y. Javed and S. R. Naqvi, "Improved strategy for human action recognition; Experiencing a cascaded design," *IET Image Processing*, vol. 14, no. 11, pp. 818–829, 2019.

[25] M. Alhaisoni, U. Tariq, N. Hussain, A. Majid and R. Damaševičius, "COVID-19 case recognition from chest CT images by deep learning, entropy-controlled firefly optimization, and parallel feature fusion," *Sensors*, vol. 21, no. 2, pp. 7286, 2021.

[26] M. Mittal, L. M. Goyal and S. Roy, "A deep survey on supervised learning based human detection and activity classification methods," *Multimedia Tools and Applications*, vol. 4, no. 1, pp. 1–57, 2021.

[27] M. Bilal, M. Maqsood, S. Yasmin, N. U. Hasan and S. Rho, "A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes," *The Journal of Supercomputing*, vol. 78, no. 21, pp. 2873–2908, 2022.

[28] A. Sarkar, A. Banerjee, P. K. Singh and R. Sarkar, "3D human action recognition: Through the eyes of researchers," *Expert Systems with Applications*, vol. 17, no. 7, pp. 116424, 2022.

[29] C. Liu, J. Ying, H. Yang, X. Hu and J. Liu, "Improved human action recognition approach based on two-stream convolutional neural network model," *The Visual Computer*, vol. 37, no. 2, pp. 1327–1341, 2020.

[30] V. A. Chenarlogh and F. Razzazi, "Multi-stream 3D CNN structure for human action recognition trained by limited data," *IET Computer Vision*, vol. 13, no. 10, pp. 338–344, 2019.

[31] A. Sharif, K. Javed, H. Gulfam, T. Iqbal and T. Saba, "Intelligent human action recognition: A framework of optimal features selection based on euclidean distance and strong correlation," *Journal of Control Engineering and Applied Informatics*, vol. 21, no. 15, pp. 3–11, 2019.

[32] N. Jaouedi, N. Boujnah and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 20, pp. 447–453, 2020.

[33] M. Sharif, T. Akram, M. Raza, T. Saba and A. Rehman, "Hand-crafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition," *Applied Soft Computing*, vol. 87, no. 21, pp. 1–26, 2020.

[34] A. Abdelbaky and S. Aly, "Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network," *Multimedia Tools and Applications*, vol. 80, no. 32, pp. 20019–20043, 2021.

[35] F. Afza, M. Sharif, S. Kadry, G. Manogaran and T. Saba, "A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection," *Image and Vision Computing*, vol. 106, no. 17, pp. 104090, 2021.

[36] A. Abdelbaky and S. Aly, "Human action recognition using short-time motion energy template images and PCANet features," *Neural Computing and Applications*, vol. 32, no. 9, pp. 12561–12574, 2020.

[37] S. P. Sahoo, S. Ari, K. Mahapatra and S. P. Mohanty, "HAR-Depth: A novel framework for human action recognition using sequential learning and depth estimated history images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 21, no. 11, pp. 1–13, 2020.

[38] K. Muhammad, U. Amin, A. S. Imran and M. Sajjad, "Human action recognition using attention based LSTM network with dilated CNN features," *Future Generation Computer Systems*, vol. 125, no. 7, pp. 820–830, 2021.

[39] M. Sharif, T. Akram, R. Damaševičius and R. Maskeliūnas, "Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization," *Diagnostics*, vol. 11, no. 1, pp. 811, 2021.

[40] S. Yadav, J. K. Sandhu, Y. Pathak and S. Jadhav, "Chest x-ray scanning based detection of COVID-19 using deepconvolutional neural network," *Future Generation Computer Systems*, vol. 125, no. 7, pp. 820–830, 2021.

[41] D. Tabernik, M. Kristan and A. Leonardis, "Spatially-adaptive filter units for compact and efficient deep neural networks," *International Journal of Computer Vision*, vol. 128, no. 61, pp. 2049–2067, 2020.

[42] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, no. 17, pp. 51–67, 2016.

[43] B. S. Kumar, S. V. Raju and H. V. Reddy, "Human action recognition using a novel deep learning approach," *Materials Science and Engineering*, vol. 1042, no. 31, pp. 1–17, 2021.

[44] S. Kiran, M. Younus Javed, M. Alhaisoni, U. Tariq and Y. Nam, "Multi-layered deep learning features fusion for human action recognition," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 4061–4075, 2021.

[45] W. Ahmed, M. H. Yousaf, A. Yasin and M. Maqsood, "Robust suspicious action recognition approach using pose descriptor," *Mathematical Problems in Engineering*, vol. 2021, no. 12, pp. 1–12, 2021.