International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 2, April, 2025, pp. 995-1005
Journal homepage: http://iieta.org/journals/ts

Remote Patient Healthcare Monitoring and Pose Recognition over Video Sensors

Bayan Alabdullah[1], Muhammad Tayyab[2], Haifa F. Alhasson[3], Naif S. Alshammari[4], Ahmad Jalal[2,5*]

[1] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[2] Department of Computer Science, Air University, Islamabad 44000, Pakistan
[3] Department of Information Technology, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia
[4] Department of Computer Sciences, College of Computer and Information Sciences, Majmaah University,
Majmaah 11952, Saudi Arabia
[5] Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, South Korea

Corresponding Author Email: ahmjal@yahoo.com

https://doi.org/10.18280/ts.420233

Received: 26 February 2025
Revised: 2 April 2025
Accepted: 10 April 2025
Available online: 30 April 2025

Keywords:
feature fusion, joint points, machine learning, MEMM, scalable keypoints, silhouettes

ABSTRACT

Considering the fact of increasing population and as a result, the number of patients is constantly increasing the delivery of medical services must be prompt and of good quality. There is no question that any perfect healthcare system should track the given activities, behaviours, schedules, and even general health of the patients. The focus of this study is in utilizing ML and DL in tracing patients and diagnosing ailments. Mounted cameras from different orientations and placements are intended for health monitoring and event surveillance for possible medical issues in the future. Classical ML and more advanced DL methods are applied for analyzing physiological and behavioral patterns to identify health conditions and stress the importance of high accuracy and real-time performance for practical purposes. Initial outcomes show that patient outcomes are enhanced by getting timely medical advice while decreasing hospital-based burdens on staff. Two challenging datasets, ETRI-Activity3D and NTU-RGB-D, are pre-processed to remove unimportant data and reduce computational load. A strong multi-feature description system can extract gradient values, and the process is followed by feature matching and feature optimization. Then, in the last Maximum Entropy Markov Model (MEMM) classifier is used, achieving an accuracy of 97.2% for ETRI-Activity3D, and 98.3% for the NTU-RGB-D dataset.

1. INTRODUCTION

Supervision of individuals in need of healthcare services is crucial for tracking their behavior and protecting their well-being, especially in caring for elderly people, rehabilitation centers, and hospitals. Given the fact that a rising number of people require constant medical supervision, the need for systems capable of creating permanent checks and responding immediately escalates. The present work proposes a health monitoring and surveillance system that incorporates multiple camera view analysis, employing machine learning and deep learning algorithms to assess the indices of physiology and behavior. Through early detection of health risks, such systems will greatly improve patient experience, particularly for one living alone, a critically injured patient, the elderly, or those in a rehabilitation center. Flashing light technology aims to enhance the quality of patient's and senior's lives due to constant activity and efficient emergency response. It reduces the pressure that healthcare providers face while delivering medical care and achieves the purpose of providing care at the right time. Apart from healthcare, the system uses are numerous; in workplace safety, one can look for hazardous behaviours in a bid to avoid accidents; in sports analytics, the system can monitor the physical and mental status of athletes and their performance.

HAR systems use multiple approaches to function, consisting of data collection followed by feature extraction and classification steps. Three main data acquisition instruments consist of RGB cameras and depth sensors together with wearable devices. Systems advance accuracy levels by processing different types of data inputs, including audio data and body signals. The traditional approach to feature extraction involved human-based methods that included Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT), and spatiotemporal descriptors. The use of deep learning techniques, including Convolutional Neural Networks and Long Short-Term Memory and Transformer models, has gained wide acceptance in recent times. Sport vector machines (SVMs), together with deep neural networks, are among the classifier choices, along with data application-dependent network selection for these techniques, which need normalization or segmentation as preprocessing before classification.

The existing systems have achieved major breakthroughs, although they operate under essential boundaries. The computational requirements of deep learning models make them unsuitable for real-time healthcare usage, even though they are effective for their tasks. The development of models that generalize to real-world situations becomes challenging because existing systems require large labelled datasets that get collected primarily from controlled environments. Using visual cues exclusively delivers insufficient results because it overlooks crucial contextual or behavioral signals that would improve diagnostic accuracy plus timely reaction ability.

This motivates for development of a real-time context-aware and efficient framework. Researchers have highlighted an important void in the field because optimization-based methods, especially evolutionary algorithms, need integration with machine learning for developing weight-efficient systems. The current research has inadequate examination of how to unite behavioral and physiological inputs alongside multi-view visual data to create more comprehensive patient context.

To address these gaps, we suggest implementing evolutionary algorithms combined with machine learning principles into our architecture because they optimize feature selection, improve classification accuracy, and decrease computational load. The proposed system achieves successful classification of patient monitoring scenarios with notable precision, according to our results. Following this introduction, the paper combines the necessary content through these sections: Section 2 presents detailed research into current HAR systems along with their known weaknesses. The system design section explains the multi-view camera system and feature extraction protocol, along with details about classifier implementation. The evaluation of experimental outcomes and performance assessments is provided in depth in Section 4, and the paper concludes through Section 5 with a finding's discussion.

2. RELATED WORK

The research community has investigated different innovative sensor methods for patient monitoring and healthcare challenges through ambient, video, and wearable systems. Secondly, these investigations demonstrate multiple approaches to monitoring with their current performance standards together with encountered operational limitations. The research of Saner et al. [1] combines ambient and wearable technology sensors, which integrate PIR motion sensors into the system. The work of Ferraris et al. [2] uses Microsoft Kinect v2 RGB-Depth camera as an ambient contactless tool to assess Parkinson's Disease (PD) postural stability in home environments. Maitre et al. [3] developed an ambient sensor-based system that included UWB radars in a constructed prototype apartment. The evaluation performance of three UWB radars depends on the new data processing technique with its implementation of a band-pass Chebyshev Type I filtering method. Abbe and O'Keeffe [4] developed a systematic approach for healthcare Continuous Video Monitoring (CVM) implementation that combines market research with steering committee formation to deploy decision tools for clinical staff. The research of Diao et al. [5] presents video-based physiologic monitoring through the application of ML and DL systems to analyze heat and light sensor data to detect patients' heart rates and respiratory rates. Herfandi et al. [6] studied real-time patient tracking systems based on WBAN technology in IoMT networks for ongoing health monitoring tasks. These research projects illustrate how sensor-based techniques monitor the health of patients through multiple sensor types, such as environment, video surveillance, and wearable devices, to solve different healthcare problems and improve healthcare outcomes.

3. METHODS AND MATERIALS

The system aims to analyze input videos and recognize patient activities. Initially, the video is segmented into frames, which are then denoised by removing unnecessary background. An averaging filter is applied to blur the frames, followed by further denoising through the conversion of the RGB images into grayscale. After completing the denoising process, the frames are forwarded to the next step: human detection. Human detection is performed using two different methods: the first involves extracting human silhouettes, while the second uses pose estimation. Four distinct features are extracted from both detection techniques, two for each. For joint points, DOF, 3D Point Cloud features are extracted, while for full body Distance Transform, and GMM are utilized. These extracted features are optimized through early fusion and the Whale Optimization Algorithm (WOA). Then, for classification, we have employed a Maximum Entropy Markov Model (MEMM) classifier, which is applied to the proposed system and demonstrates superior performance compared. Figure 1 illustrates the general design of the proposed system.

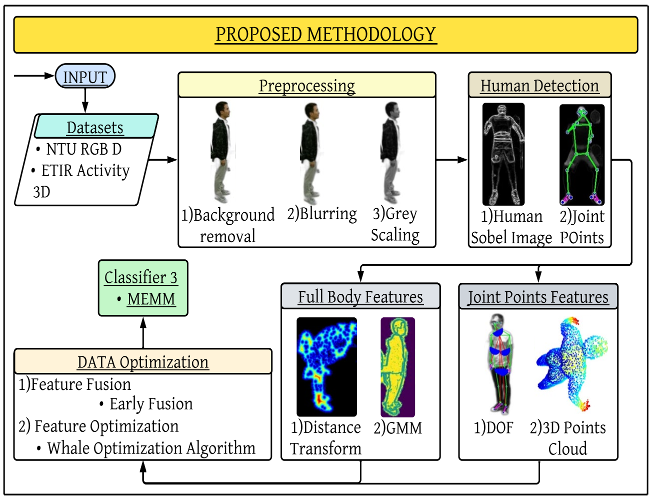


Figure 1. Proposed architecture for classification

3.1 Image pre-processing

Preprocessing is a major input in many areas, including but not limited to machine learning, image processing, and deep learning. In this context, the original input is in the form of videos, and the actions associated with this input are segmented frames [7]. These frames normally come with background noise, which makes background subtraction an imperative step to get the object of interest to improve the quality of input data for other processes. Since image backgrounds and noise distract the algorithm from learning relevant features of the target object, the application of ML/DL increases the model's efficiency and decreases the chance of over-fitting. Here, the process isolates pixels that are associated with the central object and removes unnecessary information from the background, thereby making the object of interest stand out. The other process is the smoothing process, which is used to remove detail, edges, or extra detail in an image. This is done by taking the mean of the pixel intensities in a neighborhood of each pixel, thereby resulting in an image with a smoother appearance. Reducing image noise has a great advantage in the subsequent calculations and reduces the amount of insignificant information. In image processing, ML, and AI applications, one filter type that we often use to perform smoother operations is the Gaussian blur filter. This filter then brings the center of the pixel into every image and blurs the surrounding area. In addition to this, conversion to grayscale is also common because working with 3D images will be more computationally costly [8]. A common type of this approach is quantization; it assigns a rational value between 0 and 255 to each of the pixels; in other words, it reduces a 3-dimensional picture to 2 dimensions. The outcome is the single-channel image view, which takes less perceptive power compared to the streams of the RGB model. The grayscale value is determined as a weighted average of the RGB pixel values using the values: Grey=0,2988R +0,5872G + 0,1137 B. Their weights provided concern about color intensity as being perceived by humans. These, as stated by NTPC (National Thermal Power Corporation Limited), are based on perception and could change under applications' demands. Fine-tuned methods for the RGB-to-grayscale conversion do exist for varied different purposes. Figure 2 summarizes the pre-processing techniques in sequential order from top to bottom.

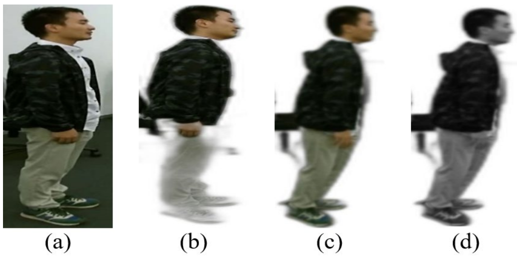


Figure 2. The steps of pre-processing are (a) Extracted Frame, (b) background removed, (c) blurred, (d) grey scaled

3.2 Human detection

Human detection is the determination of the presence and location of humans within a given frame or an image, much as a basic concept widely used in the many fields of computer vision [9]. This step is very important as it allows separate the object of our interest; thus, after enhancing procedures in the frame of our research, all the next processes will be concentrated on the identified human. To achieve this, we employ two distinct techniques: The main processes that should be mentioned are called image silhouette extraction and skeletonization. In silhouette extraction, we have a full body representation, while skeletonization paints the human figure with the barest minimum boundary as the essence of the figure isn't the flesh but the bone. Specifically, for silhouette extraction, we use the Sobel operator, which is among the most widely used operators for edge detection. Edge detection techniques such as the Sobel operator compute the rate of change in intensity between two pixels or between two neighboring pixels, areas of high spatial frequency which are areas of edges. This is done using two 3×3 convolution kernels to identify changes in intensity in the x and y axes. Due to its simplicity and efficiency, the Sobel operator is highly effective in enhancing salient features while rejecting noise which is especially important when needs to detect edges accurately This step adds value to other subsequent processes such as feature extraction because it only focuses on areas of interest in an image thereby improving the efficiency of computation [10]. When used in conjunction with skeletonization, the use of the two models guarantees a rich data capture of human features that is vital for the subsequent analytic process as in Eq. (1) and Eq. (2). Figure 3 illustrates the Sobel kernels used for this operation, which can be described as follows.

$$G_{x,y}(i,j) = \sum_{m=-1}^{1}\sum_{n=-1}^{1} G_{x,y}(m,n).I(i+m,j+n) \quad (1)$$

where,

$$G(i,j) = \sqrt{G_x(i,j)^2 + G_x(i,j)^2} \quad (2)$$

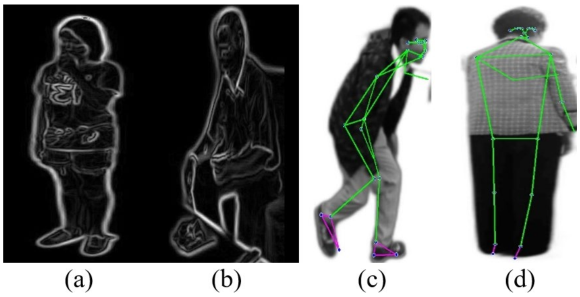


Figure 3. Binary images of human after extracting sobel image (a) NTU-RGB-D (b) ETRI-Activity3D, Illustrations of MOCAP (c) NTU-RGB-D (d) ETRI-Activity3D

These are the horizontal (*x*) and vertical (*y*) gradient kernels, respectively. For Sobel, these are typically 3×3 kernels, but for a more complex operator, you can increase to 5×5, is the summation now depends on a kernel of size (2*k*+1)×(2*k*+1), allowing for larger or more sophisticated filters, and $I(i+m,j+n)$ represents the pixel intensities in the neighborhood of pixel (*i*,*j*). Joint Motion Capture (MOCAP) is an analytical sub-component of the general motion capture technology that is more particularly related to the accurate direction and measurement of human joints or other articulated mechanisms. It is a paramount technology to applied domains such as biomechanics, sports science, computer animation and graphics, computer graphics, film and games, and virtual reality. What sets MOCAP apart is that while it delivers a detailed account of the operation of joints, it captures the highly complex motion of the skeletal structure that can accommodate complicated movements in physical modeling. The rotations at the joints are measured and modeled using quaternions, which offer an efficient and stable method of quantifying joint rotations. Eq. (3) and Eq. (4), present the quaternions that incorporate trigonometric functions and vector components that represent the axis of rotation and the angle of rotation equally well in a highly efficient method. Further, dual quaternions generalize this approach for including translations, which allow the modelling of combined rotational and translational joint movements, which is critical for effecting detailed capture and analysis of articulated systems.

$$q(t) = cos\left(\frac{\omega t}{2}\right) + sin\left(\frac{\omega t}{2}\right)\left(u_x i + u_y j + u_y k\right) \quad (3)$$

For combined joint transformations,

$$\mathfrak{Q} = q_r + \frac{\in}{2} q_r \oplus (0, t) \tag{4}$$

where, $q(t)$ is quaternion representing the rotation of a joint at a time, ω is the angular velocity of rotation, measured in radians per unit of time, t is the time variable, indicating the specific moment at which the rotation is computed, $\left(u_x i + u_y j + u_y k\right)$ is a unit vector that defines the axis of rotation in (i, j, k) 3D space, $\mathfrak{Q}$ is dual quaternion representing a joint's combined rotation and translation, and q_r rotational quaternion, describing the joint's orientation. The imagery output of human detection can be seen in Figure 3 where images a and b are Sobel images, and images c and d are MOCAP of humans.

3.3 Feature extraction

The first approach is based on the Degree of Freedom technique in which we find angles at the joint points and distances between them. This method also presents a new way of obtaining geometric features from human pose images by applying computer vision and data analysis. When using joint points as reference indicators, we calculate different geometric characteristics, including distances in distinct joint couples and angles in corresponding joints. This approach offers a more specific quantitative measurement of body postures and movements and all aspects of the biomechanics of human locomotion and provides a deeper insight into the subject. Furthermore, it increases the accuracy and robustness of pose estimation in various fields of interest, sports analytics, physical therapy, and human factors engineering. Eq. (5) and Eq. (6) use the Vector algebra geometry that brings together Euclidean distance and angle calculations to derive an all-encompassing range of features. This guarantees a solid quantitative context that facilitates understanding of human motion in two and three dimensions.

$$F_{geo} = \left\{ \left\| J_i - J_j \right\|, cos^{-1}\left(\frac{(J_i - J_k) \cdot (J_j - J_k)}{\|J_i - J_k\| \cdot \|J_j - J_k\|} \right) \right\} \tag{5}$$

Such that,

$$i, j, k \epsilon \mathcal{J}, i \neq j \neq k \tag{6}$$

where, J_i,J_j,J_k are the joint points in 2D or 3D space represented as vectors, $\left\|J_i - J_j\right\|$ is Euclidean distance between joints $i, and\ j$, $(J_i - J_k) \cdot (J_j - J_k)$ is dot product of vectors, $\mathcal{J}$ is the set of all detected joint points, and F_{geo} is set of all geometric features extracted, including distances and angles. Feature extraction from 3D point cloud is a crucial part in preparation of 3D spatial data for different uses in human motion analysis, object identification, identification of state of patient and 3D scene modeling. It includes procedures for defining and enumerating distinctive geometric and structural properties allowing the representation of intricate forms with low dimensionality [11]. Some of the identified techniques include Principal Component Analysis (PCA) which plays a role in reducing dimensions in the shape of the surface, curvature estimation for capturing the geometrical properties of the surface, and determination of the surface normal as a way of getting the orientation of points within the shape in the 3D space is also great way of characterizing the shapes in spatial sense. Recently developed potent PointNet++ employs deep learning for learning topological features, which drastically increases accuracy and provides robustness. PointNet++ is an extension of PointNet wherein we introduce hierarchical feature learning to address the issues of capturing not only the local relation of neighbors for point clouds and global features. This makes it very suitable for dense estimates from low-dimensional datasets, especially if the point density varies greatly. Moreover, 3D convolutional neural networks (CNNs) are used to improve feature extraction where the dependencies in the input domain using graph-based methods that can capture spatial dependencies and hierarchical structures in the input. The enhancement or fine-tuning of features really aids subsequent processes such as classification, segmentation, and registration, while improving recognition makes feature extraction a fundamental step in 3D data processing pipelines. Eq. (7) defines how PointNet++ aggregates feature hierarchically. At each layer, the local neighborhood is considered, and features are updated by combining spatial relationships with the features of neighboring points.

$$f^{(l+1)}(p_i) = \max_{p_i \in \mathcal{N}(p_i)} \emptyset\left(f^{(l)}\left(p_j\right), \psi(p_j - p_i)\right) \tag{7}$$

where, $\left(f^{(l)}\left(p_j\right)\right)$ is the feature vector of point p_j at layer l, $f^{(l+1)}(p_i)$ is the updated feature vector of p_j point at layer $\mathrm{l} + 1$, $\mathcal{N}(p_i)$ is the local neighborhood, $\psi(p_j - p_i)$ a positional encoding function to capture relative spatial relationships, and max is a symmetric aggregation function. Figure 4 is an imagery representation of the output of all joint point features.

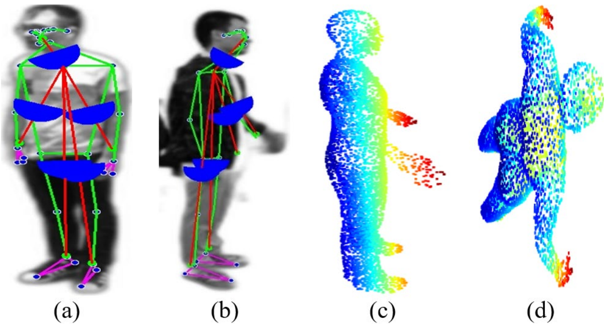


Figure 4. Feature on human MOCAP (DOF)(a) NTU-RGB-D (b) ETRI-Activity3D, 3D point cloud (c) NTU-RGB-D (d) ETRI-Activity3D

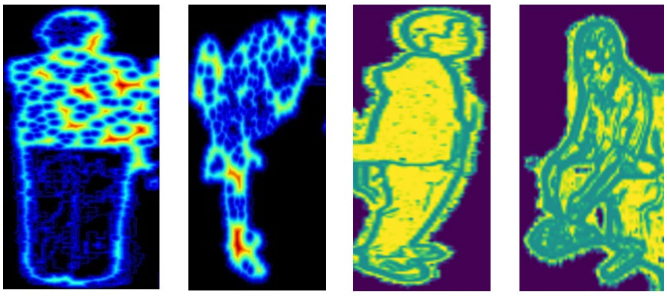
The other three features that are extracted are BRISK, HOG and ORB. These feature points are applied to the full human body (silhouettes). BRISK (Binary Robust Invariant Scalable Keypoints) is a computer vision algorithm renowned for its robust feature detection and description capabilities. Utilizing binary patterns for image patch representation, BRISK excels in scenarios with varying lighting conditions and viewpoints, ensuring robustness and reliability. BRISK constructs a scale pyramid and detects key points using a variant of FAST, refining their positions for accuracy. It generates binary descriptors based on local intensity patterns around key points, facilitating efficient image matching and recognition in computer vision tasks. BRISK's binary descriptors enable efficient feature matching and resilience to noise, enhancing its usability in tasks such as the matching of images, recognition of an object, and reconstruction of 3D objects, where robustness, accuracy, speed, and working are paramount. Its further processing is explained in Eq. (8).

$$HD(X,Y) = \sum_{i=1}^{N} x_i \oplus y_i = \sum_{i=1}^{n} (x_i, y_i) \tag{8}$$

where, $b(x_i, y_i)$ denotes bit inequality, in Eq. (15), x_i and y_i are the i^{th} bits of the descriptors X and Y respectively. Its examples of both datasets can be seen in Figure 5. The last feature extraction technique is Gaussian Mixture Model (GMM) is a probabilistic generative model, which is considered to be a soft clustering model, much of the time used in feature extraction for clustering and density estimation. GMMs model data with a mixture of several Gaussian distributions, with parameters of mean vector and covariance matrix. To estimate the above-mentioned parameters of these Gaussian components from a given dataset, expectation maximization (EM) is one of the frequently used methods. This approach provides a probability to each and every datum belonging to each Gaussian component and thus highlights the distribution structure and variability of the datasets. Possible extracted features from GMM may be the posterior probability, mean, variance, as well as weight by each of the Gaussian components or a selection thereof. These features are useful in many machines learning processes, including classification, clustering, and anomaly detection, because they can represent high-dimensional multi-modal probability density functions. Eq. (9) explains the basic concept of GMM.

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \frac{1}{(2\pi)^{d/2} |\Sigma_{k\cdot}|^{1/2}} exp \left(-\frac{1}{2}(x-\mu_k)^T \sum_{k}^{-1} (x-\mu_k)\right) \tag{9}$$

where, $p(x)$ is probability density function of the data point x, K number of Gaussian components, π_k is weight of the $K-th$ Gaussian component, where $\sum_{k=1}^{K} \pi_k = 1,\ \mu_k$ is mean vector and $\sum_k th$ covariance matrix of the $K-th$ Gaussian. Figure 5 shows the representation of full-body feature extraction points.



(a) (b) (c) (d)

Figure 5. Feature on human full body (Distance transform) (a) NTU RGB-D (b) ETRI-Activity3D, GMM (c) NTU RGB-D (d) ETRI-Activity3D

3.4 Features fusion

Early fusion is a concept in machine learning and data processing where the diverse sources of information and features are combined early, commonly before input into a learning model or algorithm. In this approach, we combine the features from the different modalities or different extraction techniques into one. This fused representation is then introduced as data for the model [12]. The strong point of this approach is that all correlations and dependencies between features from different sources are combined right after the feature extraction step, allowing the final model to work with enhanced information. For instance, in human action recognition, features like DOF, and GMM could be concatenated into one vector retaining the separate contribution of each feature but improving the capability of the model to distinguish between features of complex patterns across different modalities. However, early fusion requires careful normalization and scaling of features to ensure compatibility and prevent any single modality from dominating the combined representation. Moreover, its working can be understood with the help of Eq. (10), and Eq. (11).

$$y = f\left(\frac{[\sum_{i=1}^{n} \omega_i \cdot \emptyset_i(f_i) + Vec(H) - \mu_F]}{\sigma_F}\right) \tag{10}$$

$$H = \left(\sum_{i=1}^{n} \sum_{j=i+1}^{n} (\omega_i \cdot \emptyset_i(f_i)) \oplus \left(\omega_j \cdot \emptyset_j(f_j)\right)\right) \tag{11}$$

where, y is the final output, f_i is the feature vector extracted by the different techniques, n is the total number of feature extraction techniques, $\emptyset_i$ is the transformation function, ω_i are weight assigned to the feature vector, F is the fused feature, μ_F is mean of values, σ_F is standard deviation. Figure 6 shows the points distribution among different classes.

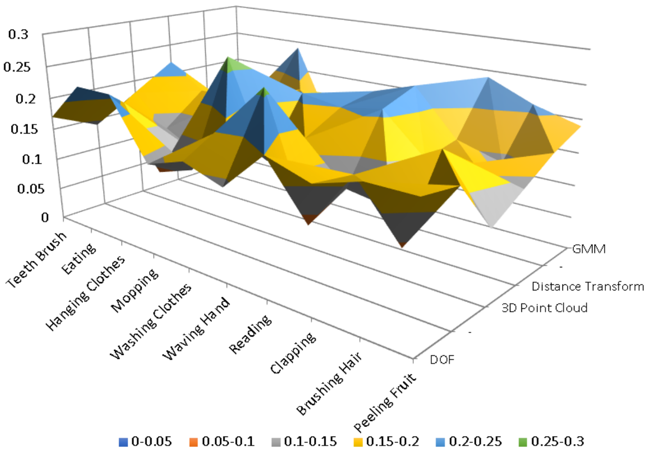


Figure 6. Graphical representation of feature fusion

3.5 Features optimization

In the context of machine learning, optimization is the process of setting the values of the parameters of a given model that best fit an objective function, usually error or loss. In performing learning, it assists in fine-tuning the weights and biases of the model in search of the required accuracy or efficiency. The Whale Optimization Algorithm (WOA) is an optimization method designed based on the hunting behavior of humpback whales and their bubble-net hunting strategy. The whale humpback whales optimization algorithm is a metaheuristic algorithm created to find the optimal solution to problems based on the feeding mechanism employed by humpback whales to exploit or explore the prey. In WOA, whales are modeled as candidate solutions within the search space. The algorithm alternates between two phases: exploitation and exploration. In exploitation, the whales advance towards a current solution, thus forming a spiral or encircling motion reminiscent of bubble-net hunting. In exploration, whales randomly search the space by moving far from the current best solution to avoid local optima. The balance of these phases is regulated by a parameter that is reduced linearly across iterations. These behaviors, in sequence, enable WOA to locate the global optima of numerous challenging optimization issues. Eq. (12) is the mathematical representation of WOA.

$$X_i^{(t=1)} = \begin{cases} X^* - A \cdot D, & if\ p < 0.5 and\ |\ A\ | \leq 1 \\ X^* + D' \cdot e^{bl} \cdot \cos(2\pi l), & if\ p < 0.5 \\ X_{rand} - A \cdot D, & if\ |A| > 1 \end{cases} \quad (12)$$

where, X^* Represents the position of the current best solution (i.e., the leader whale), X_{rand} is a randomly chosen solution from the population used for exploration, A is coefficient vector that controls the convergence behavior, D is the distance between a whale's position and the best solution, D' modified distance vector for spiral motion, p is random number in [0, 1] that determines whether to exploit (closer search) or explore (global search) the solution space, b is constant defining the shape of the spiral for encircling prey, l is a random number used in the spiral equation for diversification, e^{bl} an exponential decay factor controlling the amplitude of the spiral movement. Algorithm 1 gives full implementation of the optimizer. Figure 7 shows a 3D graph representation of WOA.

Algorithm 1: Structural and functional code of WOA

```
1. Input:
  - Objective function $f(x)$, where $x = (x1, x2, ..., xn)$
  - Population size $N$
  - Maximum number of iterations $T$
  - Search space boundaries $[lb, ub]$

2. Initialize
  - Initialize whale population $X_1 = X_1, X_2, ..., X_N)$ randomly within [lb, ub]
  - Evaluate fitness of each whale: $f(X_i)$ for each whale $X_i$
  - Determine the best whale position $X_{best}$ with the lowest fitness
  - Set initial parameter $\alpha = 2$

3. For t = 1 to T:
  - For each whale $X_i$ in population:
    - Generate random numbers $r \in [0,1], p \in [0,1]$
    - Update parameters:
      - $A = 2 \cdot a \cdot r - a$
      - $C = 2 \cdot r$

    - If $p < 0.5$ then:
      - If $|A| < 1$ then:
        // Exploitation: encircling the prey (best solution)
        - $D = |C \cdot X_{best} - X_i|$
        - $Xi = X_{best} - A \cdot D$
      - Else:
        // Exploration: search for random prey
        - Select random whale $X_{rand}$from population
        - $D = |C \cdot X_{rand} - X_i|$
        - $X_i = X_{rand}$- A · D
    - Else:
      // Exploitation: spiral updating position
      - $D = |X_{best} - X_i|$
      - $b = constant\ (e.g., 1)$
      - $l = random\ number\ in$ [-1, 1]
      - $X_i = D \cdot exp(b \cdot l) \cdot cos(2\pi \cdot l) + X_{best}$

    - Ensure $Xi$ is within $[lb, ub]$
    - Evaluate new fitness $f(Xi)$

    - If $f(Xi) < f(X_{best})$ then:
      - Update $X_{best} = X_i$

  - Update $a = 2 - (2 \cdot t / T)$// Decrease linearly
4. Output
  - Return $X_{best}$as the optimal solution
```

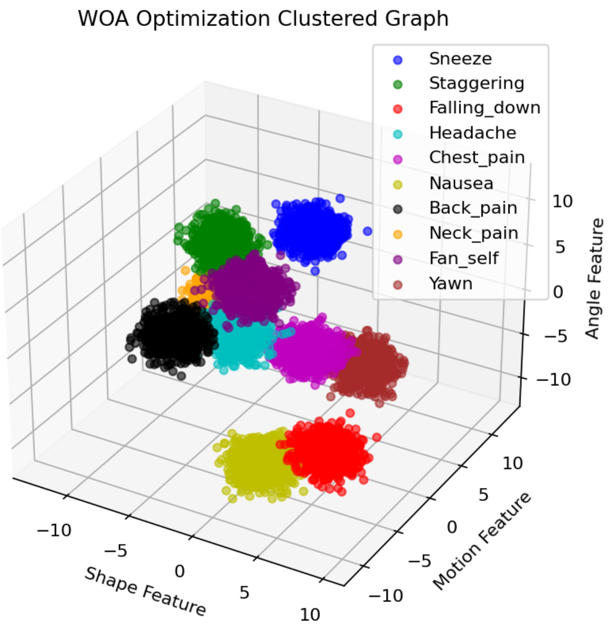


Figure 7. Accuracy optimization graph using whale optimization

3.6 Events classifier: MEMM

A MEMM classifier is a Maximum Entropy Markov Model that is a sequence model used to label as well as to segment sequences, such as in natural language processing and time series analysis. It integrates two concepts, the maximum entropy, a probability distribution theory that maximizes entropy, and the function of Markov, which holds that the conditions currently rely upon the previous condition. MEMMs estimate the probability of the next label when the current observation has been made, and the previous label known. This is done by using a conditional probability distribution commonly estimated by using logistic regression. The inputs to the model are feature sets; these are vectors that provide context as to the sequence: the model then estimates the probability by using the principle of Maximum Entropy; that is the model aims at choosing the least biased probability distribution for the data given certain constraints.

Maximum Entropy Markov Models (MEMMs) effectively solve problems that depend on sequence order or time-based events. The predictive process for MEMMs depends on both present input data and past label information. A logistic regression model estimated the necessary probabilities by using refined feature vectors. L2 regularization came into use during training as an effective measure against overfitting, which made the model more successful at generalization. The model required optimal configurations, which were discovered through a grid-searching process. The team evaluated multiple values of regularization strength alongside window dimensions and past information weight as they related to the Markov assumption. Through cross-validation, the model values achieved their final configuration because they demonstrated consistent performance across distinct data partitions.

During the prediction stage, the Viterbi decoding algorithm was applied to determine the most likely sequence of labels, using the conditional probabilities generated by the MEMM. This approach guarantees that the output sequence is the best possible one overall, considering the full context of the input. The complete process, starting from feature extraction, combining and refining the features, and leading up to classification with the MEMM, was carefully structured to ensure consistent results and strong performance across different runs. Algorithm 2 gives a complete pseudo-code for MEMM. Figure 8 shows the functional architecture.

Algorithm 2: Structural and functional code of WOA

```
Function: MEMM_Train (X, Y, Features)
Input:
  - X: Sequence data (observations)
  - Y: Ground truth label sequences
  - Features: Feature functions $\phi(x_t, y_t, y_{t-1})$

Steps:
  1. Initialize model weights $W$ randomly.
  2. For each sequence $(x_{seq}, x_{seq}) \in (X, Y)$:
     a. For each time step $t$:
        - Set $y_{prev} = y_{t-1}$ $(or\ START\ if\ t = 0)$
        - Compute scores: $Score(y_t) = W \cdot \phi(x_t, y_t, y_{prev})$
        - Compute probabilities: $P(y_t \mid x_t, y_{prev}) = softmax(Score)$

        1. Compute gradient of loss (NLL) and update weights: $W = W - \eta \cdot \nabla W$
        2. Return trained model $W$
Function: MEMM_Predict (X_test, W, Features, Label_Set)
Input:
  - X_test: Test sequences
  - W: Trained weights
  - Features: Feature functions
  - Label_Set: Set of possible labels

Steps:
  1. For each test sequence $x_{seq}$:
     a. Initialize Viterbi table $V$ and backpointer $B$
     b. For $t = 0$ to $T$:
        - For each label $y_t$:
        - For each previous label $y_{t-1}$:
          - Compute $Score = V[t-1][y_{t-1}] + logP(y_t \mid x_t, y_{t-1})$
          - Keep max score and store back pointer
     c. Backtrack to find optimal label sequence
  2. Return predicted sequences
```

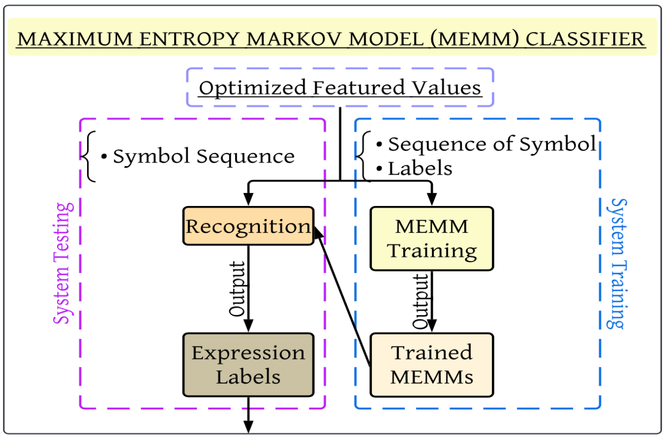


Figure 8. Functional structure of MEMM classifier

4. EXPERIMENTAL SETTINGS AND ANALYSIS

This section presents the experimental results of our approach and highlights its distinctions from previous research. Our method shows better accuracy results than earlier studies when evaluating these datasets.

4.1 Datasets description

Our research used two widely recognized datasets, ETRI-Activity3D and NTU-RGB-D, for patient action recognition purposes [13]. Both were specifically chosen due to their relevance to healthcare applications, including patient monitoring, elderly monitoring, rehabilitation analysis, and assisted living support. The ETRI-Activity3D dataset exists at the Electronics and Telecommunications Research Institute (ETRI) in South Korea as an important resource dedicated to enhancing telemedicine along with smart home innovations through daily human behavior video analysis [14]. It contains approximately 4.5TB of data across 55 diverse action classes captured in indoor environments. Our research included 10 specific classes that are frequently used during healthcare and assisted living operations. Importantly, only the video data from the dataset was used to align with our vision-based action recognition approach. The NTU-RGB-D dataset stands as one of the most extensive action recognition datasets because researchers from Nanyang Technological University (NTU) developed it. The NTU-RGB-D dataset organizes its information into three fundamental groups, which include Daily Actions, mutual Actions, and Medical Conditions. Our work analyzes only health-related activities from the Medical Conditions category that consists of 12 health-related classes that include activities such as walking and lying down because they are relevant to patient safety. The subset contains more than 50,000 video samples that establish an excellent foundation for health system training and assessment tasks. Research utilizes health-related portions of these datasets to confirm the validity of proposed methods on data that mirrors actual patient activity monitoring in clinical settings.

The experiments in our research were performed using

Python on an Intel Core i5 CPU with 8GB of RAM, we did not utilize a dedicated GPU. Cloud services served to store data, but the entire training process, together with evaluation operations, took place within local computing power. Executing 2000 images through the system took about 20-25 minutes. Our system was tested through an ablation study that examined different stages and configurations while indicating their corresponding inference times. The proposed system operates at a frame rate speed of 250–300 milliseconds when fully developed. The method proves to be efficient for execution on low-resource hardware systems, which enables its applications in real-world settings without access to high-performance computing resources. We used Eq. (13) for precision, Eq. (14) for recall, and Eq. (15) for accuracy to assess the performance of our recognition model. The findings showed a 97.2% accuracy rate on the ETRI-Activity3D dataset and 98.3% on the NTU-RGB-D dataset.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{13}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{14}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{15}$$

Figures 9 and 10 are the confusion matrices, and Tables 1 and 2 present the comparison of each class with their precision, accuracy, F1 score, and recall. Table 3 is a comparison table. Table 4 shows the ablation study of all the phases of our research.

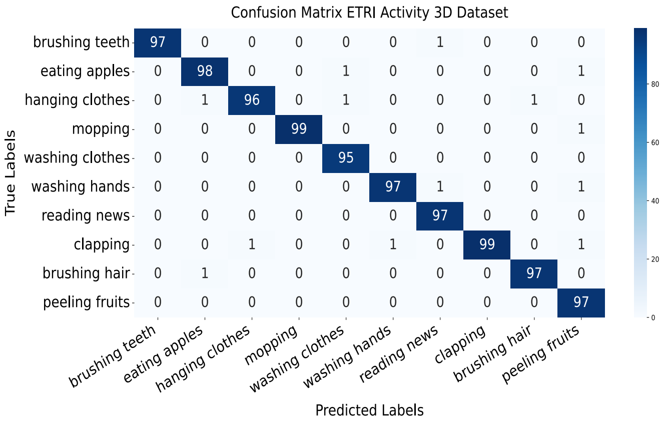


Figure 9. Confusion matrix (ETRI-Activity3D)

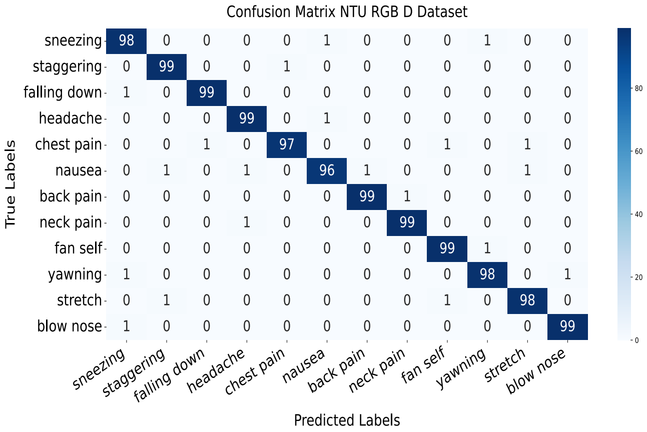


Figure 10. Confusion matrix (NTU-RGB-D)

Table 1. Accuracy table of each class of ETRI-Activity3D

Classes	Accuracy	Precision	Recall	F1 Score
Brushing Teeth	99.80	100.00	97.90	98.95
Eating Apple	99.70	98.00	98.00	98.00
Hanging Clothes	99.50	96.00	96.00	96.00
Mopping	99.90	99.00	99.00	99.00
Washing Clothes	99.40	95.00	95.00	95.00
Washing Hands	99.60	97.00	97.00	97.00
Reading Newspaper	99.60	97.00	97.00	97.00
Clapping	99.90	99.00	99.00	99.00
Brushing Hairs	99.60	97.00	97.00	97.00
Peeling Fruits	99.70	97.00	97.00	97.00
Mean	99.76	97.20	97.20	97.50

Table 2. Accuracy table of each class of YouTube

Classes	Accuracy	Precision	Recall	F1 Score
Sneezing	99.67	98.00	99.00	98.50
Staggering	99.83	99.00	99.00	99.00
Falling Down	99.83	99.00	99.00	99.00
Headache	99.83	99.00	99.00	99.00
Chest Pain	99.50	97.00	97.00	97.00
Nausea	99.50	96.00	96.00	96.00
Back Pain	99.83	99.00	99.00	99.00
Neck Pain	99.83	99.00	99.00	99.00
Fan Self	99.83	99.00	99.00	99.00
Yawing	99.67	98.00	98.00	98.00
Stretch	99.67	98.00	98.00	98.00
Blowing Nose	99.83	99.00	99.00	99.00
Mean	99.73	98.33	98.58	98.54

Table 3. Comparison of proposed model with state-of-the-art methods

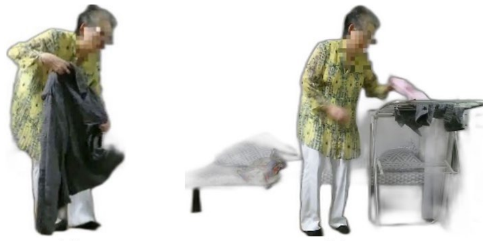
Methods	ETRI 3D (%)	NTU-RGB-D (%)
Wang et al. [15]	-	84.2
Xu et al. [16]	83.0	85.1
Li et al. [17]	-	86.5
Jang et al. [14]	-	88.0
Li et al. [18]	82.4	-
Li et al. [19]	83.3	-
Yan et al. [20]	86.8	-
Proposed System	97.2	98.3

Table 4. Ablation study of all the phases of our experiment showing tick mark for step included and cross for not including algorithms

Ablation Settings	Preprocessing	Human Detection	Full Body Features	Joint Point Features	Early Fusion	Whale Optimization	Classifiers			Accuracy		Inference Time (ms)
							CNN	LSTM	MEMM	NTU-RGB-D	ETRI-Activity3D	
Base Line	✓	✗	✗	✗	✗	✗	✓	✗	✗	88.2	87	200-250
Human Detection	✓	✓	✗	✗	✗	✗	✗	✓	✗	67.3	66.8	150-200
+ Full-body features	✓	✓	✓	✗	✗	✗	✗	✓	✗	72.1	71.2	150-200
+ Joint Points Features	✓	✓	✓	✓	✗	✗	✗	✓	✗	83.4	82.1	250-300
Human Detection + MEMM	✓	✓	✗	✗	✗	✗	✗	✗	✓	66.2	64.2	150-200
Full Body + MEMM	✓	✓	✓	✗	✗	✗	✗	✗	✓	71.2	70.7	150-200
All Features + MEMM	✓	✓	✓	✓	✗	✗	✗	✗	✓	79.7	76.4	150-200
Early Fusion	✓	✓	✓	✓	✓	✗	✗	✗	✓	85.1	85.3	200-250
Whale Optimizer	✓	✓	✓	✓	✗	✓	✗	✗	✓	91.8	90.3	200-250
Full Proposed System	✓	✓	✓	✓	✓	✓	✗	✗	✓	98.3	97.2	250-300

4.2 Failure cases

During the preprocessing stage, while the background is successfully removed, the system often fails to eliminate noise that exists within the foreground. This issue arises because the noisy pixels in the foreground are spatially connected or closely linked with the actual subject, causing the system to interpret them as relevant parts of the main object mistakenly. Consequently, this misinterpretation introduces ambiguity during further processing, which can negatively impact the system's accuracy and performance. Such confusion in distinguishing between true foreground features and irrelevant noise can lead to incorrect classification or recognition outcomes. Representative examples of these types of system failures are shown in Figure 11.



(a) (b)

Figure 11. Failure cases, (a) back ground removed properly, (b) case failure

5. CONCLUSIONS AND FUTURE WORK

A method which unites feature detection algorithms with optimal classifier systems enables the detection of key events during patient gait measurement. Mathematical models in the system track successive human body movements to perform precise behavioral analysis and event identification. The method achieves strong benchmark results. Upcoming work aims at four main enhancements for better practical use and improved stability. The next stage of development involves making the model more efficient for lower power consumption purposes and real-time edge-device capability alongside integration of biomechanical data elements such as joint motion measurement and energy expense metrics. We planned to add wearable sensor data as a supplement to visual information for enhancing recognition performance when vision is obstructed or when lighting is poor. The model development extends to analyze aerial video obtained from drones, which will permit the study of group characteristics along with event identification across extensive public areas.

As the video-based patient monitoring systems create ethical and privacy issues. This research employs only anonymized public benchmark datasets (NTU RGB+D and ETRI-Activity3D) that meet approved consent standards; however, we understand that deployment requirements in clinical or residential settings demand strict adherence to privacy rules and ethical practices. The research demands obtaining patient consent along with strong data anonymization practices combined with secure databases and processing systems. The development of edge computing alongside federated learning and pose-based abstraction solutions aims to improve user privacy protection capabilities in future research.

ACKNOWLEDGEMENT

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] Saner, H., Schütz, N., Botros, A., Urwyler, P., Buluschek, P., Du Pasquier, G., Nef, T. (2020). Potential of ambient sensor systems for early detection of health problems in older adults. Frontiers in Cardiovascular Medicine, 7: 110. https://doi.org/10.3389/fcvm.2020.00110

[2] Ferraris, C., Votta, V., Nerino, R., Chimienti, A., Priano, L., Mauro, A. (2024). At-home assessment of postural stability in parkinson's disease: A vision-based approach. Journal of Ambient Intelligence and Humanized Computing, 15(5): 2765-2778. https://doi.org/10.1007/s12652-023-04553-5

[3] Maitre, J., Bouchard, K., Gaboury, S. (2023). Data filtering and deep learning for enhanced human activity recognition from UWB radars. Journal of Ambient Intelligence and Humanized Computing, 14(6): 7845-7856. https://doi.org/10.1007/s12652-023-04596-8

[4] Abbe, J.R., O'Keeffe, C. (2021). Continuous video monitoring: Implementation strategies for safe patient care and identified best practices. Journal of Nursing Care Quality, 36(2): 137-142. https://doi.org/10.1097/NCQ.0000000000000502

[5] Diao, J.A., Marwaha, J.S., Kvedar, J.C. (2022). Video-based physiologic monitoring: Promising applications for the ICU and beyond. NPJ Digital Medicine, 5(1): 26. https://doi.org/10.1038/s41746-022-00575-z

[6] Herfandi, H., Sitanggang, O.S., Nasution, M R.A., Nguyen, H., Jang, Y.M. (2024). Real-time patient indoor health monitoring and location tracking with optical camera communications on the internet of medical things. Applied Sciences, 14(3): 1153. https://doi.org/10.3390/app14031153

[7] Wu, W., Li, B., Chen, L., Zhu, X., Zhang, C. (2017). K-ary tree hashing for fast graph classification. IEEE Transactions on Knowledge and Data Engineering, 30(5): 936-949. https://doi.org/10.1109/TKDE.2017.2782278

[8] Chelloug, S.A., Ashfaq, H., Alsuhibany, S.A., Shorfuzzaman, M., Alsufyani, A., Jalal, A., Park, J. (2023). Real objects understanding using 3D haptic virtual reality for e-learning education. Computers, Materials & Continua, 74(1): 1607-1624. https://doi.org/10.32604/cmc.2023.032245

[9] Folly, K.A. (2019). A short survey on population-based incremental learning algorithm. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, pp. 339-344. https://doi.org/10.1109/SSCI44817.2019.9002858

[10] Tayyab, M., Alateyah, S.A., Alnusayri, M., Alatiyyah, M., AlHammadi, D.A., Jalal, A., Liu, H. (2025). A hybrid approach for sports activity recognition using key body descriptors and hybrid deep learning classifier. Sensors, 25(2): 441. https://doi.org/10.3390/s25020441

[11] Ahmed, A., Jalal, A., Rafique, A.A. (2019). Salient segmentation based object detection and recognition using hybrid genetic transform. In 2019 International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, pp. 203-208. https://doi.org/10.1109/ICAEM.2019.8853834

[12] Waheed, M., Javeed, M., Jalal, A. (2021). A novel deep learning model for understanding two-person interactions using depth sensors. In 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, pp. 1-8. https://doi.org/10.1109/ICIC53490.2021.9692946

[13] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C. (2019). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2684-2701. https://doi.org/10.1109/TPAMI.2019.2916873

[14] Jang, J., Kim, D., Park, C., Jang, M., Lee, J., Kim, J. (2020). ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, pp. 10990-10997. https://doi.org/10.1109/IROS45743.2020.9341160

[15] Wang, P., Li, W., Wan, J., Ogunbona, P., Liu, X. (2018). Cooperative training of deep aggregation networks for RGB-D action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 7404-7411. https://doi.org/10.1609/aaai.v32i1.12228

[16] Xu, Y., Cheng, J., Wang, L., Xia, H., Liu, F., Tao, D. (2018). Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. IEEE Signal Processing Letters, 25(7): 1044-1048. https://doi.org/10.1109/LSP.2018.2841649

[17] Li, C., Zhong, Q., Xie, D., Pu, S. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, pp. 786-792. https://doi.org/10.24963/ijcai.2018/109

[18] Li, C., Xie, C., Zhang, B., Han, J., Zhen, X., Chen, J. (2021). Memory attention networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems, 33(9): 4800-4814. https://doi.org/10.1109/TNNLS.2021.3061115

[19] Li, C., Zhong, Q., Xie, D., Pu, S. (2017). Skeleton-based action recognition with convolutional neural networks. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, pp. 597-600. https://doi.org/10.1109/LSP.2017.2678539

[20] Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 7444-7452. https://doi.org/10.1609/aaai.v32i1.12328

NOMENCLATURE

PD	Parkinson’s Disease
WOA	Whale Optimization Algorithm
MEMM	Maximum Entropy Markov Model