


Article

Modeling Two-Person Segmentation and Locomotion for Stereoscopic Action Identification: A Sustainable Video Surveillance System

Nida Khalid ¹, Munkhjargal Gochoo ², Ahmad Jalal ¹ and Kibum Kim ^{3,*} 

¹ Department of Computer Science, Air University, Islamabad 44000, Pakistan; 190115@students.au.edu.pk (N.K.); ahmadjalal@mail.au.edu.pk (A.J.)

² Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain 15551, UAE; mgochoo@uaeu.ac.ae

³ Department of Human-Computer Interaction, Hanyang University, Ansan 15588, Korea

* Correspondence: kikum@hanyang.ac.kr

Abstract: Due to the constantly increasing demand for automatic tracking and recognition systems, there is a need for more proficient, intelligent and sustainable human activity tracking. The main purpose of this study is to develop an accurate and sustainable human action tracking system that is capable of error-free identification of human movements irrespective of the environment in which those actions are performed. Therefore, in this paper we propose a stereoscopic Human Action Recognition (HAR) system based on the fusion of RGB (red, green, blue) and depth sensors. These sensors give an extra depth of information which enables the three-dimensional (3D) tracking of each and every movement performed by humans. Human actions are tracked according to four features, namely, (1) geodesic distance; (2) 3D Cartesian-plane features; (3) joints Motion Capture (MOCAP) features and (4) way-points trajectory generation. In order to represent these features in an optimized form, Particle Swarm Optimization (PSO) is applied. After optimization, a neuro-fuzzy classifier is used for classification and recognition. Extensive experimentation is performed on three challenging datasets: A Nanyang Technological University (NTU) RGB+D dataset; a UoL (University of Lincoln) 3D social activity dataset and a Collective Activity Dataset (CAD). Evaluation experiments on the proposed system proved that a fusion of vision sensors along with our unique features is an efficient approach towards developing a robust HAR system, having achieved a mean accuracy of 93.5% with the NTU RGB+D dataset, 92.2% with the UoL dataset and 89.6% with the Collective Activity dataset. The developed system can play a significant role in many computer vision-based applications, such as intelligent homes, offices and hospitals, and surveillance systems.



Citation: Khalid, N.; Gochoo, M.; Jalal, A.; Kim, K. Modeling Two-Person Segmentation and Locomotion for Stereoscopic Action Identification: A Sustainable Video Surveillance System. *Sustainability* **2021**, *13*, 970. <https://doi.org/10.3390/su13020970>

Received: 15 December 2020

Accepted: 14 January 2021

Published: 19 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: geodesic distance; human action recognition; human locomotion; neuro-fuzzy classifier; particle swarm optimization; RGB-D sensors; trajectory features

1. Introduction

Vision-based Human-Computer Interaction (HCI) is a broad field covering many areas of computer vision, such as human action tracking, face recognition, gesture recognition, human-robot interaction and many more [1]. In our proposed methodology we focused on vision-based human motion analysis and representation for Human Action Recognition (HAR). HAR can be precisely defined as tracking the motion of each and every observable body part involved in performing human actions and identifying the activities performed by humans [2]. HAR further subdivides into atomic actions, two person interactions, multiperson interactions, human-object interactions and human-robot interactions, etc. [3,4]. However, in the proposed system, we focused on two-person interactions, i.e., human-human interaction. Extensive research has been carried out in the field of vision-based HAR systems but there remains a need for an adaptive and sustainable

HAR system that is effective regardless of the environment [5–11]. The main aim of this research work is to develop a novel, reliable and sustainable vision-based HAR system based on our unique set of features. To this end, we propose a HAR system that is highly adaptive to changing environments and variations in available light.

Activity recognition by digital monitoring systems is useful in many daily life applications such as video indexing and retrieval, virtual worlds and surveillance systems installed in houses, hospitals and public areas [12–14]. An automatic, efficient and robust surveillance system is imperative because of the elevated crime rates all over the world [15,16]. Our system is capable of detecting and identifying anomalous actions in its field of vision such as fighting, punching, pushing and kicking, etc. Moreover, the proposed system can also be used in healthcare work in hospitals, in homes for the care of the elderly and in general patient monitoring [17,18]. It can also be used in rehabilitation centers and children’s care centers [19]. Due to the wide variety of applications for human motion tracking in daily life [20], we are motivated to develop a versatile, adaptive and reliable HAR system.

In order to develop an effective HAR system, the first step is to design a sound method of pertinent data acquisition [21]. The performance of the system is wholly dependent on the quality of the data acquired by the system’s devices. If these data are compromised by inadequate processing, the reliability and efficiency of the whole system and its outcomes will be compromised. [22]. Many methods are used to acquire data. These include RGB (red, green, blue) cameras, stereoscopic RGB-D (RGB-depth) sensors, wearable marker sensors [23–25]. Many HAR systems have been proposed that consist only of RGB information [26]. These systems cannot perform efficiently in environments with crowded backgrounds or brightness variations [27,28]. So, in recent years, three-dimensional (3D) RGB-D sensors which cost-effectively tackle the limitations of RGB cameras have been developed [29–31]. These RGB-D sensors include extra stereoscopic vision which helps eradicate the confusion between foreground actors and background objects [32–34].

Inspired by various applications of vision sensors in surveillance, we propose an efficient, adaptive and sustainable system based on the fusion of RGB and depth information. We use four unique features for recording each and every motion performed by humans. We propose two full-body features, namely, geodesic distance and 3D Cartesian-plane features plus two skeletal joints-based features, namely, way-point trajectory generation and joints MOCAP (motion capture) features. Full body- and skeletal joint-based feature descriptors are combined and optimized via Particle Swarm Optimization (PSO). Optimized feature descriptors are then used to recognize human activities with a Neuro-Fuzzy Classifier (NFC). Techniques used for each phase of this research work are listed in Table 1.

Table 1. Phases of the proposed Human Action Recognition (HAR) system.

Phase	Techniques	Description
Silhouette Segmentation	Background subtraction and Morphological operations	Efficient silhouette segmentation is executed on both RGB and depth frames via frame differencing and morphological operations, respectively.
Feature Extraction	Geodesic distance	Geodesic maps are generated based on the shortest distance from the center points of two human silhouettes towards the outer boundary.
	3D Cartesian plane	RGB and depth silhouettes are projected in an altered Cartesian plane to represent features from different views.
	Joints MOCAP	The geometrical properties of each human joint are taken to record human locomotion.
	Way-point trajectory generation	The shape and motion information of each way-point trajectory generated from subsets of skeletal joints are recorded with each changing frame. Inter-silhouette and intra-silhouette trajectory generation is implemented.

Table 1. Cont.

Phase	Techniques	Description
Optimization	Particle Swarm Optimization (PSO)	The tracked motion descriptors of each action class are represented in an optimized form via a PSO algorithm. PSO is used as a feature selection algorithm to remove redundant features and to increase classification performance. For an efficient time and space computation, PSO is applied for feature selection and size reduction.
Classification	Neuro-Fuzzy Classifier(NFC)	An extensive experimental evaluation with challenging RGB-D datasets is performed with NFC. System validation is proved with an altered number of membership functions.

The main contribution of each phase of this research work is as follows:

1. Segmentation of both RGB and depth silhouettes is achieved via background subtraction and a series of morphological operations.
2. Feature extraction from full human silhouettes is performed via geodesic maps and 3D Cartesian planes. Whereas, feature extraction from skeletal joints is performed via way-points trajectories and orientation angles. These features record each movement performed by two interacting human silhouettes.
3. Feature selection is performed on the combined feature descriptors of four proposed features via PSO.
4. Extensive experimentation is performed to prove systems' validity via classification with a neuro-fuzzy inference system, effects of different numbers of membership functions, sensitivity specificity and error measures.

In the rest of the paper, Section 2 presents related work in the field of HAR. Section 3 provides the details of each phase of proposed methodology. Section 4 explains the experiments performed and their generated results. At the end, the proposed research work is concluded in Section 5.

2. Literature Review

This section describes different methodologies that have been adopted in recent years for human action tracking and recognition [35]. Vision-based human activity tracking can be subdivided at different stages: (1) first, on the basis of the source of input; (2) second, at the features extraction and recognition stage. An extensive review of related work and preceding methodologies is given in this section.

2.1. Devices for HAR Data Acquisition

On the basis of data acquisition, vision-based HAR systems are divided into two categories: (1) RGB-based HAR and (2) RGB-D-based HAR.

2.1.1. RGB-Based HAR Systems

Many HAR systems that only work on RGB datasets for experimentation and validation have been proposed in recent years [36–38]. Table 2 presents summary details of authors, datasets and the research work relevant to these systems.

Table 2. RGB-based HAR methods.

Authors	RGB Datasets	Methodology	Classification Results
Xiaobin et al. [39]	CAD Choi's Dataset	A learning-based methodology was proposed in which the interaction matrix of each activity was represented. A multitask interaction response (MIR) was computed for each class separately.	Support Vector Machine (SVM) as baseline and MIR was used for classification. Experiments proved the validation of the system. The mean accuracy achieved was 83.3% with CAD and 80.3% with Choi's dataset.
Qing et al. [40]	UT-Interaction dataset	A global feature-based approach was presented where a combination of Gaussian time-phase features was used. Multifeature fusion was performed with ResNet (Residual Network) and parallel inception.	Experiments were performed via SVM with Kalman tracking. An overall recognition rate of 91.7% was achieved with UT-Interaction dataset.
Amir et al. [41]	UCF YouTube action dataset and an IM-DailyRGBEvents	Spatiotemporal multidimensional features were used for both body part detection and action recognition.	Better system performance was achieved with Maximum entropy Markov model and activity recognition rates of 89.09% with the UCF dataset and 88.26% with the IM Event dataset were achieved.
Kishore et al. [42]	UCF 50 UCF 11 HMDB51 KTH	Scene context approach was applied. Motion features were applied along with the fusion of descriptors at early and late stages.	They achieved a performance rate of 87.19% with UCF11, 76.90% with UCF50, 27.20% with HMDB51 and 89.79% with KTH dataset via SVM.
Mahmood et al. [43]	UT-Interaction dataset	After identifying the starting and ending frame, spatiotemporal features were extracted from human key body points and from full body silhouettes, as well.	With Artificial Neural Network (ANN) and one-third training validation test, better recognition was achieved in six classes with an average accuracy of 83.5% with Set 1 and 72.5% with Set 2.

2.1.2. RGB-D-Based HAR Systems

Many HAR systems are based on datasets that combine both RGB color and depth information [44]. RGB-D sensors also provide skeletal information [45]. Table 3 shows the details of authors, datasets and research work based on RGB-D sensors, using the combination of both RGB and depth images.

Table 3. HAR systems based on RGB-D sensors.

Authors	RGB-D Datasets	Methodology	Classification Results
Rawya et al. [46]	MSR-Daily Activity 3D dataset and Online RGBD action dataset	Spatio-temporal features were extracted using a Bag-of-Features (BoF) approach. Points of interest were detected, and motion history images were created to perform this research work.	By using K-means clustering and multiclass SVM, experimental results on these publicly available datasets proved the efficacy of the system with average recognition rates of 91.1% with the MSR dataset and 92.8% with the Online RGBD dataset.
Jalal et al. [47]	MSRAAction3D	The two types of features that were extracted from human silhouettes were shape and motion features using temporal continuity constraints.	As a result of experimentation on two challenging datasets with Hidden Markov Model (HMM), this approach proved to be effective in HAR with a mean recognition rate of 82.10%.
Xiaofei et al. [48]	SBU Kinect interaction dataset UT-Interaction dataset	In this research work, interaction was divided into three stages, namely, start, middle and end. Probability fusion-based features were extracted.	Extensive experiments via HMM proved the efficacy of the system with 91.7% accuracy with the SBU and 80% with the UT-Interaction dataset.

Table 3. Cont.

Authors	RGB-D Datasets	Methodology	Classification Results
Meng et al. [49]	NTU RGB+D, SBU Kinect interaction dataset and M2I dataset	With the help of skeletal and depth data, pairwise feature learning was introduced. Relative movement between body parts was extracted.	Linear SVM was used as a classifier. All activity classes were recognized with higher accuracy rates than many state-of-the-art systems.
Claudio et al. [50]	UoL 3D social activity dataset	Determining social activity via statistical and geometrical features such as skeletal positions and motion features was proposed in this research work.	The proposed novel features with HMM proved to be very effective in social interaction recognition with a mean accuracy of 85.5%.

2.2. Division on the Basis of Feature Extraction and Recognition

Some researchers have applied hand-crafted features and machine learning methods for feature extraction and recognition, respectively, in vision-based systems. On the other hand, some researchers have applied deep learning approaches for both feature learning and activity recognition [51]. So, on the basis of feature extraction methods, HAR can be divided into two methodologies: (1) machine learning-based HAR systems and (2) deep learning-based HAR systems.

2.2.1. Machine Learning-Based HAR Systems

In this section, HAR systems based on machine learning approaches (supervised, unsupervised and semi-supervised) are presented. Table 4 shows details of authors, datasets and research work based on hand-crafted features and machine learning-based approaches.

Table 4. Machine learning-based HAR systems.

Authors	Datasets	Methodology	Results via Machine Learning
Yu et al. [52]	BIT-Interaction Dataset UT-Interaction dataset	In order to recognize interdependencies between two-person interaction, local body parts and global large-scale features were presented. Adaboost algorithm was adopted to find 3D body parts.	Linear SVM was used for classification. After testing on two benchmark datasets, the average accuracy with the BIT dataset was 82.03% and with the UT dataset it was 85%.
Yanli et al. [53]	Self-annotated CR-UESTC dataset and SBU Kinect interaction dataset	For interaction recognition a Contrastive Feature Description Model (CFDM) was proposed. Intr-a and inter-skeleton were represented.	The CFDM approach proved to be very effective with an action recognition rate of 87.6% with the CR-UESTC and 89.4% with the SBU dataset via Binary SVM.
T Subetha et al. [54]	SBU Kinect interaction dataset	Features were extracted via a Histogram of Oriented Gradients (HOG) and pyramidal approach. Constrained Weighted Dynamic Time Warping (CWDTW) was used in this work.	K-means clustering with CWDTW was used for classification. A very high recognition rate of 90.8% was achieved with this new approach towards action recognition.
Jalal et al. [55]	IM-DailyDepthActivity dataset MSRAAction3D	Spatiotemporal features of human joints and frame differentiation features were extracted.	Classification was performed with HMM. Results of the proposed system were validated via experimentation with an accuracy rate of 88.9% and 66.70% over two datasets.
Thien et al. [56]	SBU Kinect interaction dataset	Joint features were extracted via Pachinko Allocation Model. Both joint motion and distance feature were extracted.	This method outperformed many state-of-the-art methods via Binary Tree as a classifier. A mean recognition rate of 90.3% was achieved.

2.2.2. Deep Learning-Based HAR Systems

In some HAR systems, features are learned and actions are recognized automatically through deep learning models. Table 5 shows details of authors, datasets and research work based on feature learning and activity recognition via deep learning-based approaches.

Table 5. Deep learning-based HAR systems.

Authors	Datasets	Methodology	Results via Deep Learning
Amir et al. [57]	NTU RGB+D, 3D Action Pair, MSR Daily activity and Online RGBD	A new deep learning model for shared specific factorization features was introduced in this research work. Sparsity learning was introduced for classification. Two experimental settings were adopted to show the effectiveness of the proposed methodology.	The recognition rate of actions involving a single person is as high as 100% with the 3D action pair dataset, while the recognition rate with datasets that involve two-person interactions, i.e., NTU RGB+D, is 74.9%. Overall good performance is achieved with each of five datasets.
Xiangbo et al. [58]	BIT-Interaction Dataset UT-Interaction dataset	In order to capture changes in interactions between two persons over time, Concurrent Long Short-Term Memory (Co-LSTM) was proposed. Information about human action was stored in sub-memory units.	Co-LSTM produced a superior performance with both RGB datasets. A recognition rate of 92.8% was achieved with the UT-Interaction dataset and 95% with the BIT dataset was achieved.
Wentao et al. [59]	SBU Kinect interaction dataset, HDM05, Berkeley MHAD	Skeletal temporal features were extracted automatically via a Long Short-Term Memory (LSTM) network. Co-occurrence features were extracted. A novel dropout methodology was proposed.	Deep LSTM results in an average accuracy rate of 90.4% with SBU, 97.25% with HDM05 and 81.05% with the CMU dataset.
Yong et al. [60]	MSR action 3D Dataset, Berkeley MHAD and Motion Capture Dataset hdm05	Temporal long-term contextual information was learned via Hierarchical RNN (HRNN). In this approach, the human skeleton is divided into five subparts. Each subpart is separately fed into five different subnetworks.	Five different experimental settings were used with HRNN. Through experimentation, high recognition performance rate was achieved with a number of datasets with each experimental setting.
Xiangbo shu et al. [61]	CAD BIT UT	In order to overcome the limitation of LSTM in capturing changes in human interactions over time, Hierarchical LSTM (HLSTM) was used in this research work. Groups of people were observed to monitor human interactions.	Comparisons with four baselines and state-of-the-art methods were performed. The validity of the novel approach presented in this method was proved by the high accuracy achieved with three datasets.

3. Materials and Methods

A comprehensive description of each phase is given in this section. It is represented in the following phases:

- In the preprocessing phase, human silhouettes of each RGB and depth image are segmented from their backgrounds.
- In the feature descriptor generation stage, four features (geodesic distance, 3D Cartesian plane, way-point trajectory and joints MOCAP) are mined from each RGB and depth image and thus, feature descriptors are generated.
- The optimization phase results in an optimized representation of feature descriptors via PSO.
- In the final stage, each human action is classified via a neuro-fuzzy inference system.

Figure 1 shows the flow diagram of the proposed human action surveillance system.

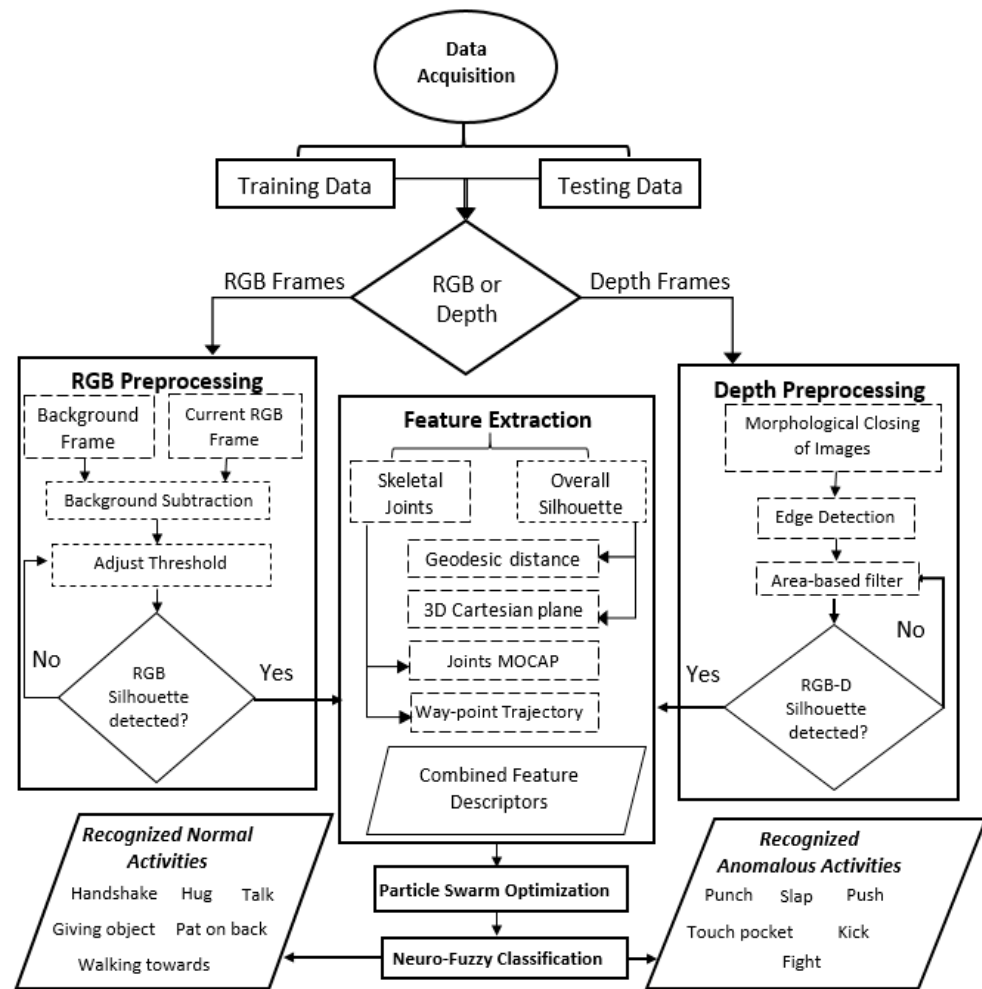


Figure 1. Proposed surveillance system architecture.

3.1. Foreground Extraction

Prior to any processing, all RGB and depth image sequences are subjected to image normalization technique to improve the quality of image [62,63]. Image contrast is adjusted, and intensity values are uniformly distributed through the entire image via histogram equalization [64,65]. After that, in order to remove noise from the image, a median filter is applied in which pixels are replaced by a median of neighboring pixels [66,67]. The most important step in any HAR system is to define and mine Regions of Interest (ROI) [68]. In our work, an ROI consists of two persons involved in an interaction in RGB-D images. These ROIs are first segmented from their background. Methods adopted to segment human silhouettes are separately given in the following subsection.

3.1.1. Background Subtraction

RGB silhouette extraction of all three datasets is achieved through a background subtraction method [69]. A frame difference technique is used in which current frames of each interaction class are subtracted from a background frame [70]. Pixels of the current frame $I(t)$ at time t , denoted by $P[I(t)]$, are subtracted from pixels of a background frame denoted by $P[B]$, as given in Equation (1):

$$P[F(t)] = P[I(t)] - P[B] \quad (1)$$

where $P[F(t)]$ is the frame obtained after subtraction. The subtracted image, i.e., the image containing human silhouettes is further processed for better foreground detection through specifying a threshold value T as given through Equation (2):

$$|P[F(t)] - P[F(t+1)]| > T \quad (2)$$

The T value is automatically selected for each subtracted image via Otsu's thresholding method [71]. In this method the subtracted frame is first converted to a grayscale image and then the best T value (a best value which differentiates the black background pixels and the white foreground pixels) is obtained through an iterative process. This T value is then used to convert the subtracted grayscale image to binary image, and then a binary silhouette is obtained as a result. Examples of RGB silhouette extraction of NTU RGB+D and UoL datasets are shown in Figure 2.

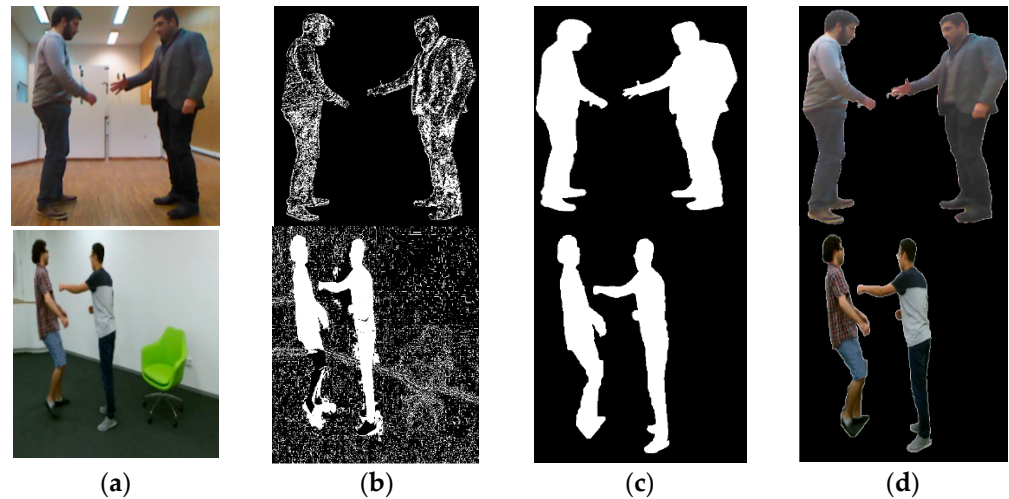


Figure 2. Background subtraction over RGB image sequences of UoL and NTU RGB+D dataset: (a) original image; (b) subtracted frame at $T = 1.5$; (c) binary silhouette obtained after adjusting T ; (d) RGB silhouette.

3.1.2. Morphological Operations

In order to extract depth silhouettes, first, threshold-based segmentation [72] is used to obtain a binary image from the original image. These segmented images are closed morphologically using binary dilation followed by a binary erosion operation [73]. Thus, binary dilation works through adding pixels to human edges while erosion works by removing boundary pixels. Binary dilation and erosion are shown through Equations (3) and (4), respectively:

$$A \oplus B = \left\{ z \mid (\hat{B})_z \cap A \neq \phi \right\} \quad (3)$$

$$A - B = \{ z \mid (B)_z \subseteq A \} \quad (4)$$

where z is a set of pixel locations where structuring element B and its reflection \hat{B} joins with pixels of foreground element A during translation to z . In this way, only the shape of the main objects in an image is maintained. Finally, Canny edge detection is applied to separate foreground pixels from the background. After the detection of the edges, smaller area objects are removed from the binary image which results in human silhouette detection. The silhouette segmentation of the depth images forms the UoL dataset is shown in Figure 3.

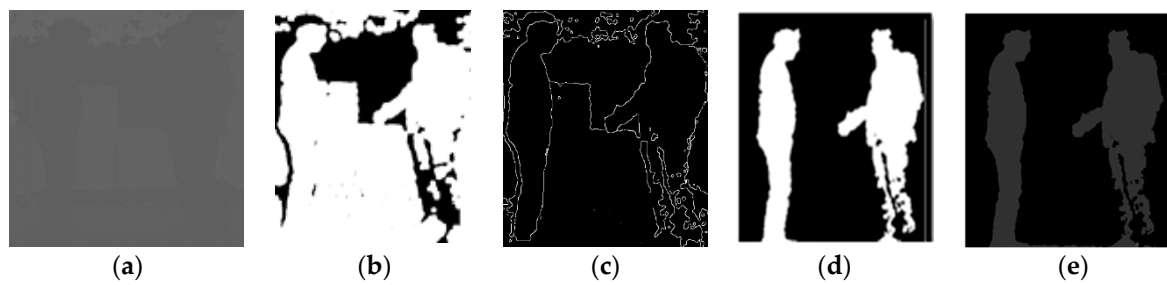


Figure 3. Morphological operations on depth images. (a) Original image; (b) image after erosion and dilation; (c) edge detection; (d) binary silhouette and (e) depth silhouette.

3.2. Feature Descriptors Mining

Segmented RGB-D silhouettes are then used for feature mining. Unique features are extracted from full silhouettes and from the skeleton joints. Two features, namely, geodesic and 3D Cartesian-plane, are applied over full human silhouettes. Two features, namely, way-point trajectory generation and joints MOCAP are applied to the skeleton joints. Each feature is explained in detail in the following subsections.

3.2.1. Geodesic Distance

In this feature, actions between two interacting humans are represented via geodesic wave maps. These maps are generated by calculating the geodesic distance (the smallest distance) which is found by a Fast Marching Algorithm (FMA) [74]. Firstly, the center point s of the two human silhouettes is located and given a distance value $d(s) = 0$. Point s is the starting point and it is marked as a visited point. All the other pixel points p on human silhouettes are given a distance value $d(p) = \infty$ and are marked as unvisited. Each unvisited point p is taken from the neighbors of s and its distance from s is measured. In this way, each neighboring pixel is taken in each iteration until all the pixel points are marked as visited. The distance calculated at each iteration is compared with the distance of each previous iteration. A priority queue is generated where the shortest distances are given priority [75]. An update in distance is defined as:

$$d = \begin{cases} \frac{d_x + d_y + \sqrt{\Delta}}{2} & \text{when } \Delta \geq 0 \\ \min(d_x, d_y) + w & \text{otherwise} \end{cases} \quad (5)$$

$$\Delta = 2w^2 - (d_x - d_y)^2 \quad (6)$$

where $d_x = \min(D_{k+1, \ell}, D_{k-1, \ell})$ and $d_y = \min(D_{k, \ell+1}, D_{k, \ell-1})$. Figure 4 demonstrates the wave propagation of the geodesic distance via FMA.

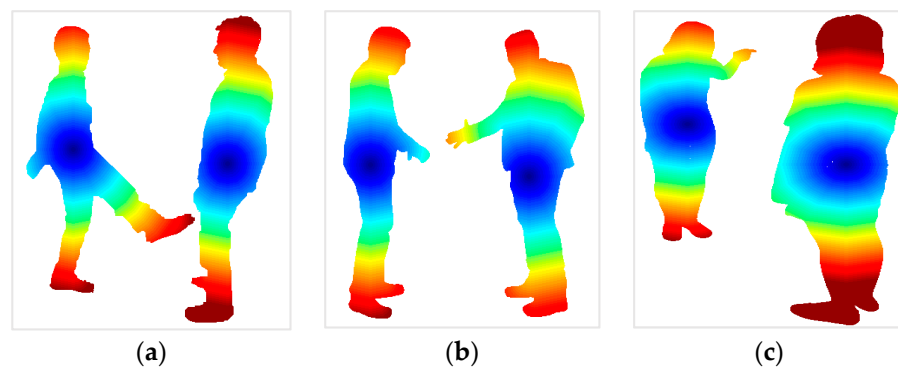


Figure 4. Wave propagation of the geodesic distance via FMA on NTR RGB-D classes of: (a) kicking; (b) shaking hands and (c) pointing finger.

3.2.2. 3D Cartesian-Plane Features

In this feature, shape as well as motion information from the two human silhouettes are taken [76]. 3D shapes of the segmented RGB-D silhouettes are created by projecting them onto a 3D Cartesian plane. Motion information from the two interacting persons is retained via a frame differencing technique that is used to take the differences in 3D shapes created between two consecutive frames. After creating 3D shapes, Histogram of Oriented Gradients (HOG) is applied to extract unique features. In order to apply HOG [77], all images are first preprocessed to make their dimensions 64×128 pixels. Bounding boxes are drawn around each human in the image and the gradient of each human in the image is calculated separately. These features measure both position and the direction of changes along each pixel. Magnitude is given as:

$$g = \sqrt{g_x^2 + g_y^2} \quad (7)$$

where g is the gradient, i.e., the change in x and y directions for each pixel and the directional angle. Pseudo code for full body feature extraction techniques (Geodesic and 3D Cartesian plane) is given in Algorithm 1. The direction of change is shown through red marks on 3D shapes in Figure 5.

Algorithm 1 Pseudo code of feature extraction from full silhouette

```

1: Input: Segmented RGB and depth silhouettes frames ( $f_1, f_2, \dots, f_n$ )
2: Output: Full body feature descriptors ( $V_1, V_2, \dots, V_n$ ) //where  $n$  is total number of frames
//Geodesic distance features//
3: for  $i = 1: n$ 
4:   mark center pixel of both human silhouettes as visited and initialize a distance equal to
   zero
5:   as  $d(x_0) = 0$ 
6:   for all the other points on human silhouette that are unvisited initialize  $d(x) = \infty$ 
7:   initialize a queue  $Q = X$  for unvisited points
8:   while  $Q \neq \emptyset$ 
9:     Step 1: Locate a vertex with a smallest distance  $d$  as  $x = \underset{x \in Q}{\operatorname{argmin}} d(x)$ 
10:    Step 2: For each neighboring unvisited vertex  $x' \in N(x) \cap Q$ 
11:       $d(x') = \min\{d(x'), d(x) + L(x, x')\}$ 
12:    Step 3: Remove  $x$  from  $Q$ 
13:   end while
14:   Return distance vector  $d(x_i) = d_L(x_0, x_i)$ 
//3D Cartesian-plane features//
15:   project each frame  $f$  in  $F$  on 3D Cartesian plane  $yz$ 
16:   for each projected 3D frame subtract current frame  $f_{yz}$  from successor frame  $(f + 1)_{yz}$  to get
   differential frame as  $\text{diff} \leftarrow (f + 1)_{yz} - f_{yz}$ 
17:   end for
18:   for each differential frame  $\text{diff}$  calculate HOG vector from gradient, magnitude, orientation
   and histogram as:
19:     Gradient ( $\text{diff}, \text{grad}_x, \text{grad}_y$ )
20:     Magnitude ( $\text{grad}_x, \text{grad}_y, \text{mag}$ )
21:     Orientation ( $\text{grad}_x, \text{grad}_y, \text{orient}$ )
22:     Histogram ( $\text{orient}, \text{mag}, \text{hist}$ )
23:     Normalization ( $\text{hist}, \text{normhist}$ )
24:     HOG vector  $\leftarrow \text{normhist}$ 
25:   end for
26:   compute full body feature descriptor  $V$  for each frame  $f$  as  $V \leftarrow$  concatenate (distance vector,
   HOG vector)
27:   end for
28:   return Full body feature descriptors ( $V_1, V_2, \dots, V_n$ )

```

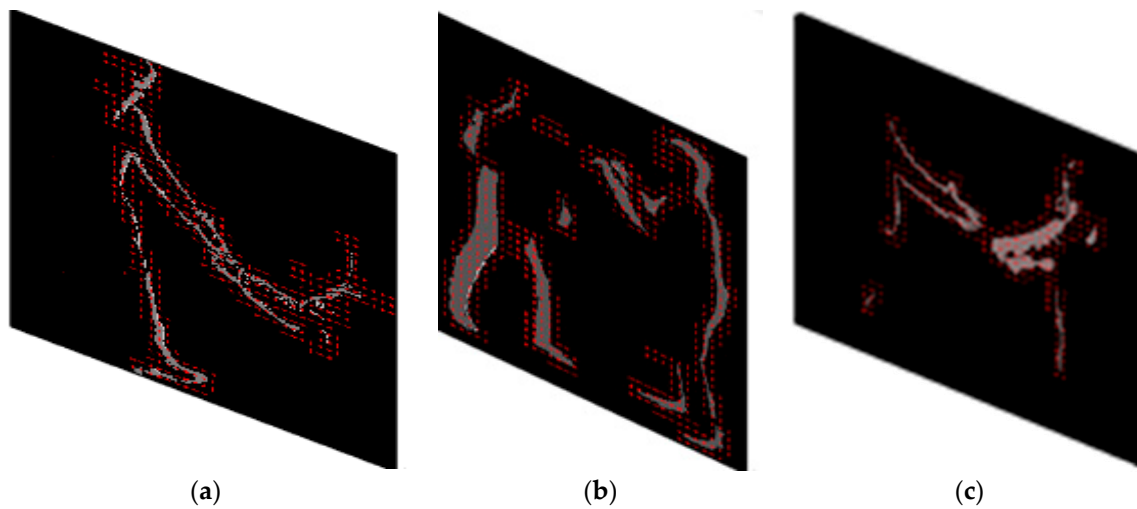


Figure 5. HOG feature extraction on 3D shapes created over action classes: (a) helping to stand up; (b) fighting and (c) shaking hands.

3.2.3. Joints MOCAP

Joints MOCAP features are used to track the movements of human joints because joints are the most significant parts involved in human movements [78]. We represent the skeleton as $S = \{J_k | k = 1, 2, \dots, n\}$ where n consists of sixteen major human joints, namely, head, neck, right shoulder, left shoulder, right elbow, left elbow, right hand, left hand, spine-mid, spine-base, left hip, right hip, right knee, left knee, right foot and left foot. A joint is represented as $J_k = (x_j, y_j)$ which specifies the coordinates location in RGB-D silhouettes. After locating all the joint positions in both human silhouettes, geometrical properties are measured between joint J_i and the rest of the joints J_k where $k \neq i$. A total of thirty-two joints (sixteen per person in an interaction) are tracked with each changing frame with time t . Two types of angular measurements that are taken to track skeletal joint movements with each changing frame are:

- **Upper body Angles:** In this type, human motion caused by the rotation of the spine's mid joint with respect to (w.r.t) all the upper body joints, namely, head, neck, left shoulder, right shoulder, left elbow, right elbow, left hand and right hand, are tracked. Four upper body angles per person, i.e., eight per frame, are tracked. The angle of the tangent between the spine's mid joint and two other joints taken from a joints set S is calculated. The inverse tangent is found by taking a dot product of two lines v_1 and v_2 , as represented by Equation (8):

$$\theta = \tan^{-1} \frac{v_1 \cdot v_2}{|v_1| |v_2|} \quad (8)$$

- **Lower body Angles:** In this type, the angle of tangent from the spine-base joint to all the lower body joints, left hip, right hip, left knee, right knee, left foot and right foot, are calculated. Three lower body angles per person, i.e., six per frame, are tracked. Figure 6 depicts angle formation in the upper body and lower body.

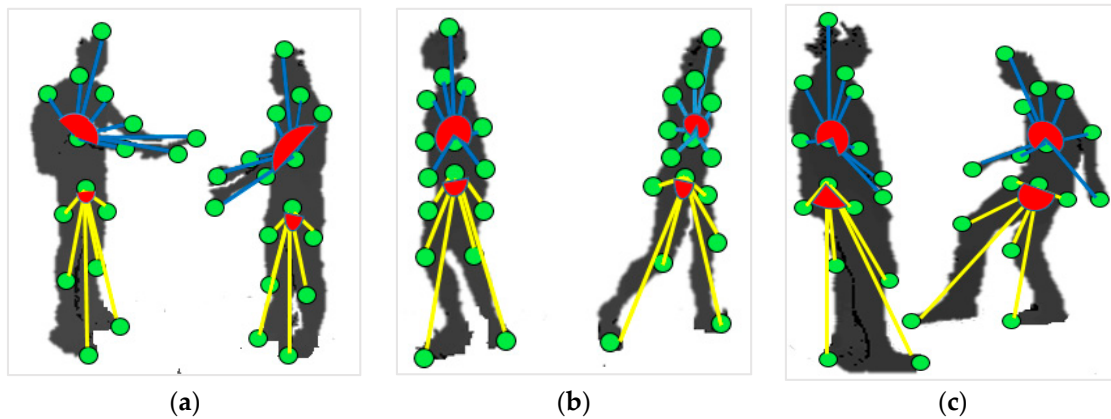


Figure 6. Eight upper body angles (shown with blue-colored lines) and six lower body angles (shown with yellow-colored lines) formation over (a) giving object; (b) walking apart and (c) kicking action classes.

3.2.4. Way-Point Trajectory Generation

A lot of research has been done on dense trajectories [79] and localized trajectories [80]. We, however, introduced the concept of new intra-silhouette and inter-silhouette localized way-point trajectories. In both of these trajectory types a subset S , containing a different number of human joints, is given as a way-point to generate trajectories. Curve trajectories are generated at a specified orientation. First of all, two joints sets, J_1 and J_2 , are created. Where $J_1 = \{j_1, j_2, \dots, j_n\}$ is constructed from n number, i.e., sixteen joints (head, neck, right shoulder, left shoulder, right elbow, left elbow, right hand, left hand, spine-mid, spine-base, left hip, right hip, right knee, left knee, right foot and left foot) of the first (left) person in an interaction. $J_2 = \{j_1, j_2, \dots, j_m\}$ is constructed from m number, i.e., from sixteen joints of the second (right) person in an interaction. In intra-silhouette trajectory generation, a subset S consists of all the way-points from a single joint set, i.e., either J_1 or J_2 . On the other hand, in inter-silhouette trajectory generation, a subset S consists of way-points from both joint sets J_1 and J_2 . Table 6 shows a detailed description of each intra-silhouette and inter-silhouette way-point trajectory cluttered around human joints.

After construction of all the trajectories over human joints, two types of feature are extracted from each trajectory [81]. Shape descriptors are described by calculating changes in displacement of the length T of the trajectory over time t . These changes are measured along the coordinate positions x and y of the joints with each changing frame given as $\Delta l_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. The normalized displacement vector is given as:

$$D_{x,y} = \frac{(\Delta l_1, \Delta l_2, \dots, \Delta l_{T-1})}{\sum_{j=1}^{T-1} \|\Delta l_j\|} \quad (9)$$

Motion descriptors are computed by tracking changes in velocity w.r.t time. Velocity is measured by changes in position (i.e., displacement) of trajectories over time t . So, a first- and second-order derivative of the position of trajectory (coordinates) is taken as x'_t, y'_t, x''_t and x''_t respectively. The final curvature C over space time coordinates x and y is defined as:

$$C_t = \frac{x'_t y''_t - y'_t x''_t}{(x_t'^2 + y_t'^2 + 1)^{3/2}} \quad (10)$$

Pseudo code of feature extraction from skeletal joints is given in Algorithm 2. Figure 7 displays curved intra-silhouette and inter-silhouette way-point trajectories over human joints.

Table 6. Intra-silhouette way-points trajectory generation.

No. of Trajectories	No. of Way-Points	Subsets of Joints	
		Intra-Silhouette	Inter-Silhouette
1	Three	{H ¹ , N ¹ , SM ¹ }	{H ¹ , N ¹ , H ² }
2		{H ¹ , N ¹ , SB ¹ }	{H ² , N ² , H ¹ }
3		{H ² , N ² , SM ² }	{RS ¹ , LS ¹ , RE ² }
4		{H ² , N ² , SB ² }	{RS ² , LS ² , RE ¹ }
5		{RS ¹ , RE ¹ , RH ¹ }	{RH ¹ , LH ¹ , LE ² }
6		{LS ¹ , LE ¹ , LH ¹ }	{RH ² , LH ² , LE ¹ }
7		{RS ² , RE ² , RH ² }	{SM ¹ , SB ¹ , SM ² }
8		{LS ² , LE ² , LH ² }	{SM ² , SB ² , SM ¹ }
9		{RHP ¹ , RK ¹ , RF ¹ }	{RHP ¹ , LHP ¹ , RK ² }
10		{LHP ¹ , LK ¹ , LF ¹ }	{RHP ² , LHP ² , RK ¹ }
11		{RHP ² , RK ² , RF ² }	{RF ¹ , LF ¹ , LK ² }
12		{LHP ² , LK ² , LF ² }	{RF ² , LF ² , LK ¹ }
13	Four	{H ¹ , N ¹ , SM ¹ , SB ¹ }	{H ¹ , N ¹ , H ² , N ² }
14		{H ² , N ² , SM ² , SB ² }	{H ² , N ² , H ¹ , N ¹ }
15		{N ¹ , RS ¹ , RE ¹ , RH ¹ }	{RS ¹ , LS ¹ , RS ² , LS ² }
16		{N ¹ , LS ¹ , LE ¹ , LH ¹ }	{RE ¹ , LE ¹ , RE ² , LE ² }
17		{N ² , RS ² , RE ² , RH ² }	{RH ¹ , LH ¹ , RH ² , LH ² }
18		{N ² , LS ² , LE ² , LH ² }	{SM ¹ , SB ¹ , SM ² , SB ² }
19		{SB ¹ , RHP ¹ , RK ¹ , RF ¹ }	{RHP ¹ , LHP ¹ , RHP ² , LHP ² }
20		{SB ¹ , LHP ¹ , LK ¹ , LF ¹ }	{RK ¹ , LK ¹ , RK ² , LK ² }
21		{SB ² , RHP ² , RK ² , RF ² }	{RF ¹ , LF ¹ , RF ² , LF ² }
22		{SB ² , LHP ² , LK ² , LF ² }	
23	Five	{H ¹ , N ¹ , RS ¹ , RE ¹ , RH ¹ }	{H ¹ , N ¹ , SM ² , N ² , H ² }
24		{H ¹ , N ¹ , LS ¹ , LE ¹ , LH ¹ }	{H ² , N ² , SM ¹ , N ¹ , H ¹ }
25		{H ² , N ² , RS ² , RE ² , RH ² }	{RS ¹ , LS ¹ , SB ² , RS ² , LS ² }
26		{H ² , N ² , LS ² , LE ² , LH ² }	{RS ² , LS ² , SB ¹ , RS ¹ , LS ¹ }
27		{SM ¹ , SB ¹ , RHP ¹ , RK ¹ , RF ¹ }	{RE ¹ , LE ¹ , RH ² , RE ² , LE ² }
28		{SM ¹ , SB ¹ , LHP ¹ , LK ¹ , LF ¹ }	{RE ¹ , LE ¹ , LH ² , RE ² , LE ² }
29		{SM ² , SB ² , RHP ² , RK ² , RF ² }	{RK ² , RF ² , RHP ¹ , RK ¹ , RF ¹ }
30		{SM ² , SB ² , LHP ² , LK ² , LF ² }	{LK ² , LF ² , LHP ¹ , LK ¹ , LF ¹ }

¹ Joints of the first (left) silhouette, ² Joints of the second (right) silhouette, H = head, N = neck, SM = spine-mid, SB = spine-base, RS = right shoulder, LS = left shoulder, RE = right elbow, LE = left elbow, RH = right hand, LHP = left hand, RHP = right hip, LH = left hip, RK = right knee, LK = left knee, RF = right foot, LF = left foot.

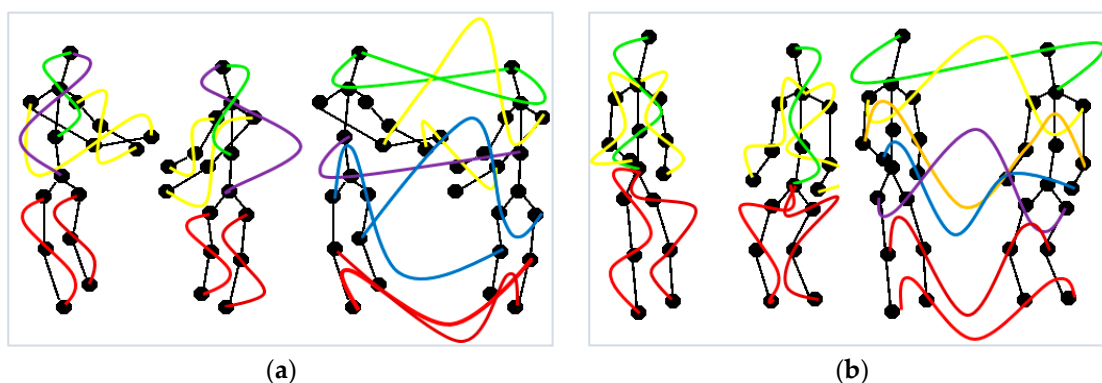


Figure 7. Cont.

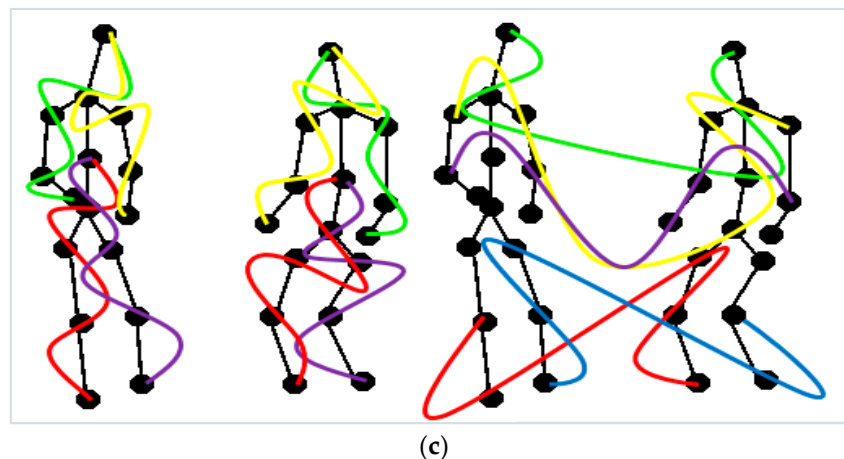


Figure 7. Intra-silhouette and inter-silhouette trajectories over human joints: (a) three way-points; (b) four way-points and (c) five way-points.

Algorithm 2. Pseudo code of feature extraction from skeletal joints

```

1: Input: RGB and depth silhouette frames  $(f_1, f_2, \dots, f_N)$  // where  $N$  is total number of frames.
   Skeleton  $S$  consisting of sixteen human joints as  $\{j_1, j_2, \dots, j_n\}$  // where  $n$  is total number of joints
2: Output: Skeletal joints feature descriptors from all silhouette frames as  $(D_1, D_2, \dots, D_N)$ 
   //Joints MOCAP feature descriptors//
3: for  $i = 1: N$ 
4:   for  $j = 1: n$ 
5:     calculate angle of tangent  $\theta_{up}$  from spine mid joint to all the upper body joints
6:     calculate angle of tangent  $\theta_{low}$  from spine base joint to all the lower body joints
7:     compute joints MOCAP feature descriptor  $J_{MOCAP} \leftarrow \text{concatenate}(\theta_{up}, \theta_{low})$ 
8:   end for
   //way-point trajectory feature descriptors//
9:   for  $i = 1: n$ 
10:    compute subsets  $Sub_3, Sub_4$  and  $Sub_5$  consisting of sets of three, four and five
11:    number of joints, respectively
12:    generate trajectories as three-way  $T_3$  from  $Sub_3$ , four-way  $T_4$  from  $Sub_4$  and four-way  $T_5$ 
13:    from  $Sub_5$ 
14:    compute displacement  $d_{x,y}$  and motion  $C_t$  vector from trajectories  $T_3, T_4$  and  $T_5$  with
15:    time  $t$ 
16:    generate way-point trajectory descriptor  $T \leftarrow \text{concatenate}(d_{x,y}, C_t)$ 
17:   end for
18:   skeletal joints feature descriptor  $D \leftarrow \text{concatenate}(J_{MOCAP}, T)$ 
19: end for
20: return Skeletal joints feature descriptors from all silhouette frames as  $(D_1, D_2, \dots, D_N)$ 

```

3.3. Particle Swarm Optimization (PSO)

After combining RGB-D descriptors to recognize human activities, a very complex representation is generated. So, for an efficient time and space computation, PSO is applied for feature selection and dimensionality reduction. PSO belongs to a stochastic optimization technique category [82]. This algorithm is based on the communal behavior of groups of animals such as birds, insects and fishes [83]. At first, optimization is initialized by randomly selecting a swarm, i.e., a sample of candidate solutions from feature descriptors. The t position of this swarm in dimension D is constantly regulated by a position vector \vec{x}_i and velocity vector \vec{v}_i defined as:

$$\bar{x}_i(t) = (x_{i1}(t), x_{i2}(t), \dots, x_{iD}(t)) \quad (11)$$

$$\bar{v}_i(t) = (v_{i1}(t), v_{i2}(t), \dots, v_{iD}(t)) \quad (12)$$

where $i = 1, 2, 3 \dots N$. N is the total number of particles. A movement of this selected swarm is initialized, and the direction of this movement is toward the best found position in the search space. During this whole optimization process, the three types of variables that are retained by every candidate of optimization are current velocity, current position and personal best position. The personal best position called $pbest$ is maintained in a vector $\bar{p}_i = (p_{i1}, p_{i2}, \dots, p_{iD})$ and gives the optimal fitness value. However, the global best position ($gbest$) is also maintained in a vector as $\bar{p}_g = (p_{g1}, p_{g2}, \dots, p_{gD})$ and gives the best particle from all the N particles. Both the position and velocity of particles are updated in the search space according to the new best position, thus:

$$\vec{x}_i(t+1) = \bar{x}_i(t) + \bar{v}_i(t+1) \quad (13)$$

$$\vec{v}_i(t+1) = \vec{v}_i(t) + \varphi_1(\bar{p}_i - \bar{x}_i(t)) + \varphi_2(\bar{p}_g - \bar{x}_i(t)) \quad (14)$$

where φ_1 and φ_2 can be defined as random numbers. All the particles finally converge to local minima after calculating best values. This is an iterative process which continues until a best solution is learned. Then, original dimension of NTU RGB+D feature descriptors is 5360×550 , for UoL it is 5360×400 and for CAD it is 5360×250 . The length of the combined feature vector of all four proposed features is 5360 which is reduced to 4796×550 for NTU RGB+D, 4796×400 for UoL and 4796×250 for CAD dataset. At the end, all the particles are assigned the best place in the search space. Movement of each particle is influenced by both the local best position and global best position. All the swarm particles try to get closer to the global best position by moving towards and getting closer to it. Movement of swarm particles that are trying to achieve global best position by moving towards $gbest$ is displayed in Figure 8.

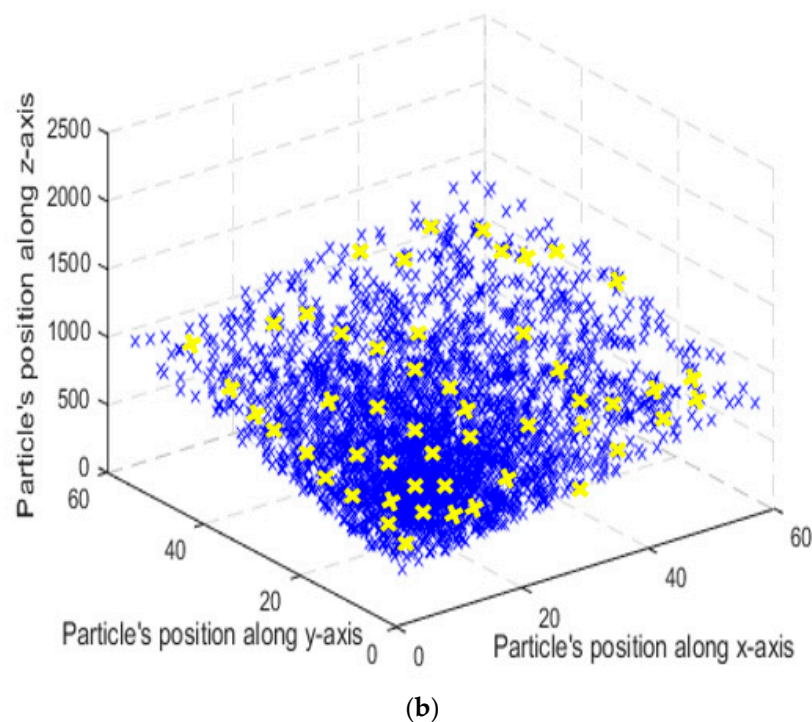
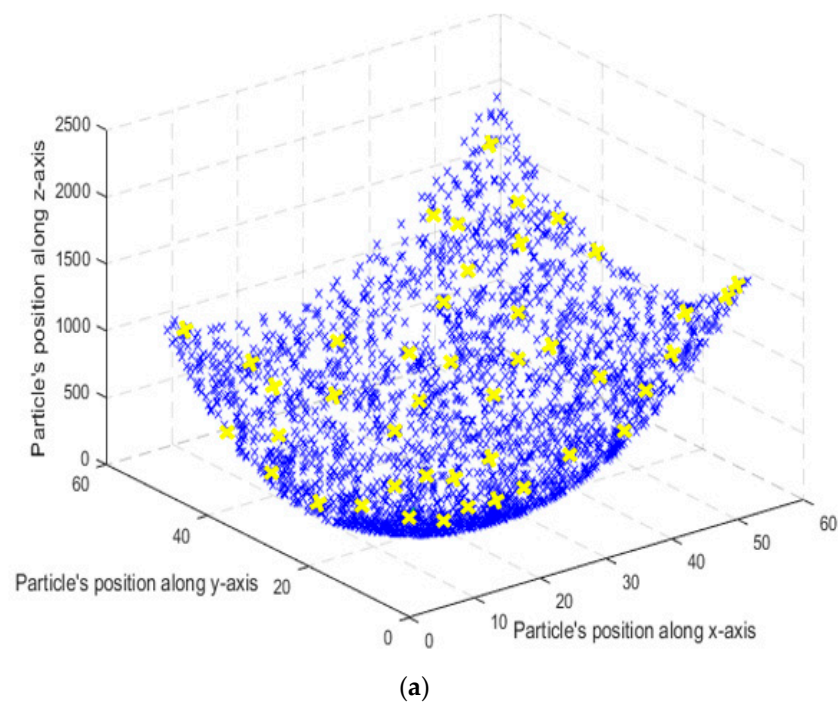


Figure 8. Movement of the swarm particles (shown in blue color) towards global best positions (shown in yellow color) after applying Particle Swarm Optimization (PSO) on action classes of (a) the NTU RGB+D dataset and (b) the UoL dataset.

3.4. Neuro-Fuzzy Classifier (NFC)

In order to accelerate the recognition rate of human actions, NFC, i.e., the hybrid of fuzzy set theory and Artificial Neural Networks (ANN) is applied. This hybrid classifier results in an intelligent inference system which is capable of both reasoning and self-learning [84]. Many action recognition systems based on NFC have been proposed in recent years [85]. This is a six-layer architecture, as displayed in Figure 9.

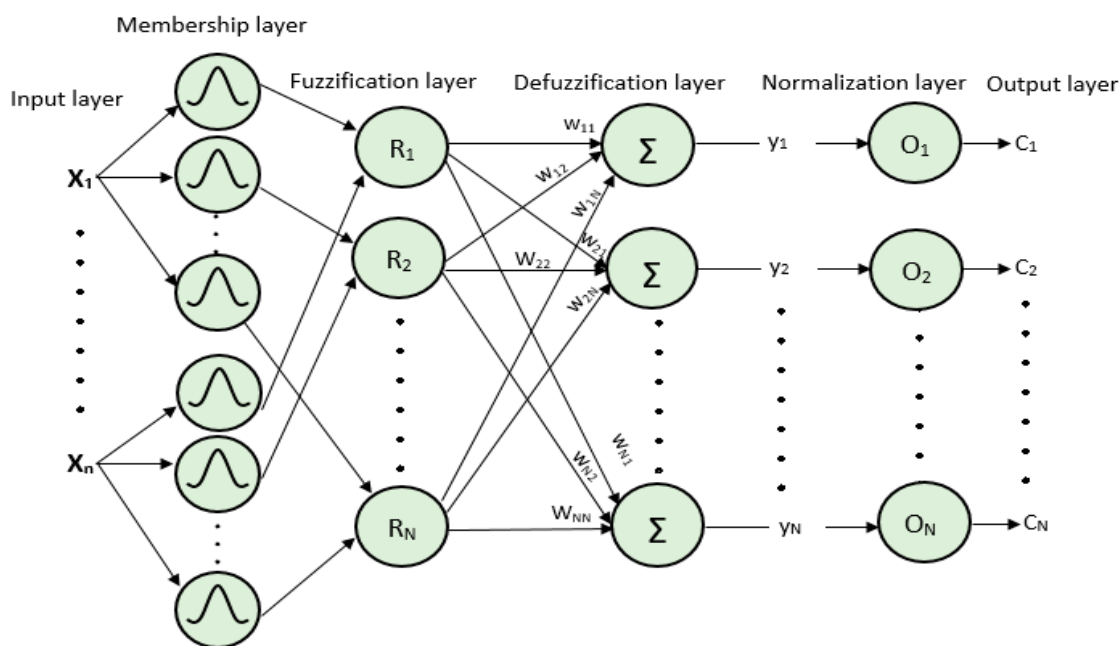


Figure 9. The six-layer architecture of the NFC.

First of all we fed our training data to the input layer: $\{(x^1, c^1) \dots (x^k, c^k)\}$, where $x^k = [x_1^k \dots x_m^k]^K \in R^m$ is the vector of dimension m , and $c^k = [1, 2, \dots, n]$ is the label of the class to which it belongs and n is the total number of classes in a training dataset. The second layer is the membership layer. In this layer, the membership function of each input vector is recognized. We applied a Gaussian membership function. The membership function u_{ij} for sample x_{sj} with mean c and standard deviation σ is defined by:

$$u_{ij}(x_{sj}) = \exp\left(-\frac{(x_{sj} - c_{ij})^2}{2\sigma_{ij}^2}\right) \tag{15}$$

where j, i and s represents the feature, the rule of the corresponding feature and the sample, respectively. Results of the Gaussian function for each input are fed into the third layer, i.e., the fuzzification layer [86]. The firing strength of each generated fuzzy rule w.r.t each corresponding sample x_s from all features N is calculated in this layer as:

$$\alpha_{is} = \prod_{j=1}^N u_{ij}(x_{sj}) \tag{16}$$

where α is the firing strength for i th rule. The fourth layer is called the defuzzification layer. Nodes in this layer are equal to the total number of action classes in the training data. In this layer, output is generated by integrating the results of the preceding layers, i.e., firing strength α_{is} with weight values w_{ik} . The output y_{sk} for a sample s from class k is generated by:

$$y_{sk} = \sum_{i=1}^M \alpha_{is} w_{ik} \tag{17}$$

This weighted summation is completed from rule i to overall generated rules M . At the last layer, all the output values are normalized by dividing the output of each sample s from each class k with the sum of the output for all the classes K . The normalized output o_{sk} is given as:

$$o_{sk} = \frac{y_{sk}}{\sum_{l=1}^K y_{sl}} \tag{18}$$

In this way, the class label of each s sample is obtained by maximum o_{sk} value. When all the testing vectors are fed to the NFC architecture, the resultant output accurately predicts its class label for the input vector. The steps involved in predicting class labels from input data are given in Algorithm 3.

Algorithm 3. Pseudo code of Neuro-fuzzy Classification of Optimized Vectors

```

1: Step 1: for each input node in first layer do fuzzification by calculating the membership
grade as
2:    $u_{ij} \leftarrow \text{gussM}(x, \text{sig}, c)$ 
3:   end for
4: Step 2: for each node in second layer calculate fuzzy strength by product of each sample with
5:   antecedents of previous layer  $\alpha_{is} \leftarrow \text{rule-layer}(u_{ij})$ 
6:   end for
7: Step 3: for each node in third layer defuzzify each node and generate output by weighted
sum
8:   of firing strenghts  $y_{sk} \leftarrow \text{sum}(\alpha_{is}, w_{ik}.)$ 
9:   end for
10: Step 4: for each node in this layer normalize each output class by dividing it with sum of all
11:   output classes  $o_{sk} \leftarrow y_{sk}/\text{sum}(y_{sl})$ 
12:   end for
13: Step 5: Assign class label to each sample  $C_s \leftarrow \max\{o_{sk}\}$ 

```

4. System Validation and Experimentation

This section first gives a brief description of the datasets used for training and testing the proposed system. Then, the parameters used for evaluation of the system and the generated results are given. All the experiments are performed on MATLAB (R2017a). Four parameters are used to validate system performance. First, the recognition rate of the individual activities of all three datasets is given. The second parameter is the effect of the number of membership functions on evaluation time and performance. The third parameters are used for testing system sensitivity, specificity and error measures. The fourth parameter is the comparison of the proposed system with other systems that have been reported in recent years. Each parameter is explained in detailed subsections.

4.1. Dataset Descriptions

Table 7 gives the name, type of input data and description of each dataset used for the training and testing of the proposed system.

Table 7. Description of datasets used for evaluation and experimentation.

Name of Dataset	Type of Input Data	Action Classes
NTU RGB+D	RGB videos, depth map sequences, 3D skeletal data and infrared video	This dataset contains sixty action classes and 56,880 video samples. The eleven mutual action classes that we used in the proposed system are: punch/slap, kicking, pushing, pat on back, point finger, hugging, giving object, touch pocket, shaking hands, walking towards and walking apart. Dataset descriptions and samples are given in [87].
UoL 3D social activity dataset	RGB-D images with tracked skeletons	This dataset contains eight two-person social interaction activities: handshake, hug, help walk, help stand-up, fight, push, conversation and call attention. The rest of the details and dataset samples are given in [88].
Collective Activity Dataset (CAD)	RGB	This dataset consists of RGB sequences of five actions classes: crossing, walking, talking, queueing and waiting. Actions are performed in both indoor and outdoor environments. Dataset descriptions and samples are given in [89].

4.2. Recognition Accuracy

In order to validate the system's performance, descriptors of action classes from each dataset are given individually to NFC to identify the recognition rate. The percentage of accuracies for each class is given separately in the form of a confusion matrix. Each activity class of all three datasets used for experimentation achieved very good performance results with the proposed system. Table 8 shows the confusion matrix of action classes of the NTU RGB+D dataset.

Table 8. Confusion matrix for action classes of the NTU RGB+D dataset.

		Predicted Action Classes										
		S	K	P	PB	PF	H	GO	TP	SH	WT	WA
Actual Action Classes	S ¹	97	0	1	0	2	0	0	0	0	0	0
	K ²	0	96	0	0	0	0	0	1.5	0	2.5	0
	P ³	1	0	98	0.5	0	0	0.5	0	0	0	0
	PB ⁴	0	0	3	89	1	4	0	0	0	1	2
	PF ⁵	1.5	0	3	2.5	88.5	0	2	0.5	2	0	0
	H ⁶	0	0	3	0	2	93	0	0	1	1	0
	GO ⁷	0	0.2	0	3	2	2	89	0	3.8	0	0
	TP ⁸	0	0	0	3.8	0	1	0	92.9	0	0.7	1.6
	SH ⁹	0	1.2	0	0	2.6	1.4	0	0	94.8	0	0
	WT ¹⁰	1	0	0	0.8	0	0	2	0	0.2	96	0
	WA ¹¹	0	0	0	4	0	0	0	2	0	0	94

Mean Accuracy = 93.5%

¹ slap/punch, ² kicking, ³ pushing, ⁴ pat on back, ⁵ point finger, ⁶ hugging, ⁷ giving object, ⁸ touch pocket, ⁹ shaking hands, ¹⁰ walking towards, ¹¹ walking apart.

It is inferred from Table 8 that the average recognition rate for the NTU RGB+D dataset is 93.5%. Each activity class is recognized with a high recognition accuracy. Due to our robust features set, the proposed system has achieved excellent accuracies of 98%, 97% and 96% with slap/punch, kicking and pushing interactions, respectively. Thus, it is proved that our system is capable of detecting anomalous activities from environment. Regular activities like pointing the finger and hugging are also recognized with very high accuracy rates. The lowest accuracy rates are observed in activities such as pat on back, point finger and giving object due to the repetition of similar movements involved in these activities. For example, the actions giving object and shaking hands are performed with similar movements of the same body parts (the hands). Table 9 shows the confusion matrix for action classes of the UoL dataset.

When a testing set of action classes from the UoL dataset is given to NFC, an average recognition rate of 92.2% is achieved. It is inferred from Table 9 that anomalous activities from this dataset are also detected with excellent recognition rates. This is because of the strong set of skeletal joints data and full body features which enable our system to detect continuous activities such as hug and handshake with very high accuracy rates. However, a slight drop in the recognition rate is observed with conversations and call attention activities due to similarities in human body gestures and postures. Table 10 shows the confusion matrix for action classes of the CAD dataset.

Table 9. Confusion matrix for action classes of the UoL dataset.

		Predicted Action Classes							
		Handshake	Hug	Help Walk	Help Stand-Up	Fight	Push	Conversation	Call Attention
Actual Action Classes	Handshake	96.5	1.5	2	0	0	0	0	0
	Hug	2	95	1	0	0	0	2	0
	Help walk	0	2	92	3	1	0	2	0
	Help stand-up	2	0	0	91.2	4.8	0	0	2
	Fight	0	2	0	1	92	5	0	0
	Push	0	1	1	0	3	94	0	1
	Conversation	2	5	0	0	2	0	89	2
	Call Attention	3	0	0	4	0	0	5	88
Mean Accuracy = 92.2%									

Table 10. Confusion matrix for action classes of the CAD dataset.

		Predicted Action Classes				
		crossing	talking	walking	queueing	waiting
Actual Action Classes	crossing	88	0	10	2	0
	talking	0	92	3	0	5
	walking	8	0	84	3	5
	queueing	0	2	4	90	4
	waiting	1	0	1	4	94
Mean Accuracy = 89.6%						

CAD is a very challenging outdoor dataset with highly occluded backgrounds. The average recognition rate with CAD is slightly less compared to the NTU RGB+D and the UoL datasets. Nevertheless, our system is capable of recognizing some activities such as talking and waiting with 92% and 94% recognition rates, respectively. Actions involved in all classes of the CAD dataset are strongly related to each other so a confusion rate as high as 10% is observed in activities such as crossing and walking. A mean performance rate of 89.6% is achieved with the CAD dataset. In summary, this experiment proved the effectiveness of the proposed system by achieving high recognition rates with all three datasets.

4.3. The Effects of Numbers of Membership Functions

In this experiment, the effect of different numbers of Membership Functions (MF) over computation time, Root Mean Square Error (RMSE) and accuracy is observed. A Gaussian membership function is used. During experimentation, the number of MFs is varied from 2, 3, 5, to 8. The number of epochs is changed from 200 to 300 and 500. This experiment is performed with all the three datasets. Tables 11–13 demonstrate the effects of different numbers of MFs on NTU RGB+D, UoL and CAD datasets, respectively.

Table 11. Effects of different numbers of Membership Function (MF) on performance with the NTU RGB+D dataset.

Parameters		Performance		
No. of Epochs	No. of MF	Computational Time (s)	RMSE	Accuracy (%)
200	3	20	0.065	90.5
	5	27.4	0.059	91.2
	8	35.2	0.060	90.9
300	3	25	0.058	92.0
	5	32.8	0.055	93.5
	8	38	0.055	93.0
500	3	32	0.059	92.9
	5	47.2	0.055	93.1
	8	58	0.056	93.1

Table 12. Effects of different numbers of MF on performance with the UoL dataset.

Parameters		Performance		
No. of Epochs	No. of MF	Computational Time (s)	RMSE	Accuracy (%)
200	3	15	0.069	88.7
	5	24	0.059	90.5
	8	32	0.060	90
300	3	27	0.070	89
	5	33	0.056	92
	8	45	0.061	91.9
500	3	29	0.070	90
	5	43	0.059	92
	8	57	0.066	92

Table 13. Effects of different numbers of MF on performance with the CAD dataset.

Parameters		Performance		
No. of Epochs	No. of MF	Computational Time (s)	RMSE	Accuracy (%)
200	3	20	0.125	82.2
	5	23	0.097	87.8
	8	31	0.098	85
300	3	25	0.099	88.7
	5	27	0.096	89.5
	8	34	0.096	89
500	3	31	0.111	87
	5	35	0.098	89
	8	40	0.099	88.5

It is observed from the results given in Tables 11–13 that increases in the number of membership functions affect the performance and computation time of the system. Increases in the number of MFs up to some points result in increased performance. However, after a certain limit increases in the number of MFs will result in the repetition of similar patterns. For example, in Table 11, increases in the number of MFs from five to eight results in increases of RMSE and decreases in the system's recognition rate. This is because increases in number of MFs after a certain limit will result in increased in fuzzy rules and the problem of overfitting occurs. However, if we use very few numbers of MFs, then fewer numbers of fuzzy rules will be compared and system performance will decrease. The minimum RMSE is observed with five MFs at the cost of computation time with NTU RGB+D, UoL, and also with CAD dataset. It is also observed that an increase in

the number of MF and iterations results in increased computation time. However, increases in the number of iterations above a certain limit start to produce similar results to previous iterations. The best performance is achieved with 300 iterations. Thus, it is inferred that the number of MF and iterations effects the performance of the system.

4.4. Sensitivity, Specificity and Error Measures

For an in-depth evaluation and validation of the proposed system, we calculated sensitivity, specificity and error measures. Sensitivity measures the probability of detection, i.e., the True Positive Rate (TPR), while specificity measures the True Negative Rate (TNR). In order to represent false classifications, False Positive Rates (FPR) or fall-out rate and False Negative Rates (FNR) or miss-rate are calculated. FPR and FNR identify errors or misclassification rates. Sensitivity, specificity, FPR and FNR for each activity class of NTU RGB+D, UoL and CAD dataset are displayed in the form of bar graphs in Figures 10–12, respectively.

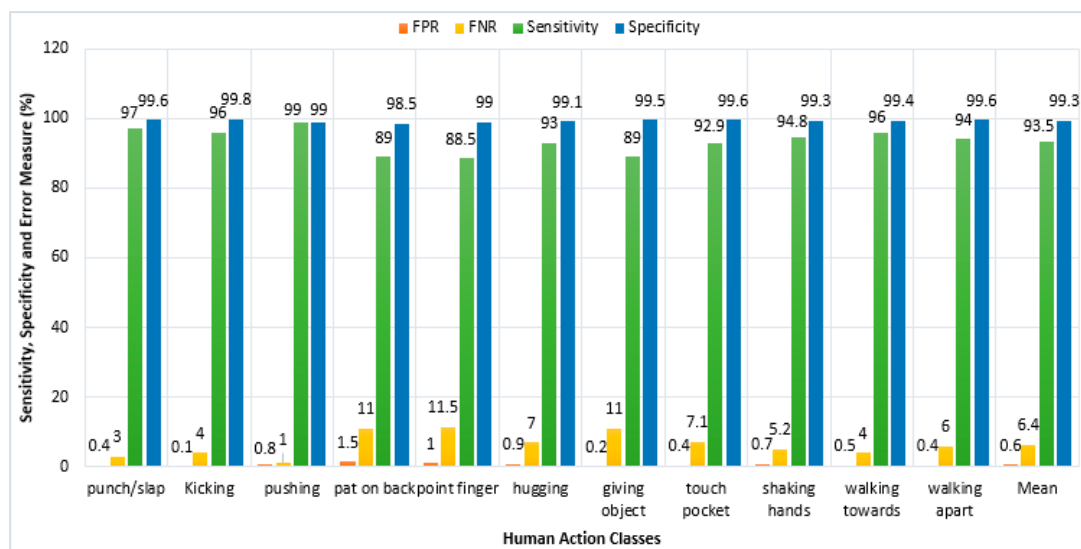


Figure 10. Bar graph displaying sensitivity, specificity, False Positive Rates (FPR) and False Negative Rates (FNR) measures with classes of the NTU RGB+D dataset.

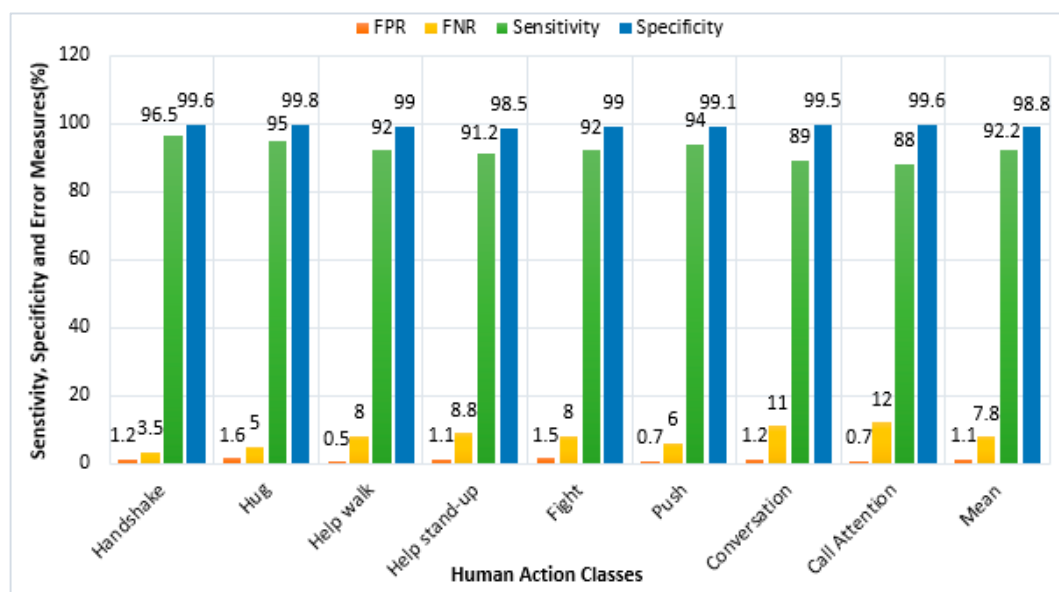


Figure 11. Bar graph displaying sensitivity, specificity, FPR and FNR measures with classes of the UoL dataset.

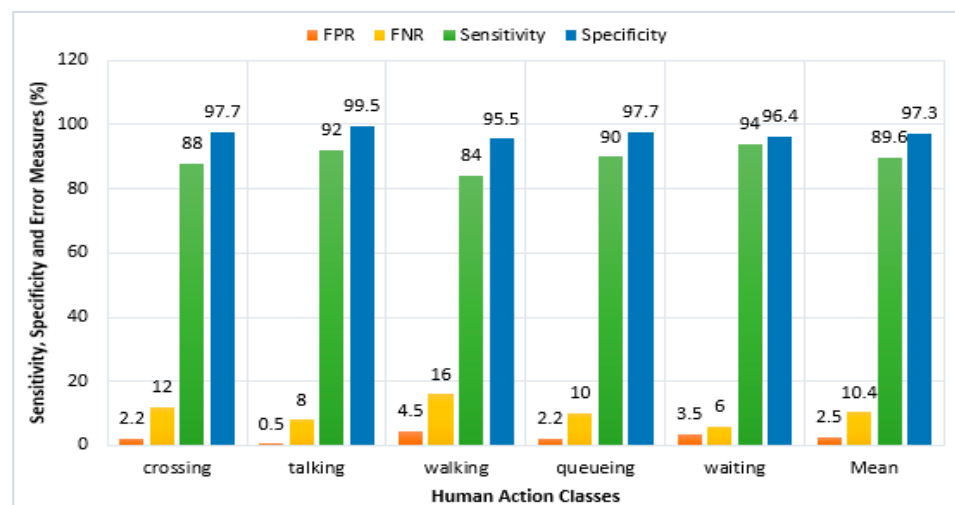


Figure 12. Bar graph displaying sensitivity, specificity, FPR and FNR measures with classes of the CAD dataset.

It is observed from Figure 10 that the sensitivity ratio of all action classes of the NTU RGB+D dataset is very high. The proposed system can clearly distinguish between the classes and accurately predict the true label of the class to which it belongs. The overall sensitivity for all action classes is as high as 93.5%, 92.2% and 89.6% for the NTU RGB+D, UoL and CAD datasets, respectively. This shows that the system has a very small failure rate. The mean true negative rate of the system is 99.3% with NTU RGB+D dataset, 98.8% with the UoL dataset and 97.3% with the CAD dataset. Most of the specificity ratios that we obtained with all three datasets are as high as 99%. Thus, our system has a high ability to reject a testing sample if it does not belong to a specific class.

When the FPR of action classes for all three datasets are compared, it is observed that the mean FPR of NTU RGB+D, UoL and CAD datasets is 0.61%, 1.06% and 2.58%, respectively, as seen in Figures 10–12. On the other hand, the FNR of all three datasets is 6.43% with NTU RGB+D, 7.78% with UoL and 10.4% with CAD datasets. Hence, the error rates are very low, compared to sensitivity and specificity. So, a robust system is produced with high TPR and TNR, and low FPR and FNR.

4.5. Computational Time Complexity

In order to demonstrate the efficiency of a system, an experiment is performed to compute the computational time of the system. This experiment investigates the running time with respect to given input in the form of frames. A Core i5-4300U CPU (Control Processing Unit of speed 1.90 GHz and MATLAB (R2017a)) is used to compute the running time. The testing set of single activity class consists of 30 frames per action. When a testing sample of each activity class was given to the system, it took 3.3 s to recognize the action and assign a class label to a given input. For one frame, the computational time for recognition of the human action was 0.11 s. So, our system is capable of providing real-time human action recognition of 10 frames per second. Furthermore, in this experiment, computational time of the proposed system was compared with Artificial Neural Networks (ANN) as a classifier. First of all, the action classes from all three datasets were given individually as an input to the proposed system and the computational time in which the system classified all the action classes is measured. Then action classes from all three datasets were given individually as an input to the proposed system and classified via ANN instead of NFC. The proposed system with NFC provided results faster than ANN approach. Figure 13 shows the computational time with action classes of NTU RGB+D dataset, UoL dataset and CAD dataset classified via NFC and ANN.

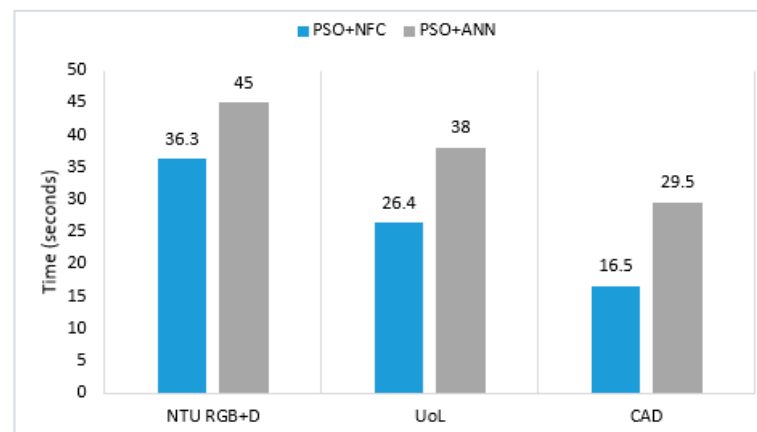


Figure 13. Comparison of computational time between NFC and ANN.

4.6. Comparison with Other Recently Proposed Systems

Finally, we compared the performance of the proposed system with other recently developed systems. We compared the recognition rates of our system with other system using same activities from all three datasets. In [49] each interaction is divided into interaction of different body parts. Two unique features called RLTP (Relative Location Temporal Pyramid) and PCTP (Physical Contact Temporal Pyramid) are produced. In [90], two libraries, OpenPose and 3D-baseline, are used to extract human joints while Convolutional Neural Networks (CNN) is used to classify the results. In [57], a feature factorization network is proposed. For better classification, a sparsity learning machine is produced. In [91], a system is proposed for both pose estimation and action recognition. In this method, action heat maps are generated via CNN. In [92], a scale and translation invariant transformation of skeletal images to color images is performed. In order to adjust frequency, a multi-scale CNN is used.

In [50], a method to temporally detect human action is presented by tracking the movements of upper body joints. HMM is used to detect and classify human actions. They used proxemics theory which depends on the usage of space during social interactions. In [93], spatiotemporal features are extracted from each person interacting in an action class, while social features are extracted between two interacting persons. In [94], human actions are tracked with skeletal data and by human body postures. SVM and X-means are used in the training and testing phase. In [95], a method based on the fusion of RGB, depth and wearable inertial sensors is presented. HOG and statistical features are extracted to record human actions. In [39], a connection between atomic activities is measured and interaction responses are formulated. A multi-task interaction response (MIR) was computed for each class separately. In [61], inter-related dynamic among different persons is identified via LSTM. First the static features of one person are given to Single-person LSTM and then its output is given to Concurrent-LSTM. In [96], a graphical model is used to find relationships between individual persons in an interaction. Furthermore, structured learning is introduced to connect with right output. In [97], the relationships among individuals as well as the atomic activity performed by each individual are measured. Table 14 shows the comparison of performances on the NTU RGB+D, UoL and CAD datasets.

It is observed from Table 14 that the proposed system performed better than many action recognition systems of recent years. The proposed system works well for HAR because of the features used to track each and every movement made by both persons involved in an interaction. The incorporation of depth sensors makes it possible to predict even complex human-to-human actions accurately. The data obtained after extracting features are in a more structured form to make decisions which improve the performance of the system. In a very short time, our system can give results with high sensitivity and accuracy. On the other hand, deep learning methods presented in the comparison consist

of very complex data models that take more computational time to predict results. A large amount of data is used in the compared deep learning-based approaches for training and then predicting the right outcome. By contrast, the proposed system can be used as real-time surveillance system which can learn from a small number of data samples and produce results in less time.

Table 14. Comparison of the proposed system with other recently proposed systems.

Datasets	Authors	Methodology	Recognition Accuracy (%)
NTU RGB+D	Meng Li et al. [49]	Pairwise features	88.6
	Junwoo et al. [90]	Mobile robot platform	75.0
	Amir et al. [57]	Deep multimodal features	74.9
	Diogo et al. [91]	Multitask deep learning	85.5
	Bo et al. [92] Proposed Methodology	Skeletal based action recognition RGB-D skeletal and full body features	88.6 93.5
UoL 3D activity	Claudio et al. [50]	Statistical and geometrical features	85.5
	Claudio et al. [93]	Probabilistic merging of skeleton features	85.1
	Alessandro et al. [94]	Skeletal data	87.0
	Muhammad et al. [95]	Multimodal feature level fusion	85.1
	Proposed Methodology	RGB-D skeletal and full body features	92.2
CAD	Xiaobin et al. [39]	Interaction response from atomic actions	83.3
	Xiangbo et al. [61]	Hierarchical Long Short-Term Concurrent Memory	83.7
	Zhiwei et al. [96]	Relationship in group activity	81.2
	Wongun et al. [97]	Multitarget tracking	73.3
	Proposed Methodology	RGB-D skeletal and full body features	89.6

5. Discussion

A sustainable system with high stability and uniformity towards different challenges faced during performance is proposed in this research work. We used three challenging datasets in both indoor and outdoor environments. Our system produced uniformly good performance with all three datasets by tackling problems of varying environment conditions such as various brightness and lightning conditions due to the incorporation of depth sensors. Actions of all three datasets used in the proposed system are very complex because the movements involved in performing most of the actions are quite similar to each other. For example, walking towards, shaking hands, giving an object are actions in which two persons move towards each other. However, our system remained stable and reliable in differentiating all similar actions; this is due to the robust set of features. Our features resulted in high accuracy, sensitivity and specificity ratios.

The challenge of silhouette overlapping is faced during the system's execution. Silhouette segmentation of both RGB and depth images is not affected by overlapping silhouettes. However, in the feature extraction phase, there are some images where the silhouette of one person either slightly or completely overlaps the silhouette of another person. For example, in classes such as shaking hands and giving objects, hands of two silhouettes do not overlap at the beginning of the interaction. However, at the end of interaction, the hands of two persons overlap with each other and it is difficult to distinguish and mark the hand joints of each person. In the case of slight silhouette overlapping, blob extraction is performed through connected component analysis and specifying height and width of human. Through blob extraction, the silhouettes of both humans are extracted individually and then the feature extraction is performed. So, the performance of these actions is not very much affected. In some actions such as pat on the back, performance is affected by overlapping of the silhouettes and it is slightly lower (89%) compared to other classes. This is because in this action class, there is constant overlapping of hand of one person with shoulder of other person from start of the interaction until end. Moreover, in instances

where one silhouette overlaps more significantly with the other silhouette (e.g., hugging interaction), connected component analysis fails. In such instances, full-body features such as geodesic distance and 3D Cartesian-plane features are still being computed which is why recognition accuracy is not very much affected. For example, in hugging interaction, the geodesic maps are created by taking a single point of origin, i.e., single centroid is used for both persons. However, in skeletal joints features, the joints of one person are detected on the silhouette of the other person, and the skeleton is deformed. As shown in Figure 14, human joints are not identified in the correct positions.

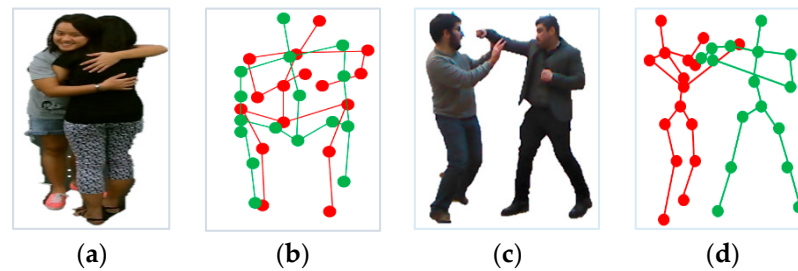


Figure 14. Deformed skeleton due to overlapping of silhouettes. (a,c) original images of hugging and fight action, respectively; (b,d) deformed skeleton of hugging and fight action, respectively.

6. Conclusions

In this research work, an efficient and sustainable human surveillance system is proposed. We developed an action recognition system that is capable of good performance due to the deployment of RGB-D sensors and regardless of varying environments. We proposed four novel features in this research work. These features track each and every movement made by humans. Two features, namely, geodesic distance and 3D Cartesian-plane features are extracted from full human body images. Two of the proposed features, joints MOCAP and way-point trajectory, are extracted from the skeletal joints of humans. By combining all four feature descriptors, we are able to track human locomotion. These feature descriptors are optimized via PSO and, at the end, a hybrid neuro-fuzzy classifier is used to recognize human actions.

The proposed system has been validated via extensive experimentation. At first, feature descriptors of each dataset are separately fed into NFC and the mean recognition rate for each action class is calculated. Mean recognition accuracy with the NTU RGB+D dataset is 93.5%, with the UoL dataset it is 92.2% and with CAD it is 89.6%. We also evaluated our system with different numbers of membership functions over different numbers of iterations. The best performance was obtained with five membership functions but at the cost of computation time. The resulting RMSE values at 300 iterations are 0.55 with the NTU RGB+D dataset, 0.056 with the UoL dataset and 0.096 with CAD. Sensitivity and specificity measures for each activity class were taken to measure system performance from the true positive rate and true negative rate, respectively. Overall, the true positive rate for all action classes was 93.5%, 92.2% and 89.6% with the NTU RGB+D, UoL and CAD datasets, respectively. The overall true negative rate is 99.3%, 98.8% and 97.3% with the NTU RGB+D, UoL and CAD datasets, respectively. Finally, the performance of the proposed system was compared with other systems and these comparisons showed that the proposed system performed better than many state-of-the-art systems.

In future, we have plans to further evaluate our model with deep learning concepts over more challenging human action datasets as well as for group interaction recognition.

Author Contributions: Conceptualization, N.K.; methodology, N.K., M.G. and A.J.; software, N.K.; validation, M.G. and A.J.; formal analysis, K.K. and M.G.; resources, A.J. and K.K.; writing—review and editing, A.J. and K.K.; funding acquisition, A.J. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2018R1D1A1A0208 5645). Also, this work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 202012D05-02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, Y.; Xu, C.; Li, G.; Xu, W.; Kong, J.; Jiang, D.; Tao, B.; Chen, D. Intelligent human computer interaction based on non-redundant EMG signal. *Alex. Eng. J.* **2020**, *59*, 1149–1157. [[CrossRef](#)]
2. Zank, M.; Nescher, T.; Kunz, A. Tracking human locomotion by relative positional feet tracking. In Proceedings of the IEEE Virtual Reality (VR), Arles, France, 23–27 March 2015. [[CrossRef](#)]
3. Jalal, A.; Akhtar, I.; Kim, K. Human posture estimation and sustainable events classification via pseudo-2D stick model and K-ary tree hashing. *Sustainability* **2020**, *12*, 9814. [[CrossRef](#)]
4. Jalal, A.; Kamal, S.; Kim, D. A depth video sensor-based life-logging human activity recognition system for elderly care in Smart indoor environments. *Sensors* **2014**, *14*, 11735–11759. [[CrossRef](#)] [[PubMed](#)]
5. Batool, M.; Jalal, A.; Kim, K. Sensors technologies for human activity analysis based on SVM optimized by PSO algorithm. In Proceedings of the IEEE International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 27–29 August 2019. [[CrossRef](#)]
6. Susan, S.; Agrawal, P.; Mittal, M.; Bansal, S. New shape descriptor in the context of edge continuity. *CAAI Trans. Intell. Technol.* **2019**, *4*, 101–109. [[CrossRef](#)]
7. Shokri, M.; Tavakoli, K. A review on the artificial neural network approach to analysis and prediction of seismic damage in infrastructure. *Int. J. Hydromechatron.* **2019**, *4*, 178–196. [[CrossRef](#)]
8. Tingting, Y.; Junqian, W.; Lintai, W.; Yong, X. Three-stage network for age estimation. *CAAI Trans. Intell. Technol.* **2019**, *4*, 122–126. [[CrossRef](#)]
9. Zhu, C.; Miao, D. Influence of kernel clustering on an RBFN. *CAAI Trans. Intell. Technol.* **2019**, *4*, 255–260. [[CrossRef](#)]
10. Wiens, T. Engine speed reduction for hydraulic machinery using predictive algorithms. *Int. J. Hydromechatron.* **2019**, *1*, 16–31. [[CrossRef](#)]
11. Osterland, S.; Weber, J. Analytical analysis of single-stage pressure relief valves. *Int. J. Hydromechatron.* **2019**, *2*, 32–53. [[CrossRef](#)]
12. Rafique, A.A.; Jalal, A.; Kim, K. Automated sustainable multi-object segmentation and recognition via modified sampling consensus and kernel sliding perceptron. *Symmetry* **2020**, *12*, 1928. [[CrossRef](#)]
13. Mahmood, M.; Jalal, A.; Kim, K. WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors. *Multimed. Tools Appl.* **2020**, *79*, 6919–6950. [[CrossRef](#)]
14. Jalal, A.; Khalid, N.; Kim, K. Automatic recognition of human interaction via hybrid descriptors and maximum entropy Markov model using depth sensors. *Entropy* **2020**, *22*, 817. [[CrossRef](#)] [[PubMed](#)]
15. Prati, A.; Shan, C.; Wang, K.I.-K. Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. *J. Ambient Intell. Smart Environ.* **2019**, *11*, 5–22. [[CrossRef](#)]
16. Sreenu, G.; Saleem Durai, M.A. Intelligent video surveillance: A review through deep learning techniques for crowd analysis. *J. Big Data* **2019**, *6*, 48. [[CrossRef](#)]
17. Xu, H.; Pan, Y.; Li, J.; Nie, L.; Xu, X. Activity recognition method for home-based elderly care service based on random forest and activity similarity. *IEEE Access* **2019**, *7*, 16217–16225. [[CrossRef](#)]
18. Park, S.U.; Park, J.H.; Al-masni, M.A.; Al-antari, M.A.; Uddin, M.Z.; Kim, T.S. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. *Procedia Comput. Sci.* **2016**, *100*, 78–84. [[CrossRef](#)]
19. Zhao, W.; Lun, R.; Espy, D.D.; Reinthal, M.A. Rule based real time motion assessment for rehabilitation exercises. In Proceedings of the IEEE Symposium Computational Intelligence in Healthcare and E-Health, Orlando, FL, USA, 9–12 December 2014. [[CrossRef](#)]
20. Barnachon, M.; Bouakaz, S.; Boufama, B.; Guillou, E. Ongoing human action recognition with motion capture. *Pattern Recognit.* **2014**, *47*, 238–247. [[CrossRef](#)]
21. Bersch, S.D.; Azzi, D.; Khusainov, R.; Achumba, I.E.; Ries, J. Sensor data acquisition and processing parameters for human activity classification. *Sensors* **2014**, *14*, 4239–4270. [[CrossRef](#)]
22. Schrader, L.; Vargas Toro, A.; Konietzny, S.; Rüping, S.; Schäpers, B.; Steinböck, M.; Krewer, C.; Müller, F.; Güttler, J.; Bock, T. Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Popul. Ageing* **2020**, *13*, 139–165. [[CrossRef](#)]
23. Li, J.; Tian, L.; Wang, H.; An, Y.; Wang, K.; Yu, L. Segmentation and recognition of basic and transitional activities for continuous physical human activity. *IEEE Access* **2019**, *7*, 42565–42576. [[CrossRef](#)]

24. Jalal, A.; Batoool, M.; Kim, K. Stochastic recognition of physical activity and healthcare using tri-axial inertial wearable sensors. *Appl. Sci.* **2020**, *10*, 7122. [[CrossRef](#)]
25. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 4405–4425. [[CrossRef](#)]
26. Mahjoub, A.B.; Atri, M. Human action recognition using RGB data. In Proceedings of the International Design & Test Symposium (IDT), Hammamet, Tunisia, 18–20 December 2016. [[CrossRef](#)]
27. Nadeem, A.; Jalal, A.; Kim, K. Human actions tracking and recognition based on body parts detection via artificial neural network. In Proceedings of the International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020. [[CrossRef](#)]
28. Jalal, A.; Mahmood, M.; Hasan, A.S. Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments. In Proceedings of the IEEE IBCAST, Islamabad, Pakistan, 8–12 January 2019. [[CrossRef](#)]
29. Ali, H.H.; Moftah, H.M.; Youssif, A.A.A. Depth-based human activity recognition: A comparative perspective study on feature extraction. *Future Comput. Inform. J.* **2018**, *3*, 51–67. [[CrossRef](#)]
30. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [[CrossRef](#)]
31. Jalal, A.; Kamal, S.; Kim, D. Human depth sensors-based activity recognition using spatiotemporal features and hidden Markov model for smart environments. *J. Comput. Netw. Commun.* **2016**, *1026*, 2090–7141. [[CrossRef](#)]
32. Ince, Ö.F.; Ince, I.F.; Yildirim, M.E.; Park, J.S.; Song, J.K.; Yoon, B.W. Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor. *ETRI J.* **2020**, *42*, 78–89. [[CrossRef](#)]
33. Tahir, S.B.; Jalal, A.; Kim, K. Wearable inertial sensors for daily activity analysis based on Adam optimization and the maximum entropy Markov model. *Entropy* **2020**, *22*, 579. [[CrossRef](#)]
34. Ahmed, A.; Jalal, A.; Kim, K. A novel statistical method for scene classification based on multi-object categorization and logistic regression. *Sensors* **2020**, *20*, 3871. [[CrossRef](#)]
35. Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: A survey. *Multimed. Tools Appl.* **2020**, *79*, 30509–30555. [[CrossRef](#)]
36. Nguyen, N.; Yoshitaka, A. Human interaction recognition using hierarchical invariant features. *Int. J. Semant. Comput.* **2015**, *9*, 169–191. [[CrossRef](#)]
37. Cho, N.; Park, S.; Park, J.; Park, U.; Lee, S. Compositional interaction descriptor for human interaction recognition. *Neurocomputing* **2017**, *267*, 169–181. [[CrossRef](#)]
38. Bibi, S.; Anjum, N.; Sher, M. Automated multi-feature human interaction recognition in complex environment. *Comput. Ind.* **2018**, *99*, 282–293. [[CrossRef](#)]
39. Chang, X.; Zheng, W.-S.; Zhang, J. Learning person-person interaction in collective activity recognition. *IEEE Trans. Image Process.* **2015**, *24*, 1905–1918. [[CrossRef](#)] [[PubMed](#)]
40. Ye, Q.; Zhong, H.; Qu, C.; Zhang, Y. Human interaction recognition based on whole-individual detection. *Sensors* **2020**, *20*, 2346. [[CrossRef](#)] [[PubMed](#)]
41. Nadeem, A.; Jalal, A.; Kim, K. Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness. *Symmetry* **2020**, *12*, 1766. [[CrossRef](#)]
42. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [[CrossRef](#)]
43. Mahmood, M.; Jalal, A.; Siddiqui, M.A. Robust spatio-temporal features for human interaction recognition via artificial neural network. In Proceedings of the International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018. [[CrossRef](#)]
44. Sharif, M.; Khan, M.A.; Akram, T.; Younus, M.J.; Saba, T.; Rehman, A. A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. *EURASIP J. Image Video Process.* **2017**, *2017*, 89. [[CrossRef](#)]
45. Kao, J.; Ortega, A.; Tian, D.; Mansour, H.; Vetro, A. Graph based skeleton modeling for human activity analysis. In Proceedings of the International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019. [[CrossRef](#)]
46. Al-Akam, R.; Paulus, D. Local feature extraction from RGB and depth videos for human action recognition. *Int. J. Mach. Learn. Comput.* **2018**, *8*, 274–279. [[CrossRef](#)]
47. Jalal, A.; Kamal, S.; Kim, D. Shape and motion features approach for activity tracking and recognition from kinect video camera. In Proceedings of the IEEE International Conference on Advanced Information Networking and Applications Workshops, Gwangju, Korea, 24–27 March 2015. [[CrossRef](#)]
48. Ji, X.; Wang, C.; Ju, Z. A new framework of human interaction recognition based on multiple stage probability fusion. *Appl. Sci.* **2017**, *7*, 567. [[CrossRef](#)]
49. Li, M.; Leung, H. Multi-view depth-based pairwise feature learning for person-person interaction recognition. *Multimed. Tools Appl.* **2019**, *78*, 5731–5749. [[CrossRef](#)]
50. Coppola, C.; Cosar, S.; Faria, D.R.; Bellotto, N. Automatic detection of human interactions from RGB-D data for social activity classification. In Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN), Lisbon, Portugal, 28 August–1 September 2017. [[CrossRef](#)]

51. Jalal, A.; Quaid, M.A.K.; Kim, K. A wrist worn acceleration based human motion analysis and classification for ambient smart home system. *J. Electr. Eng. Technol.* **2019**, *14*, 1733–1739. [[CrossRef](#)]
52. Kong, Y.; Liang, W.; Dong, Z.; Jia, Y. Recognizing human interaction from videos by a discriminative model. *IET Comput. Vis.* **2014**, *8*, 277–286. [[CrossRef](#)]
53. Ji, Y.; Cheng, H.; Zheng, Y.; Li, H. Learning contrastive feature distribution model for interaction recognition. *J. Vis. Commun. Image Represent.* **2015**, *33*, 340–349. [[CrossRef](#)]
54. Subetha, T.; Chitrakala, S. Recognition of human-human interaction using CWDTW. In Proceedings of the International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, India, 18–19 March 2016. [[CrossRef](#)]
55. Jalal, A.; Kamal, S.; Azurdia-Meza, C.A. Depth maps-based human segmentation and action recognition using full-body plus body color cues via recognizer engine. *J. Electr. Eng. Technol.* **2019**, *14*, 455–461. [[CrossRef](#)]
56. Huynh-The, T.; Banos, O.; Le, B.-V.; Bui, D.-M.; Lee, S.; Yoon, Y.; Le-Tien, T. PAM-based flexible generative topic model for 3D interactive activity recognition. In Proceedings of the International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh, Vietnam, 14–16 October 2015. [[CrossRef](#)]
57. Shahroudy, A.; Ng, T.; Gong, Y.; Wang, G. Deep multimodal feature analysis for action recognition in RGB+D videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1045–1058. [[CrossRef](#)]
58. Shu, X.; Tang, J.; Qi, G.-J.; Song, Y.; Li, Z.; Zhang, L. Concurrence-aware long short-term sub-memories for person-person action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017. [[CrossRef](#)]
59. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-16), Beijing, China, 24 March 2016.
60. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [[CrossRef](#)]
61. Shu, X.; Tang, J.; Qi, G.; Liu, W.; Yang, J. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, 1–8. [[CrossRef](#)]
62. Yao, Y.; Zhang, S.; Yang, S.; Gui, G. Learning attention representation with a multi-scale CNN for gear fault diagnosis under different working conditions. *Sensors* **2020**, *20*, 1233. [[CrossRef](#)]
63. Li, T.; Shi, J.; Li, X.; Wu, J.; Pan, F. Image encryption based on pixel-level diffusion with dynamic filtering and DNA-level permutation with 3D Latin cubes. *Entropy* **2019**, *21*, 319. [[CrossRef](#)]
64. Veluchamy, M.; Subramani, B. Image contrast and color enhancement using adaptive gamma correction and histogram equalization. *Optik* **2019**, *183*, 329–337. [[CrossRef](#)]
65. Zhuang, L.; Guan, Y. Image enhancement via subimage histogram equalization based on mean and variance. *Comput. Intell. Neurosci.* **2017**, *2017*, 12. [[CrossRef](#)]
66. Khan, S.; Lee, D.H. An adaptive dynamically weighted median filter for impulse noise removal. *EURASIP J. Adv. Signal. Process.* **2017**, *67*, 14. [[CrossRef](#)]
67. Erkan, U.; Gökrem, L.; Enginoğlu, S. Different applied median filter in salt and pepper noise. *Comput. Electr. Eng.* **2018**, *70*, 789–798. [[CrossRef](#)]
68. Ahmed, A.; Jalal, A.; Kim, K. RGB-D images for object segmentation, localization and recognition in indoor scenes using feature descriptor and Hough voting. In Proceedings of the IEEE IBCAST, Islamabad, Pakistan, 14–18 January 2020. [[CrossRef](#)]
69. Jalal, A.; Quaid, M.A.K.; Tahir, S.B.u.d.; Kim, K. A study of accelerometer and gyroscope measurements in physical life-log activities detection systems. *Sensors* **2020**, *20*, 6670. [[CrossRef](#)] [[PubMed](#)]
70. Jalal, A.; Batool, M.; Kim, K. Sustainable wearable system: Human behavior modeling for life-logging activities using K-ary tree hashing classifier. *Sustainability* **2020**, *12*, 10324. [[CrossRef](#)]
71. Truong, M.T.N.; Kim, S. Automatic image thresholding using Otsu’s method and entropy weighting scheme for surface defect detection. *Soft Comput.* **2018**, *22*, 4197–4203. [[CrossRef](#)]
72. Rizwan, S.A.; Jalal, A.; Kim, K. An accurate facial expression detector using multi-landmarks selection and local transform features. In Proceedings of the International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020. [[CrossRef](#)]
73. Abid Hasan, S.M.; Ko, K. Depth edge detection by image-based smoothing and morphological operations. *J. Comput. Des. Eng.* **2016**, *3*, 191–197. [[CrossRef](#)]
74. Treister, E.; Haber, E. A fast marching algorithm for the factored eikonal equation. *J. Comput. Phys.* **2016**, *324*, 210–225. [[CrossRef](#)]
75. Garrido, S.; Alvarez, D.; Moreno Luis, E. Marine applications of the fast marching method. *Front. Robot. AI* **2020**, *7*, 2. [[CrossRef](#)]
76. Jalal, A.; Nadeem, A.; Bobasu, S. Human body parts estimation and detection for physical sports movements. In Proceedings of the International Conference on Communication, Computing and Digital Systems (C-CODE), Islamabad, Pakistan, 6–7 March 2019. [[CrossRef](#)]
77. Nguyen, N.; Bui, D.; Tran, X. A novel hardware architecture for human detection using HOG-SVM co-optimization. In Proceedings of the APCCAS, Bangkok, Thailand, 11–14 November 2019. [[CrossRef](#)]
78. Muralikrishna, S.N.; Muniyal, B.; Dinesh Acharya, U.; Holla, R. Enhanced human action recognition using fusion of skeletal joint dynamics and structural features. *J. Robot.* **2020**, *2020*, 16. [[CrossRef](#)]

79. Abdul-Azim, H.A.; Hemayed, E.E. Human action recognition using trajectory-based representation. *Egypt. Inform. J.* **2015**, *16*, 187–198. [[CrossRef](#)]
80. Papadopoulos, K.; Demisse, G.; Ghorbel, E.; Antunes, M.; Aouada, D.; Ottersten, B. Localized trajectories for 2D and 3D action recognition. *Sensors* **2019**, *19*, 3503. [[CrossRef](#)] [[PubMed](#)]
81. Ouyed, O.; Allili, M.S. Group-of-features relevance in multinomial kernel logistic regression and application to human interaction recognition. *Expert Syst. Appl.* **2020**, *148*, 113247. [[CrossRef](#)]
82. Wang, D.; Tan, D.; Liu, L. Particle swarm optimization algorithm: An overview. *Soft Comput.* **2018**, *22*, 387–408. [[CrossRef](#)]
83. Berlin, S.J.; John, M. Particle swarm optimization with deep learning for human action recognition. *Multimed. Tools Appl.* **2020**, *79*, 17349–17371. [[CrossRef](#)]
84. Do, Q.H.; Chen, J.-F. A neuro-fuzzy approach in the classification of students' academic performance. *Comput. Intell. Neurosci.* **2013**, *2013*, 179097. [[CrossRef](#)]
85. Mohamed, G.; Lotfi, A.; Pourabdollah, A. Human activities recognition based on neuro-fuzzy finite state machine. *Technologies* **2018**, *6*, 110. [[CrossRef](#)]
86. Subramanian, K.; Suresh, S.; Sundararajan, N. A metacognitive neuro-fuzzy inference system (McFIS) for sequential classification problems. *IEEE Trans. Fuzzy Syst.* **2013**, *21*, 1080–1095. [[CrossRef](#)]
87. Shahroudy, A.; Liu, J.; Ng, T.; Wang, G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [[CrossRef](#)]
88. Coppola, C.; Cosar, S.; Faria, D.R.; Bellotto, N. Social activity recognition on continuous RGB-D video sequences. *Int. J. Soc. Robot.* **2020**, *12*, 201–215. [[CrossRef](#)]
89. Choi, W.; Shahid, K.; Savarese, S. What are they doing? Collective activity classification using spatio-temporal relationship among people. In Proceedings of the International Conference on Computer Vision Workshops (ICCV), Kyoto, Japan, 27 September–4 October 2009. [[CrossRef](#)]
90. Lee, J.; Ahn, B. Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors* **2020**, *20*, 2886. [[CrossRef](#)]
91. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3D pose estimation and action recognition using multitask deep learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [[CrossRef](#)]
92. Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017. [[CrossRef](#)]
93. Coppola, C.; Faria, D.R.; Nunes, U.; Bellotto, N. Social activity recognition based on probabilistic merging of skeleton features with proximity priors from RGB-D data. In Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Daejeon, South Korea, 9–14 October 2016. [[CrossRef](#)]
94. Manzi, A.; Fiorini, L.; Limosani, R.; Dario, P.; Cavallo, F. Two-person activity recognition using skeleton data. *IET Comput. Vis.* **2018**, *12*, 27–35. [[CrossRef](#)]
95. Ehatisham-Ul-Haq, M.; Javed, A.; Awais, M.A.; Hafiz, M.A.M.; Irtaza, A.; Hyun, I.L.; Tariq, M.M. Robust human activity recognition using multimodal feature-level fusion. *IEEE Access* **2019**, *7*, 60736–60751. [[CrossRef](#)]
96. Deng, Z.; Vahdat, A.; Hu, H.; Mori, G. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [[CrossRef](#)]
97. Choi, W.; Savarese, S. A unified framework for multi-target tracking and collective activity recognition. In Proceedings of the ECCV LNCS, Berlin, Germany, 23–28 August 2012. [[CrossRef](#)]