Computers, Materials & Continua
DOI: 10.32604/cmc.2022.024422
Article

Tech Science Press

An Intelligent HealthCare Monitoring Framework for Daily Assistant Living

Yazeed Yasin Ghadi[1], Nida Khalid[2], Suliman A. Alsuhibany[3], Tamara al Shloul[4], Ahmad Jalal[2] and Jeongmin Park[5,*]

[1]Department of Computer Science and Software Engineering, Al Ain University, Al Ain, 15551, UAE
[2]Department of Computer Science, Air University, Islamabad, 44000, Pakistan
[3]Department of Computer Science, College of Computer, Qassim University, Buraydah, 51452, Saudi Arabia
[4]Department of Humanities and Social Science, Al Ain University, Al Ain, 15551, UAE
[5]Department of Computer Engineering, Korea Polytechnic University, Siheung-si, Gyeonggi-do, 237, Korea
*Corresponding Author: Jeongmin Park. Email: jmpark@kpu.ac.kr
Received: 16 October 2021; Accepted: 12 January 2022

Abstract: Human Activity Recognition (HAR) plays an important role in life care and health monitoring since it involves examining various activities of patients at homes, hospitals, or offices. Hence, the proposed system integrates Human-Human Interaction (HHI) and Human-Object Interaction (HOI) recognition to provide in-depth monitoring of the daily routine of patients. We propose a robust system comprising both RGB (red, green, blue) and depth information. In particular, humans in HHI datasets are segmented via connected components analysis and skin detection while the human and object in HOI datasets are segmented via saliency map. To track the movement of humans, we proposed orientation and thermal features. A codebook is generated using Linde-Buzo-Gray (LBG) algorithm for vector quantization. Then, the quantized vectors generated from image sequences of HOI are given to Artificial Neural Network (ANN) while the quantized vectors generated from image sequences of HHI are given to K-ary tree hashing for classification. There are two publicly available datasets used for experimentation on HHI recognition: Stony Brook University (SBU) Kinect interaction and the University of Lincoln's (UoL) 3D social activity dataset. Furthermore, two publicly available datasets are used for experimentation on HOI recognition: Nanyang Technological University (NTU) RGB-D and Sun Yat-Sen University (SYSU) 3D HOI datasets. The results proved the validity of the proposed system.

Keywords: Artificial neural network; human-human interaction; human-object interaction; k-ary tree hashing; machine learning

1 Introduction

Recent years have seen an advanced use of multi-vision sensors to attain robustness and high-performance rates while tackling many of the existing challenges in visual recognition systems [1].

Moreover, low-cost depth sensors such as Microsoft Kinect [2] are being used extensively ever since their introduction. In comparison with conventional visual systems, depth maps are unaffected by varying brightness and lighting conditions [3] which motivate reflection over a wide variety of applications of Human Activity Recognition (HAR). These applications include assisted living, behaviour understanding, security systems, human-robot interactions, e-health care, smart homes, and others [4].

To monitor the daily lifecare routine of humans thoroughly, this paper proposes a system that integrates the recognition of Human-Human Interaction (HHI) and Human-Object Interaction (HOI). In the proposed system, the silhouette segmentation of red, green, blue (RGB) and depth images from HHI and HOI datasets is carried out separately. After silhouette segmentation, there is the feature extraction phase which consists of mining two unique features, namely thermal and orientation features. Both HHI and HOI descriptors are combined and processed via Linde-Buzo-Gray (LBG) algorithm for compact vector representation. In the end, K-ary tree hashing is used for the classification of HHI classes, while Artificial Neural Network (ANN) is applied for the classification of HOI classes.

We have used two publicly available datasets for experimentation on HHI recognition: Stony Brook University (SBU) Kinect interaction and the University of Lincoln's (UoL) 3D social activity datasets. Furthermore, we have used two different publicly available datasets for experimentation on HOI recognition: Nanyang Technological University (NTU) RGB+D and Sun Yat-Sen University (SYSU) 3D HOI datasets.

The main contributions of this paper are:

- Developing an efficient way of segmenting human silhouettes from both RGB and depth images via connected components, skin detection, morphological operations, and saliency maps.
- Designing a high-performance recognition system based on the extraction of unique orientation and thermal features.
- Accurate classification of HHI classes via K-ary tree hashing and HOI classes via ANN.

The rest of this paper is organized as follows: Section 2 explains and analyzes the research work relevant to the proposed system. Section 3 describes the proposed methodology of the system which involves an extensive pre-classification process. Section 4 describes the datasets used in the proposed work and proves the robustness of the system through different experiments. Section 5 concludes the paper and notes some future works.

2 Related Work

The related work can be divided into two subsections including some recently developed recognition systems for both HHI and HOI.

2.1 HHI Recognition Systems

In recent years, many RGB-D (red, green, blue, and depth) human-human interaction recognition systems have been proposed [5]. Prati et al. [6] proposed a system in which multiple camera views were used to extract features from depth maps using regression learning. However, despite the use of multiple cameras, their system had restricted applicability on large areas and was not robust against occlusion. Q. Ye et al. [7] proposed a system comprising of Gaussian time-phase features using ResNet (Residual Network). A high performance rate was achieved but their system also had a high complexity rate. In [8], Ouyed et al. extracted motion features from the joints of two persons involved in an

interaction. They used multinomial kernel logistic regression to evaluate HIR but the system lacked spatiotemporal context for interaction recognition. Moreover, Ince et al. [9] proposed a system based on skeletal joints movement using Haar-wavelet. In this system, some confusion was observed due to the similarities in angles and positions of various actions. Furthermore, Bibi et al. [10] proposed an HIR system with local binary patterns using multi-view cameras. A high confusion rate was observed in similar interactions.

Yanli et al. [11] proposed an HIR system that benefits from contrastive feature distribution. The authors extracted skeleton-based features and calculated the probability distribution. In [12], Subetha et al. extracted Histogram of Oriented Gradient (HOG) and pyramidal features. Besides, a study in [13] presented an action recognition system based on way-points trajectory, geodesic distance, joints motion and 3D Cartesian-plane features. This system achieved better performance but there was a slight decrease in accuracy due to the factor of silhouette overlapping. In addition to this study, an action representation was performed with shape, spatio-temporal angular-geometric and energy-based features [14]. Therefore, a high recognition rate was achieved but the performance of the system is reduced in those environments where human posture changes rapidly. Also, human postures were extracted in [15] using an unsupervised dynamic X-means clustering algorithm. Features were extracted from skeleton joints obtained via depth sensors. The system lacked in the identification of static actions. Waheed et al. [16] generated 3D human postures and obtained their heat kernels to identify key body points. Then they extracted topological and geometric features using these key points. The authors also extracted full body features using CNN [17]. In [18], a time interval at which social interaction is performed was detected and spatio-temporal and social features were extracted to track human actions. Moreover, a study in [19] proposed a fusion of multiple sensors. That is, they extracted HOG and statistical features. However, this system only worked on pre-segmented activities.

2.2 HOI Recognition Systems

Various methodologies have been adopted by researchers for identifying human activities in the past few years [20]. For example, Meng et al. [21] proposed an HOI recognition system based on inter-joint and joint-object distances. However, they did not identify the object individually but considered it one of the human body joints. In [22], joint distances that were invariant to the human pose were measured for feature extraction. This system was tested with only one dataset that consists of six simple interactions. Jalal et al. [23] proposed a HAR system based on the frame differentiation technique. In this system, human joints were extracted to record the spatio-temporal movements of humans. Moreover, Yu et al. [24] proposed a discriminative orderlet, i.e., a middle-level feature that was used for the visual representation of actions. A cascaded HOI recognition system was proposed by Zhou et al. [25]. It was a multi-stage architecture in which each stage of HOI was refined and then fed to the next network. In [26], zero-shot learning was used to accurately identify a relationship between a verb and an object. Their system lacked a spatial context and was tested on a simple verb-object pair. All the methodologies mentioned above in related work are either tested on RGB data or represented by a very complex set of features which increases the time complexity of the system.

Inspired by these approaches and mindful of their limitations, the proposed system has been designed. Because of the high accuracies achieved by systems that used depth sensors or RGB-D data, our system also takes RGB-D input. Different researchers have extracted different types of features from human silhouettes. To make our model unique, we have chosen orientation and thermal features. Moreover, these features are robust against occlusion and rapid posture changes, the two major issues faced by most researchers. Furthermore, the approaches that have used multiple features have increased time complexity. To solve this issue, we used vector quantization. It was also noted that systems tested

on only one dataset or limited number of classes fail to prove their general applicability. Therefore, the proposed system has been validated on four large datasets including two HHI and two HOI datasets.

3 Material and Methods

This section describes the proposed framework for active monitoring of the daily life of humans. Fig. 1 shows the general overview of the proposed system architecture. This architecture is explained in the following sections.

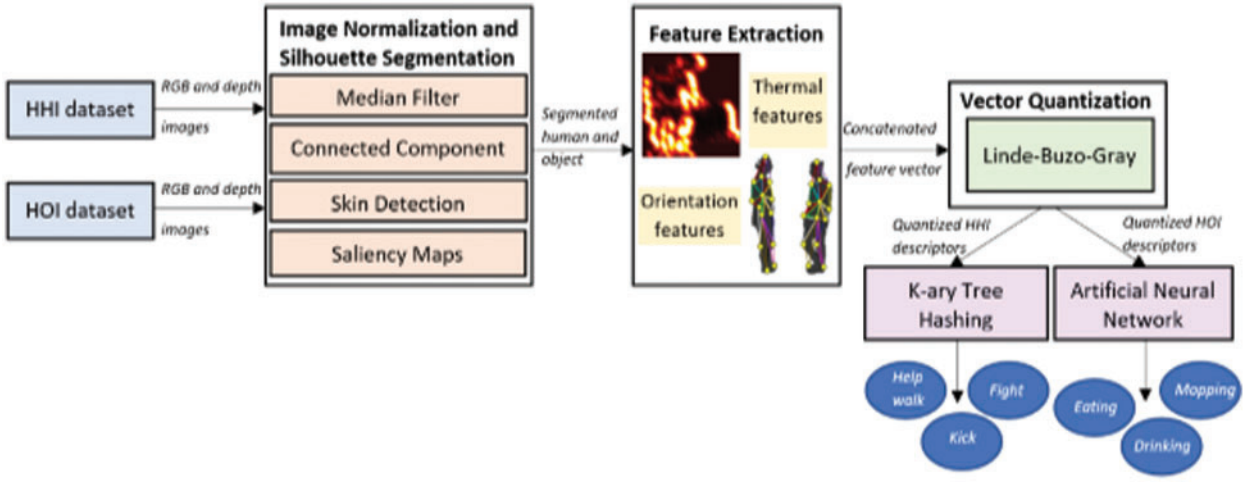

Figure 1: System architecture of the proposed system

3.1 Image Normalization and Silhouette Segmentation

The images in HHI and HOI datasets are first filtered to enhance the image features. Then, a median filter is applied to both RGB and depth image sequences to remove noise [27] using the formula in Eq. (1):

$$y[m,n] = median\{x[i,j], (i,j) \in w\} \tag{1}$$

where i and j belong to a window w a having specified neighborhood centered around the pixels $[m, n]$ in an image.

3.1.1 Silhouette Segmentation of Human-Human Interactions

The silhouette segmentation of RGB and depth frames of HHIs is performed separately. At first, connected components are located in an image via 4-connected pixel analysis as given through Eq. (2):

$$N_4(p) = \{(x+1,y), (x-1,y), (x,y+1), (x,y-1)\} \tag{2}$$

where x and y are coordinates of pixel p. After labeling of connected components, a threshold limit that determines the area (height and width) of the human body is specified. Then, a bounding box is drawn on only those labeled components that are within the specified limit. As a result, all the humans in a frame are identified and enclosed in a bounding box. After identification, human skin is detected inside a bounding box via HSV (hue, saturation, value) color model [28]. To improve the thresholding process, all the light color intensities (white, skin, and yellow) are converted into black color having

an intensity value of *0*. Then threshold-based segmentation is applied to generate binary silhouettes. Silhouette segmentation of RGB images is shown in Fig. 2.

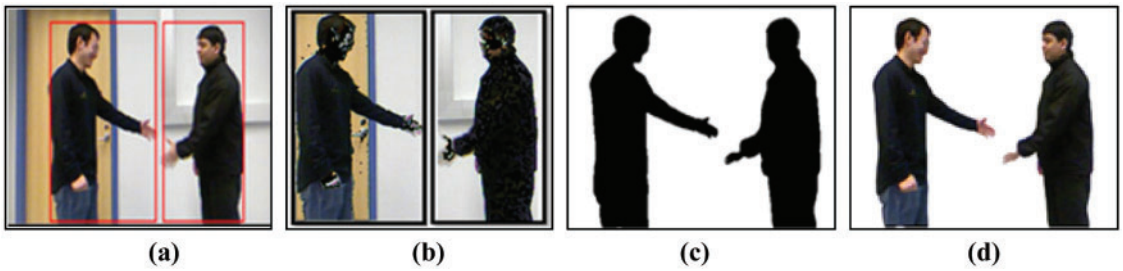

Figure 2: Silhouette segmentation of an HHI image. (a) Detected humans via connected components, (b) skin color detection of right and left human, (c) binary silhouette, and (d) RGB silhouette

The segmentation of depth images of HHI is performed via morphological operations [29]. The first step is to convert a depth image into a binary image via Otsu's thresholding method. Then morphological operations of dilation and erosion are applied which result in retaining the contour of key objects in an image. The Canny edge detection technique is then applied to detect the edges of the two people involved in HHI. In the end, the smaller detected objects are removed from the image.

3.1.2 Silhouette Segmentation of HOI

Spectral residual saliency map-based segmentation technique [30] is used to segment humans and objects from RGB and depth images. The saliency map is generated by evaluating the log spectrum of images and Fourier transform is used to obtain the frequency f of the grayscale images $G(x)$. The amplitude $A(f) = abs(f)$ and phase spectrum $P(f) = angle(f)$ are computed through frequency image. Then, spectral residual regions of RGB and depth images are computed from Eq. (3) as follows:

$$R(f) = L(f) - h(f) \times L(f) \tag{3}$$

where $L(f)$ is the log of $A(f)$ and $h(f)$ is an averaging filter. After residual regions are computed, saliency map S is generated from Eq. (4) as follows:

$$S = F^{-1}[\exp((R(f) + P(f))]^2 \tag{4}$$

where F^{-1} is inverse Fourier transform. In the end, the saliency map is converted into a binary image via the binary thresholding method. After segmenting salient regions from the background, humans and objects are detected separately in an image using K-means algorithm. This process of silhouette segmentation on a depth image is depicted in Fig. 3.

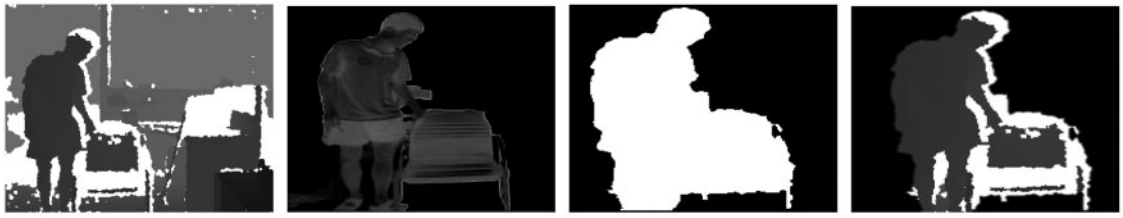

Figure 3: Silhouette segmentation of an HOI image. (a) Original depth image, (b) saliency map, (c) binary silhouette, and (d) segmented depth human and object silhouette

3.2 Feature Extraction

The proposed system exploits two types of features: orientation features and thermal features. The details and results of these features are described in the following subsections.

3.2.1 Orientation Features

After obtaining human silhouettes, fourteen human joints (head, neck, right shoulder, left shoulder, right elbow, left elbow, right hand, left hand, torso, torso base, right knee, left knee, right foot, and left foot) are identified. Eight human joints are identified by Algorithm 1 (detection of key-body points in human silhouette) proposed in [14]. In this algorithm, eight human joints are detected by finding the topmost, bottommost, rightmost and leftmost pixels from the boundary of a human silhouette. The rest of the six human joints are identified by taking the average of the pixel locations of already identified joints. For example, the location of the neck joint is identified by taking the mean of the location of the head and torso joint. After locating joint points, a combination of three joints is taken to form a triangular shape and as a result, fourteen triangles are formed in HOI images (See Fig. 4a). In HOI silhouettes, the orientation features of objects are also extracted. Four triangles (twelve angles) are formed from the centroid to all the four extreme points (See Fig. 4b). While in HHI images, two people are involved so the number of triangles is twenty-eight (fourteen for each person) as shown in Fig. 4c. The angle of tangent is measured between three sides of each triangle from Eq. (5) as follows:

$$tan\theta = \frac{u.v}{|u||v|} \tag{5}$$

where $u.v$ are obtained by taking a dot product of two vectors u *and* v which are any two sides of a triangle. A total of three angles are calculated for each triangle. The first angle is calculated by taking AB as u and AC as v. The second angle is calculated by taking AB as u and BC as v. The third angle is calculated by taking BC as u and AC as v as shown in Fig. 4d.

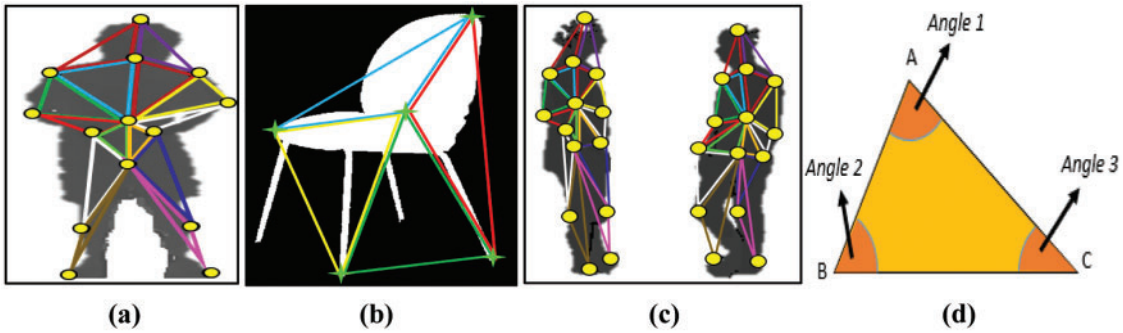

Figure 4: Triangular shapes formed by combining human joints. (a) Single person frame of playing phone HOI in SYSU dataset, (b) triangle formation on object, (c) two-person frame of conversation HHI in UoL dataset, and (d) three angles of a triangle

3.2.2 Thermal Features

The movements of different body parts as the human silhouettes move from one frame to the next are captured in the form of thermal maps. The parts having greater movements during an interaction have higher heat values (yellowish colors) on these maps. On the other hand, parts of a human silhouette that show lesser movements, i.e., they are less involved in performing an interaction, are displayed in a reddish or blackish color. A matrix of index values ranging from 0 to 8000 shows heat values in thermal maps. These index values are used to extract heat of only those parts that are involved in an HOI and are represented from Eq. (6) as follows:

$$TM(v) = \sum_{0}^{K} \ln R(K) \tag{6}$$

where v is a 1D vector in which the extracted values are stored, K represents index values and ln R refers to the RGB values that are extracted from K. The thermal maps of different HOI are shown in Fig. 5.

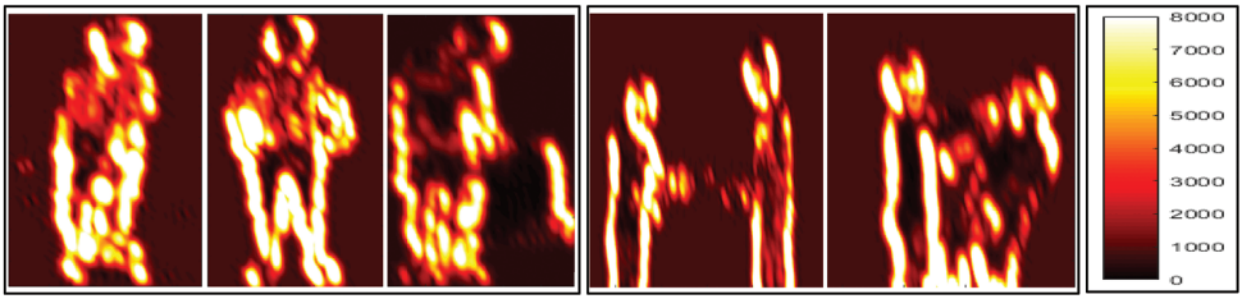

Figure 5: Thermal maps of HOI and HHI along with the scale of thermal values

After extracting the two types of features from both HHI and HOI datasets, they are concatenated, resulting in a matrix.

3.3 Vector Quantization

After extracting the two types of features from all the images of the HHI and HOI datasets, the features are added as descriptors of each interaction class, separately. However, this results in a very large feature dimension size [31]. Therefore, we generate an organized feature vector by considering a codebook of size 512.

3.4 HHI Classification via K-ary Tree Hashing

The quantized feature vectors of HHI classes are fed to the K-ary tree hashing classifier. The optimized features are represented in the form of a graph $G = \{g_i\}$, where $i = 1 \ldots N$ and N represents the number of objects in the graph [32]. The graph comprises of vertices V and undirected edges E. Moreover there is a label function $l: V \rightarrow L$ to assign labels to nodes in g_i where g_i represents the whole graph with V, E, and l. A class label y_i is set for each g_i based on graph's structure. The graph structure means the values in the nodes of the graph based on the values of feature vectors. Each feature vector of the testing class is represented in the form of a graph and then used for predicting the correct label for each class. Also, a size of MinHashes $\{D^{(r)}\}_{r=1}^{R}$ for R iterations and the traversal table K is also defined. For MinHashes, random permutations $\{\pi_d^{(r)}\}$ are also generated. The process of classifying various HHIs is given in Algorithm 2 which takes the graph, the traversal table, and the size of MinHashes as input. It is divided into four sections: traversal table construction, MinHash selection, recursive leaf extension, and leaf sequence. The traversal table is constructed to find the subtree patterns in using k-ary trees. Like a binary tree in which each node has two children, each node in a k-ary tree has k children. The leaf node labels of the k-ary trees can identify the patterns hidden in the data. Then the MinHash scheme is used to classify the interactions based on the identified patterns.

Algorithm 1: HHI Classification via K-ary tree hashing

Input: $g_i = (V, E, l)$, K, $\{D^{(r)}\}_{r=1}^{R}$
// R specifies total iterations//
Output: $\{x^{(r)}\}_{r=1}^{R}$
//Traversal Table Construction//
1 $V \leftarrow |V|$
2 $l(V+1) \leftarrow \infty$
3 $T \leftarrow (V+1) * ones(V+1, 1+K)$
for v = 1: V **do**
4 $N_v \leftarrow neighbour(v)$
//MinHash Selection//
5 $temp \leftarrow sort(l(N_v))$
6 $k \leftarrow \min(K, |N_v|)$
7 $T(v) \leftarrow [v, index(temp(1{:}k))]$
end for
// Recursive Leaf Extension//
8 $z^{(1)} \leftarrow [1{:}V]^T$
9 $S^{(1)} \leftarrow l(z^{(1)})$
for i = 1: R **do**
if r > 1 **then**
10 $z^{(r)} \leftarrow reshape(T(z^{(r-1)}, :), [1, *])$

(Continued)

Algorithm 1: Continued

$S^{(r)} \leftarrow reshape(l(z^{(r)}), [V, *])$
end if
//Leaf Sequence//
$f^{(r)} \leftarrow [h(S^{(r)}(1,:)), \ldots h(S^{(r)}(V,:))]^T$
$x^{(r)} \leftarrow [min(\pi_1^{(r)}(f^{(r)})), \ldots min(\pi_{D^{(r)}}^{(r)}(f^{(r)}))]^T$
end for

3.5 HOI Classification via Artificial Neural Networks

The Quantized vectors of HOI are then fed to ANN for training and predicting accurate results. The final vector dimension of the SYSU 3D HOI dataset is 6054×480 while for NTU RGB+D dataset is 6054×530. There are 6054 rows that represent the feature values for both thermal and orientation features. Whereas there are 480 and 530 columns representing the number of images in the SYSU 3D HOI and NTU RGB+D datasets respectively. In the LOSO validation technique, one subset is used for testing and all the remaining subsets are used to train the system. The system is then validated by taking another subset for testing and the remaining subsets for training. In this way, the system is trained and tested with all the subjects in both datasets and avoids sampling bias. There are three layers: input layer, hidden layer, and output layer in ANN [33]. These layers are interconnected to each other and weights are associated with each connection. The net input at the neuron of each layer is computed using a transfer function T_j given in Eq. (7) as follows:

$$T_j = \sum_i w_{i,j} \times x_i + b_j \tag{7}$$

where $w_{i,j}$ are the weights, x_i represents the inputs and b_j is the added bias term. An input layer is fed with feature descriptors. After adjusting weights, adding bias, and processing through hidden layer, it predicts accurate HOI classes of both datasets.

4 Performance Evaluation

This section gives a brief description of the four datasets used for HHI and HOI, results of the experiments conducted to evaluate the proposed HAR system and its comparison with other systems.

4.1 Datasets Description

The description of each dataset used for HHI recognition and HOI recognition is given in Tab. 1. Each class of HHI and HOI dataset is performed by different number of subjects as described in the dataset description table. So the proposed system is trained with the different number of subjects of varying appearances resulting in high-performance rate in the testing phase. Each subject is used for both training and testing of a system via the LOSO validation technique. Cross-validation is used to avoid sampling bias via using new image classes for testing of a system other than those used for training.

Table 1: Datasets description for HHI and HOI recognition

Dataset name	Type of interaction	Description
SBU Kinect interaction dataset [34]	Eight RGB-D human-human interactions	SBU Kinect interaction dataset is a two-person interaction dataset consisting of 8 interaction classes, i.e., approaching, departing, kicking, punching, pushing, shaking hands, exchanging object, and hugging. Each class of the SBU dataset is performed by 21 subjects. Further details of the dataset are given in [34].
Uol 3D activity dataset [35]	Eight RGB-D human-human interactions	Kinect 2 sensor is used to capture eight interactions that are: handshake, hug, help walk, help stand-up, fight, push, conversation, and call attention. Each class of the UoL dataset is performed by six different persons. The rest of the details are given in [35].
SYSU 3D HOI dataset [36]	Twelve RGB-D human-object interactions	A Kinect sensor is used to collect RGB and depth images. Twelve human-object activities performed in this dataset are sweeping, mopping, taking from a wallet, taking out wallet, moving chair, sitting chair, packing backpacks, wearing backpacks, playing phone, calling phone, pouring, and drinking. Each class of the SYSU 3D HOI dataset is performed by forty subjects. Details of the dataset are given in [36].

(Continued)

Table 1: Continued

Dataset name	Type of interaction	Description
NTU RGB+D dataset [37]	Ten RGB-D human-object interactions	We used 3781 video samples of twelve human-object interactions for experimentation such as drink water, eat a meal, tear up paper, put on the jacket, take off the jacket, put on a hat/cap, take off a hat/cap, phone call, play with phone/tablet and taking a selfie. There are 56,880 video samples provided for 60 action classes performed by many different number of subjects. The rest of the details are given in [37].

4.2 Experimental Settings and Results

All the processing and experimentation are performed on MATLAB (R2018a). The hardware system used is Intel Core i5 with 64-bit Windows-10. The system has an 8 GB ram and 5 (GHz) CPU. To evaluate the performance of the proposed system, we used a Leave One Subject Out (LOSO) cross-validation method. The results section is divided into two sections: experimental results on HHI datasets and experimental results on HOI datasets.

4.2.1 Experimental Results on HHI Datasets

Experiment I: Recognition Accuracies

At first, classes of SBU and UoL datasets are given to the K-ary tree hashing classifier separately. The results of classification with classes of the SBU and UoL dataset is shown in the form of confusion matrices in Tabs. 2 and 3 respectively.

Experiment II: Precision, Recall and F1 Measures

The precision is the ratio of correct positive predictions to the total positives while the recall is the true positive rate and it is the ratio of correct positives to the total predicted positives. The F1 score is the mean of precision and recall. The precision, recall and F1 score for classes of SBU and UoL dataset are given in Tabs. 4 and 5 respectively.

Table 2: Confusion matrix showing recognition accuracies over classes of SBU dataset

Interaction classes	Approaching	Departing	Kicking	Punching	Pushing	Hugging	EO	SH
Approaching	**0.93**	0.02	0	0	0	0	0.01	0.04
Departing	0.03	**0.93**	0	0	0	0.01	0.02	0.01
Kicking	0	0	**0.97**	0.01	0.02	0	0	0
Punching	0	0	0.03	**0.95**	0.02	0	0	0
Pushing	0	0.01	0.01	0.05	**0.92**	0.01	0	0
Hugging	0.01	0	0	0.02	0.01	**0.96**	0	0
EO	0.03	0	0	0	0.01	0	**0.91**	0.05
SH	0.05	0	0	0	0	0	0.05	**0.90**

Mean recognition accuracy rate = 93.38%

Note: EO = exchanging object, SH = shaking hands.

Table 3: Confusion matrix showing recognition accuracies over classes of UoL dataset

Interaction classes	Handshake	Hug	Help walk	Help stand-up	Fight	Push	Conversation	Call attention
Handshake	**0.94**	0.02	0.03	0	0	0	0.01	0
Hug	0	**0.96**	0.02	0	0	0.02	0	0
Help walk	0	0.02	**0.94**	0.04	0	0	0	0
Help stand-up	0.02	0	0.04	**0.94**	0	0	0	0
Fight	0	0.01	0	0	**0.96**	0.02	0.01	0
Push	0	0	0.01	0	0.04	**0.94**	0.01	0
Conversation	0	0	0	0.02	0.04	0	**0.89**	0.05
Call attention	0.03	0	0	0	0	0.4	0.05	**0.88**

Mean Recognition Accuracy rate = 93.1%

Table 4: Precision, Recall and F1 score over classes of SBU dataset

Interaction classes	Precision	Recall	F1 score	Interaction classes	Precision	Recall	F1 score
Approaching	0.89	0.93	0.91	Pushing	0.94	0.92	0.93
Departing	0.97	0.93	0.95	Hugging	0.98	0.96	0.97
Kicking	0.96	0.97	0.97	Exchanging Object	0.92	0.91	0.91
Punching	0.92	0.95	0.94	Shaking hands	0.90	0.90	0.90

Mean precision = 0.935 Mean recall = 0.933 Mean F1 score = 0.935

Table 5: Precision, Recall and F1 score over classes of UoL dataset

Interaction classes	Precision	Recall	F1 score	Interaction classes	Precision	Recall	F1 score
Handshake	0.95	0.94	0.94	Fight	0.92	0.96	0.94
Hug	0.95	0.96	0.96	Push	0.92	0.94	0.93
Help walk	0.90	0.94	0.92	Conversation	0.92	0.89	0.90
Help stand-up	0.94	0.94	0.94	Call attention	0.95	0.88	0.91
Mean precision = 0.931 Mean recall = 0.931 Mean F1 score = 0.930							

Experiment III: Comparison with Other Systems

This section compares the proposed methodology with other recent methods as shown in Fig. 6. These methods have been discussed in Section 2.

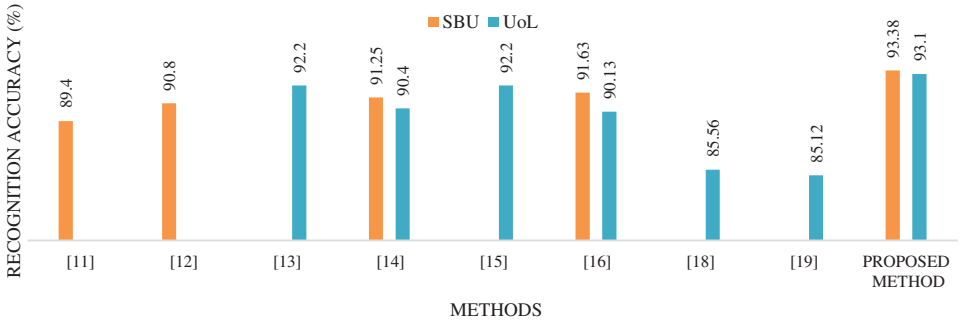

Figure 6: Comparison of mean recognition accuracy of the proposed method with other recent methods over HHI datasets

4.2.2 Experimental Results on HOI Datasets

Experiment I: Recognition Accuracies

The results of classification with classes of SYSU and NTU dataset are shown in the form of confusion matrices in Tabs. 5, 6 and 7, respectively. It is observed during experimentation that the interactions, which involve similar objects like packing backpacks and wearing backpacks, are confused with each other.

Experiment II: Precision, Recall and F1 Measures

The precision, recall and F1 scores for classes of the SYSU and the NTU dataset are given in Tabs. 8 and 9 respectively. Hence an accurate system is developed which is able to recognize each HOI with high precision.

Table 6: Confusion matrix showing recognition accuracies over classes of SYSU 3D HOI dataset

HOI classes	S	M	TFW	TOW	MC	SC	PB	WB	PP	CP	P	D
S	**0.89**	0.07	0	0	0.02	0.02	0	0	0	0	0	0
M	0.06	**0.90**	0	0	0.04	0	0	0	0	0	0	0
TFW	0	0	**0.88**	0.08	0	0	0.02	0	0.01	0.01	0	0
TOW	0	0	0.07	**0.89**	0	0	0	0	0.02	0	0.02	0
MC	0	0.01	0	0	**0.95**	0.04	0	0	0	0	0	0
SC	0	0	0	0	0.06	**0.94**	0	0	0	0	0	0
PB	0	0	0	0	0	0	**0.90**	0.08	0	0	0.02	0
WB	0	0	0	0	0	0	0.08	**0.89**	0	0	0.01	0.02
PP	0	0	0.03	0	0	0	0	0	**0.88**	0.07	0.02	0
CP	0	0	0	0	0	0	0	0.02	0.08	**0.87**	0.00	0.03
P	0	0	0	0	0	0	0.03	0	0.02	0	**0.90**	0.05
D	0	0	0	0	0	0	0	0	0	0.04	0.08	**0.88**

Mean recognition accuracy = 89.75%

Note: S = sweeping, M = mopping, TFW = taking from wallet, TOW = taking out wallet, MC = moving chair, SC = Sitting chair, PB = packing backpacks, WB = wearing backpacks, PP = playing phone, CP = calling pone, P = pouring, D = drinking.

Table 7: Confusion matrix showing recognition accuracies over classes of NTU RGB+D dataset

HOI classes	DW	EM	BT	BH	TP	PJ	TJ	PH	TH	PC	PP	TS
DW	**0.90**	0.06	0.04	0	0	0	0	0	0	0	0	0
EM	0.05	**0.91**	0.02	0.02	0	0	0	0	0	0	0	0
BT	0.02	0.03	**0.87**	0.08	0	0	0	0	0	0	0	0
BH	0.01	0.01	0.09	**0.89**	0	0	0	0	0	0	0	0
TP	0	0	0	0.01	**0.94**	0	0	0.02	0	0	0.03	0
PJ	0	0	0	0	0	**0.92**	0.06	0.02	0	0	0	0
TJ	0	0	0	0	0	0.07	**0.91**	0	0.02	0	0	0
PH	0	0	0	0	0	0.02	0	**0.90**	0.08	0	0	0
TH	0	0	0	0	0	0	0.02	0.07	**0.91**	0	0	0
PC	0	0	0	0.02	0	0	0	0	0	**0.91**	0.05	0.02
PP	0	0	0	0	0.02	0	0	0	0	0.04	**0.91**	0.03
TS	0	0	0	0	0	0	0	0	0	0.02	0.06	**0.92**

Mean recognition accuracy = 90.75%

Note: DW = drink water, EM = eat meal, BT = brush teeth, BH = brush hair, TP = tear up paper, PJ = put on jacket, TJ = take off jacket, PH = put on a hat, TH = take off a hat, PC = phone call, PP = play with phone, TS = taking a selfie.

Table 8: Precision, recall and F1 score over classes of SYSU dataset

Interaction classes	Precision	Recall	F1 score	Interaction classes	Precision	Recall	F1 score
Sweeping	0.94	0.89	0.91	Packing backpacks	0.87	0.90	0.89
Mopping	0.92	0.90	0.91	Wearing backpacks	0.90	0.89	0.89
Taking from wallet	0.90	0.88	0.89	Playing phone	0.87	0.88	0.88
Taking out wallet	0.92	0.89	0.90	Calling phone	0.88	0.87	0.87
Moving chair	0.89	0.95	0.92	Pouring	0.86	0.90	0.88
Sitting chair	0.94	0.94	0.94	Drinking	0.90	0.88	0.89
Mean precision = 0.899 Mean recall = 0.897 Mean F1 score = 0.897							

Table 9: Precision, Recall and F1 score over classes of NTU dataset

Interaction classes	Precision	Recall	F1 score	Interaction classes	Precision	Recall	F1 score
Drink water	0.92	0.90	0.91	take off jacket	0.92	0.91	0.91
Eat meal	0.90	0.91	0.91	put on a hat	0.89	0.90	0.90
Brush teeth	0.85	0.87	0.86	take off a hat	0.90	0.91	0.91
Brush hair	0.89	0.89	0.89	phone call	0.94	0.91	0.92
Tear up paper	0.98	0.94	0.96	play with phone	0.87	0.91	0.89
Put on jacket	0.89	0.92	0.91	taking a selfie	0.95	0.92	0.93
Mean precision = 0.907 Mean recall = 0.908 Mean F1 score = 0.907							

Experiment III: Comparison with Other Systems

This section compares the proposed methodology over HOI datasets with other recent methods as shown in Fig. 7. In [36], a RGB-D HOI system based on joint heterogeneous features based learning was proposed. Also, an RGB-D HOI system based on SIFT regression was proposed in [38]. A feature map was constructed by Local Accumulative Frame Feature (LAFF). Furthermore, a study in [39] explained graph regression, whereas multi-modality learning convolutional network was proposed in [40]. In [41], the skeletal joints extracted via depth sensors were represented in the form of key poses

and temporal pyramids. A mobile robot platform-based HIR was performed in [42] using skeleton-based features. Moreover, the overall human interactions are divided into interactions of different body parts [43]. In this work, pairwise features were extracted to track human actions. In [44], the authors introduced a semi-automatic rapid upper limb assessment (RULA) technique using Kinect v2 to evaluate the upper limb motion.

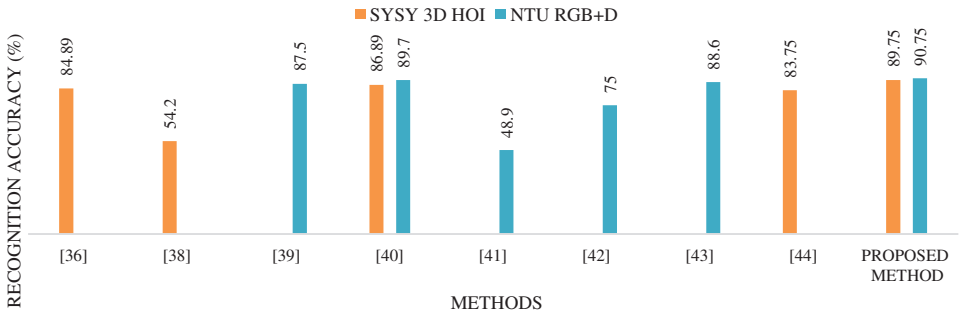

Figure 7: Comparison of mean recognition accuracy of different recent methods over HOI datasets

5 Discussion

A comparison of the proposed system with other systems showed that the proposed system performed better than many other systems proposed in the recent years. Moreover, the high accuracy scores justify the need of additional depth information along with RGB information. Similar findings were also presented in [16] and [17]. However, there are some limitations of the systems, such as during skeletal joints extraction, it was challenging to locate the joints of occluded body parts. In order to overcome this limitation, we have adopted the methodology of dividing the silhouette into four halves and then locating the top, bottom, left and right pixels for identifying joints in each half. Moreover, most of the interactions are performed in standing positions in the datasets used in the proposed system. Due to this reason, there is less occlusion of human body parts with objects or other body parts and the performance rate is not very much affected.

6 Conclusion and Future Works

This paper proposes a real-time human activity monitoring system that recognizes the daily activities of humans using multi-vision sensors. This system integrates two types of HAR systems: HHI recognition systems and HOI recognition systems. After silhouette segmentation, two unique features are extracted: thermal and orientation features. In order to validate the proposed system's performance, three types of experiments are performed. The comparison of the proposed system with other state-of-the-art systems is also provided which clearly shows the better performance of the proposed system. In real life, the proposed system should be applicable to many applications such as assisted living, behavior understanding, security systems and human-robot interactions, e-health care and smart homes.

We are working on integrating more types of human activity recognition and developing a system that monitors human behavior in both indoor and outdoor environments as part of our future works.

Funding Statement: This research was supported by a grant (2021R1F1A1063634) of the Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Education, Republic of Korea.

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

References

[1] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang *et al.,* "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, pp. 1–20, 2019.

[2] A. Jalal and S. Kamal, "Real-time life logging via a depth silhouette-based human activity recognition system for smart home services," in *Proc. Int. Conf. on Advanced Video and Signal Based Surveillance*, Seoul, South Korea, pp. 74–80, 2014.

[3] S. Kamal, A. Jalal and D. Kim, "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM," *Journal of Electrical Engineering and Technology*, vol. 11, no. 6, pp. 1921–1926, 2016.

[4] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu *et al.,* "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–40, 2021.

[5] M. Ajmal, F. Ahmad, M. Naseer and M. Jamjoom, "Recognizing human activities from video using weakly supervised contextual features," *IEEE Access*, vol. 7, pp. 98420–98435, 2019.

[6] A. Prati, C. Shan and K. I. -K. Wang, "Sensors, vision and networks: From video surveillance to activity recognition and health monitoring," *Journal of Ambient Intelligence and Smart Environments*, vol. 11, no. 1, pp. 5–22, 2019.

[7] Q. Ye, H. Zhong, C. Qu and Y. Zhang, "Human interaction recognition based on whole-individual detection," *Sensors*, vol. 20, no. 8, pp. 1–18, 2020.

[8] O. Ouyed and M. A. Said, "Group-of-features relevance in multinomial kernel logistic regression and application to human interaction recognition," *Expert Systems with Applications*, vol. 148, pp. 1–22, 2020.

[9] Ö. F. İnce, I. F. Ince, M. E. Yıldırım, J. S. Park, J. K. Song *et al.,* "Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor," *Electronics and Telecommunications Research Institute Journal*, vol. 42, no. 1, pp. 78–89, 2020.

[10] S. Bibi, N. Anjum and M. Sher, "Automated multi-feature human interaction recognition in complex environment," *Computers in Industry Elsevier*, vol. 99, pp. 282–293, 2018.

[11] Y. Ji, H. Cheng, Y. Zheng and H. Li, "Learning contrastive feature distribution model for interaction recognition," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 340–349, 2015.

[12] T. Subetha and S. Chitrakala, "Recognition of human-human interaction using CWDTW," in *Proc. Int. Conf. on Circuit, Power and Computing Technologies*, Nagercoil, India, pp. 1–5, 2016.

[13] N. Khalid, M. Gochoo, A. Jalal and K. Kim, "Modeling Two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system," *Sustainability*, vol. 13, no. 2, pp. 970, 2021.

[14] A. Jalal, N. Khalid and K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy markov model using depth sensors," *Entropy*, vol. 22, no. 8, pp. 1–33, 2020.

[15] A. Manzi, L. Fiorini, R. Limosani, P. Dario and F. Cavallo, "Two-person activity recognition using skeleton data," *Institute of Engineering and Technology Computer Vision*, vol. 12, no. 1, pp. 27–35, 2018.

[16] M. Waheed, A. Jalal, M. Alarfaj, Y. Y. Ghadi, T. Shloul *et al.,* "An LSTM-based approach for understanding human interactions using hybrid feature descriptors over depth sensors," *IEEE Access*, vol. 9, pp. 1–6, 2021.

[17] M. Waheed, M. Javeed and A. Jalal, "A novel deep learning model for understanding two-person interactions using depth sensors," in *Proc. Int. Conf. on Innovative Computing*, Lahore, Pakistan, 2021.

[18] C. Coppola, S. Cosar, D. R. Faria and N. Bellotto, "Automatic detection of human interactions from RGB-D data for social activity classification," in *Proc. Int. Conf. on Robot and Human Interactive Communication*, Lisbon, Portugal, pp. 871–876, 2017.
[19] M. E.-Haq, A. Javed, M. A. Awais, H. M. A. Malik, A. Irtaza *et al.,* "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019.
[20] Y. Gao, Z. Kuang, G. Li, W. Zhang and L. Lin, "Hierarchical reasoning network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8306–8317, 2021.
[21] M. Meng, H. Drira, M. Daoudi and J. Boonaert, "Human object interaction recognition using rate-invariant shape analysis of inter joint distances trajectories," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops*, Las Vegas, USA, pp. 999–1004, 2016.
[22] M. Meng, H. Drira, M. Daoudi and J. Boonaert, "Human-object interaction recognition by learning the distances between the object and the skeleton joints," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia, pp. 1–6, 2015.
[23] A. Jalal, S. Kamal and C. A. Azurdia-Meza, "Depth maps-based human segmentation and action recognition using full-body plus body color cues via recognizer engine," *Journal of Electrical Engineering and Technology*, vol. 14, pp. 455–461, 2019.
[24] G. Yu, Z. Liu and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Proc. Int. Asian Conf. on Computer Vision*, Singapore, pp. 50–65, 2014.
[25] T. Zhou, W. Wang, S. Qi, H. Ling and J. Shen, "Cascaded human-object interaction recognition," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 4262–4271, 2020.
[26] L. Shen, S. Yeung, J. Hoffman, G. Mori and L. Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. Int. Conf. on Applications of Computer Vision*, Lake Tahoe, NV, USA, pp. 1568–1576, 2018.
[27] G. George, R. M. Oommen, S. Shelly, S. S. Philipose and A. M. Varghese, "A survey on various median filtering techniques for removal of impulse noise from digital image," in *Proc. Int. Conf. on Emerging Devices and Smart Systems*, Tiruchengode, India, pp. 235–238, 2018.
[28] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat and J. Jatakia, "Human skin detection using RGB, HSV and YCbCr color models," in *Proc. Int. Conf. on Communication and Signal Processing*, Lonere, Raigad, India, pp. 324–332, 2016.
[29] S. M. A. Hasan and K. Ko, "Depth edge detection by image-based smoothing and morphological operations," *Journal of Computational Design and Engineering*, vol. 3, no. 3, pp. 191–197, 2016.
[30] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang *et al.,* "An iterative co-saliency framework for RGBD images," *IEEE Transactions on Cybernetics*, vol. 49, pp. 233–246, 2019.
[31] S. D. Thepade, R. H. Garg, S. A. Ghewade, P. A. Jagdale and N. M. Mahajan, "Performance assessment of assorted similarity measures in gray image colorization using LBG vector quantization algorithm," in *Proc. Int. Conf. on Industrial Instrumentation and Control*, Pune, India, pp. 332–337, 2015.
[32] K. Atighehchi and R. Rolland, "Optimization of tree modes for parallel Hash functions: A case study," *IEEE Transactions on Computers*, vol. 66, no. 9, pp. 1585–1598, 2017.
[33] P. Lubina and M. Rudzki, "Artificial neural networks in accelerometer-based human activity recognition," in *Proc. Int. Conf. on Mixed Design of Integrated Circuits & Systems*, Torun, Poland, pp. 63–68, 2015.
[34] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops*, RI, USA, pp. 28–35, 2012.
[35] C. Coppola, D. R. Faria, U. Nunes and N. Bellotto, "Social activity recognition based on probabilistic merging of skeleton features with proximity priors from RGB-D data," in *Proc. Int. Conf. on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, pp. 5055–5061, 2016.
[36] J. F. Hu, W. S. Zheng, J. Lai and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Transaction on PAMI*, vol. 39, no. 11, pp. 2186–2200, 2017.
[37] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan *et al.,* "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Transactions on PAMI*, vol. 42, no. 10, pp. 2684–2701, 2020.

[38] J. F. Hu, W. S. Zheng, L. Ma, G. Wang and J. Lai, "Real-time RGB-D activity prediction by soft regression," in *Proc. European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 280–296, 2016.
[39] X. Gao, W. Hu, J. Tang, J. Liu and Z. Guo, "Optimized skeleton-based action recognition via sparsified graph regression," in *Proc. Int. Conf. on Multimedia*, Nice, France, pp. 601–610, 2019.
[40] Z. Ren, Q. Zhang, X. Gao, P. Hao and J. Cheng, "Multi-modality learning for human action recognition," *Multimedia Tools and Applications*, vol. 20, pp. 1–19, 2020.
[41] E. Cippitelli, E. Gambi, S. Spinsante and F. F.-Revuelta, "Evaluation of a skeleton-based method for human activity recognition on a large-scale RGB-D dataset," in *Proc. Int. Conf. on Technologies for Active and Assisted Living*, London, UK, pp. 1–6, 2016.
[42] J. Lee and B. Ahn, "Real-time human action recognition with a low-cost RGB camera and mobile robot platform," *Sensors*, vol. 20, no. 10, pp. 1–12, 2020.
[43] M. Li and H. Leung, "Multi-view depth-based pairwise feature learning for person-person interaction recognition," *Multimedia Tools and Applications*, vol. 78, pp. 5731–5749, 2019.
[44] C. Wenming, J. Zhong, G. Cao and Z. He, "Physiological function assessment based on kinect V2," *IEEE Access*, vol. 7, pp. 105638–105651, 2019.