IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 4, 2021, accepted November 21, 2021, date of publication November 25, 2021, date of current version December 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3130613

An LSTM-Based Approach for Understanding Human Interactions Using Hybrid Feature Descriptors Over Depth Sensors

MANAHIL WAHEED[1], AHMAD JALAL[1], MOHAMMED ALARFAJ[2], YAZEED YASIN GHADI[3], TAMARA AL SHLOUL[4], SHAHARYAR KAMAL[1], AND DONG-SEONG KIM[5], (Senior Member, IEEE)

[1]Department of Computer Science, Air University, Islamabad 44000, Pakistan
[2]Department of Electrical Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[3]Department of Computer Science and Software Engineering, Al Ain University, Al Ain, United Arab Emirates
[4]Department of Humanities and Social Science, Al Ain University, Al Ain, United Arab Emirates
[5]Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, South Korea

Corresponding author: Dong-Seong Kim (dskim@kumoh.ac.kr)

This work was supported in part by the Priority Research Centers Program through the NRF by the MEST under Grant 2018R1A6A1A03024003; and in part by the Grand Information Technology Research Center Support Program supervised by the IITP by the MSIT, South Korea, under Grant IITP-2020-2020-0-01612.

ABSTRACT Over the past few years, automatic recognition of human interactions has drawn significant attention from researchers working in the field of Artificial Intelligence (AI). And feature extraction is one of the most critical tasks in developing efficient Human Interaction Recognition (HIR) systems. Moreover, recent researches in computer vision suggest that robust features lead to higher recognition accuracies. Hence, an improved HIR system has been proposed in this paper that combines 2D and 3D features extracted using machine learning and deep learning techniques. These discriminative features result in accurate classification and help avoid misclassification of similar interactions. Ten keyframes have been extracted from each video to reduce computational complexity. Next, these frames have been preprocessed using image normalization and noise removal techniques. The Region Of Interest (ROI), which contains the two humans involved in the interaction, has been extracted using motion detection. Then, the human silhouettes have been segmented using the GrabCut algorithm. Next, the extracted silhouettes have been converted into 3D meshes and their heat kernel signatures (HKS) have been obtained to extract key body points. A Convolutional Neural Network (CNN) has been used to extract full-body features from 2D full-body silhouettes. Then, topological and geometric features have been extracted from the key body points. Finally, the combined feature vector has been fed into Long Short-Term Memory (LSTM) and each interaction has been recognized using a Softmax classifier. The proposed system has been validated via extensive experimentation on three challenging RGB+D datasets. The recognition accuracies of 91.63%, 90.54%, and 90.13% have been achieved with the SBU Kinect Interaction, NTU RGB+D, and ISR-UoL 3D social activity datasets respectively. The results of extensive experiments performed on the proposed system suggest that it can be used effectively for various applications, such as security, surveillance, health monitoring, and assisted living.

INDEX TERMS 3-D mesh, depth videos, geodesic distance, heat kernel signature, human interaction recognition, RGB videos, topological features.

I. INTRODUCTION

The task of Human-Human Interaction (HHI) recognition involves detecting and understanding social interactions between two humans. These interactions can be everyday activities like talking, passing objects, hugging, and waving. Similarly, these can be assisted living activities such as helping a person stand up, helping another person walk, or drawing another person's attention. Moreover, suspicious activities including touching someone's pocket, pushing

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio Piccinno.

someone, or fighting are also of interest for researchers in this field. HIR has become a trending topic in the field of artificial intelligence because of its wide range of applications, including security [1]–[3], content-based video retrieval [4]–[6], healthcare [7]–[11], and surveillance [12]–[15].

Even though significant progress has been made in this regard and many efficient HIR systems have been developed for various purposes, detecting human interactions remains challenging because of multiple reasons, such as different viewpoints, change of clothing, poor lighting, different interactions containing similar motions, and unavailability of large datasets. However, low-cost depth sensors, such as Microsoft Kinect [16] are now being used excessively since these are not as affected by lighting conditions as RGB cameras. Moreover, many interactions seem similar and are often misclassified. For example, two humans exchanging a very small object may look very similar to two people shaking hands. On the contrary, the same interaction appears different when viewed from various viewpoints. Hence, it is very important to extract distinctive features from images that can easily differentiate between two interactions that look the same.

This research paper proposes a novel approach for efficient video-based human interaction recognition using both machine learning and deep learning techniques. Human silhouettes have been extracted from both RGB and depth frames using GrabCut. Additional masking has been used to improve the output of the GrabCut algorithm in complex scenarios. Next, the full-body RGB and depth silhouettes have been fed into two separate CNN models and the extracted features have been concatenated. Then 3D meshes have been generated from the full-body silhouettes and their heat kernel signatures have been obtained. These have been used to extract six key body points. These key points have been used to extract topological and geometric features. The three different types of features have been combined and fed into LSTM. Finally, the Softmax classifier has been used for interaction recognition. Three publically available datasets have been used that provide RGB, depth, and skeletal information of human interactions. The major contributions of this research work include:

- Silhouette segmentation from both RGB and depth images using GrabCut algorithm.
- Training and concatenation of two separate CNN models for RGB and depth images.
- 3-D mesh generation from 2-D silhouettes.
- Detection of key points via heat kernel signatures based on geodesic distance.
- Extraction of topological and geometric features using key body points.
- Extensive experimentation on three large and challenging RGBD video datasets.

Section II of the paper describes similar research work and the proposed system architecture has been discussed in Section III. Section IV presents the implementation details and results of the proposed method. Section V contains a discussion on various aspects of the designed system and section VI contains the conclusion of this paper and proposes future work of the authors.

II. RELATED WORK

Researchers have been actively contributing to the development of efficient HIR systems. The existing systems have been divided into two categories: marker-based systems and video-based systems. Researches falling into each category have been discussed in detail below:

A. MARKER-BASED HIR SYSTEMS

In marker-based HIR systems, different types of sensors, for example, reflective spheres, light-emitting diodes, and infrared markers, are mounted on the bodies of the humans whose movements are being monitored. These systems are commonly used for rehabilitation treatments [17]. For example, a marker-based motion tracking system is proposed in [18] to analyze the movement of various body parts. The authors have argued that accurate detection of movement of different parts can result in better therapeutic decisions. However, the system was evaluated on a small dataset of only 10 real patients. Similarly, the authors in [19] attached an IR camera and an infrared emitter with a passive hand skateboard training device for conventional upper limb training. The proposed device was used to train eight patients with abnormal upper limb function. After four weeks of training, all the patients were able to move the hand skateboard along the designated ‘figure of eight’ path.

Capturing body movements is also critical for sports. Hence, researchers have used marker-based sensors for movement detection in walking gait [20], discus [21], dressage [22], and swimming [23] activities. Esfahani *et al.* [24] developed a trunk motion system (TMS) using printed body-worn sensors (BWS). Twelve BWSs were printed on stretchable clothing to measure the 3D trunk movements and a neural network data fusion algorithm was used to integrate the data from sensors. However, one shortcoming of these marker-based techniques is that they require the installation and calibration of multiple cameras. Hence, these systems are quite expensive. Moreover, they can only encode two-dimensional motion information.

B. VIDEO-BASED HIR SYSTEMS

In video-based HIR systems, video cameras are used to record human interactions. In such systems, the first step is to extract important features or interest points [25], [26]. Based on these distinctive features, the interaction that has been performed in the video is identified. Khan *et al.* [27] proposed a deformable part-based modeling technique to detect the body parts of a patient and track them in subsequent frames. Their system then performed movement analysis to detect various movement disorders in infants. They captured the data in a local hospital using Microsoft Kinect but it was only RGB data. Khan *et al.* [28] proposed a system for analyzing a patient’s body movements during Vojta therapy. They proposed the use of color features and pixel locations

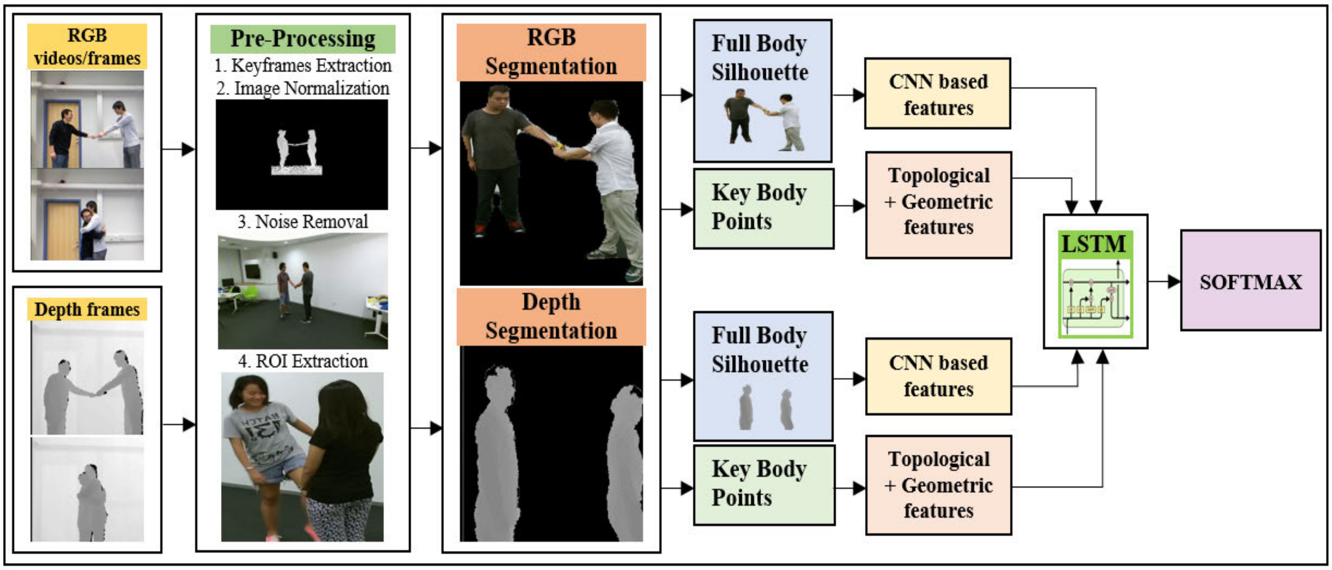


FIGURE 1. The architecture of the proposed HIR system.

for segmenting the patient's body in the images. Then they employed a multi-dimensional feature vector to classify the correct movements using multiclass SVM.

Some researchers [29], [30] also prefer extracting various features and then combining them since hybrid features have yielded better classification results in the past. For example, Jalal *et al.* [31] combined four different types of features including blobs, multiple orientations, Fourier transforms, and geometrical points. Similarly, the hybrid features introduced in [32] included energy, sine, distinct body parts movements, and 3D Cartesian views of smoothing gradients. The authors of [33] also used a hybrid of four different local descriptors: spatio-temporal features, energy-based features, shape-based angular and geometric features, and Motion-Orthogonal Histograms of Oriented Gradients (MO-HOG). However, all these approaches only employed 2D features.

III. THE PROPOSED APPROACH

The proposed system can be divided into four main sections: image preprocessing, image segmentation, feature extraction, and interaction recognition. The used methodologies and results of each section are discussed in detail below. Fig. 1 shows a flow chart of the proposed system architecture.

A. IMAGE PREPROCESSING

The RGB videos taken from the NTU RGB+D dataset have been converted into image frames at the rate of 31 frames per second. The image frames of the other two datasets were already available. Since there are multiple videos for each interaction class and each video consists of a large number of image frames, 10 keyframes have been extracted from each video to reduce complexity. The extracted frames have been normalized and noise has been removed from them. Finally, regions of interest (ROI) have been extracted from each frame. These four subsections are explained in detail below. Moreover, Algorithm 1 explains each step of the pre-processing stage.

1) KEYFRAME EXTRACTION

The number of frames varies from video to video. So, to get a fixed number of frames, 10 keyframes have been extracted from each video of every dataset. To extract the keyframes of a video, the histograms of all the image frames have been obtained. The histogram of an image x can be computed using (1).

$$P_x(i) = \frac{n_i}{N}, \quad i = 0, 1, 2 \ldots 256 \tag{1}$$

where n_i is the number of pixels with intensity i and N is the total number of pixels in the input image. Then the histograms of every two consecutive frames have been compared and their differences have been stored in a sorted array. The indices corresponding to the top ten differences have been fetched and the images at those indices are referred to as keyframes. In other words, these frames are the ones with the highest differences in their histograms.

2) IMAGE NORMALIZATION

The purpose behind image normalization is to change the pixel values of an image to a common scale so that the image appears more normal to the senses. The depth images in two out of the three datasets used are too dark to be seen by the naked eye. Moreover, features on drastically different scales can be problematic for an HIR system. In other words, features with a larger scale will dominate others and cause the system to make inaccurate assumptions. Hence, all images have been normalized. Each pixel x_i in the normalized

Algorithm 1 Preprocessing

```
Input: raw frames
Output: ROI coordinates (x,y,w,h) in preprocessed frames
            %key frame extraction%
for i in range(total frames)
    diff(i) ← hist(frame(i))-hist(frame(i + 1))
    indices ← nlargestindex(10, range(len(difference)))
    key_frame(j) ← frame(indices[j])
end
            %normalization and noise removal%
img ← key_frames(i)
norm_img ← zscore_norm(img)
denoised_img ← nonlocalmeans_denoising(norm_img)
            %ROI extraction%
diff_img ← absdiff(key_frame(i), key_frame(i + 1))
contours = FindContours(diff_img)
for contour in contours:
    (x, y, w, h) ← boundingRect(contour)
    if contourArea(contour)> min_area:
        draw_rectangle(img1, (x, y), (x + w, y + h))
        coordinates.append(x, y, w, h)
end
returncoordinates
```

image I_{norm} is normalized using (2).

$$I_{norm}(x_i) = \frac{I_{org}(x_i) - E(I_{org})}{Var(I_{org})} \tag{2}$$

where $E(I_{org})$ and $Var(I_{org})$ are the mean and variance of the original image I_{org}.

3) NOISE REMOVAL

The technique of "non-local means denoising" has been used to remove noise from the images. Local-means filters replace the value of a pixel with the mean of a group of pixels surrounding it. However, a non-local means filter takes the weighted mean of all the pixels in the image. The weight of each pixel depends on how similar it is to the target pixel. A pixel in the denoised image $u(p)$ at point p after applying non-local means denoising technique on a pixel at point q in the original image $v(q)$, is defined by (3).

$$u(p) = \frac{1}{C(p)} \int v(q) f(p, q)\, dq \tag{3}$$

where $f(p, q)$ is the weight and $C(p)$ is a normalization factor defined by (4).

$$C(p) = \int f(p, q)\, dq \tag{4}$$

4) ROI EXTRACTION

The regions of interest have been extracted from images through motion detection. Frame differencing technique has been used to detect motion in subsequent frames and rectangular boxes have been drawn over the points where motion has been detected. Since different body parts show different movements, rectangular regions of various sizes for different body parts have been obtained. Each region has a starting point x, y, width w, and height h. A minimum area condition has also been set for these rectangular regions to be considered valid. The minimum and the maximum values of x and y and the maximum values of w and h have been obtained. Finally, one rectangular region comprising all the smaller regions has been extracted as the region of interest. Its starting position is the minimum value of x and y obtained from all the smaller regions and its width is equal to the maximum value of x added to the maximum value of w. Similarly, its height is equal to the maximum value of y added to the maximum value of h.

B. IMAGE SEGMENTATION

Image segmentation is the process of segmenting the image into two parts: foreground and background. The GrabCut algorithm proposed by Rother *et al.* [34] has proven to be an efficient foreground extraction technique. It takes a rectangular region as input and assumes that all the pixels outside that region belong to the background. Then it uses a Gaussian Mixture Model (GMM) to define the area inside the rectangle by labeling each pixel as probable background and probable foreground depending upon their relation to the provided data.

Using this pixel distribution, a weighted graph is created. All pixels are treated as nodes in the graph. Then two additional nodes are added: the Source node and the Sink node. Every foreground pixel is connected to the Source node and every background pixel is connected to the Sink node. The weights of edges connecting pixels to the Source node depend on the probability of a pixel of belonging to the foreground or background. The weights between the pixels depend on pixel similarity, that is, if there is a large difference in pixel color, the edge between them will get a low weight and vice versa. Next, the graph is segmented using a Min-Cut algorithm. The graph is cut separating the Source node and the Sink node with a minimum cost function. The cost function is the sum of all weights of the edges that are cut. After cutting the graph, all the pixels connected to the Source node are labeled foreground and those connected to the Sink node are labeled background. The process continues until convergence.

However, in some cases, the extracted foreground contains portions that belong to the background. This problem came up while segmenting images from the NTU RGB+D dataset. In all those images, a major portion of the region of interest is the floor. Hence, a floor mask has been created by extracting a certain range of the intensity values from the original image and the GrabCut output is masked to get accurate results. The results of the segmentation process are shown in Fig. 2.

C. KEY BODY POINTS SELECTION

First, the full-body silhouettes have been converted into 3d meshes [35] as shown in Fig. 3. The center points of the 3d meshes have been considered the source and then geodesic distance-based heat kernel signatures (HKS) of the 3d meshes have been achieved as shown in Fig. 4. HKS, as introduced by Sun *et al.* [36], is based on a heat kernel, which is a

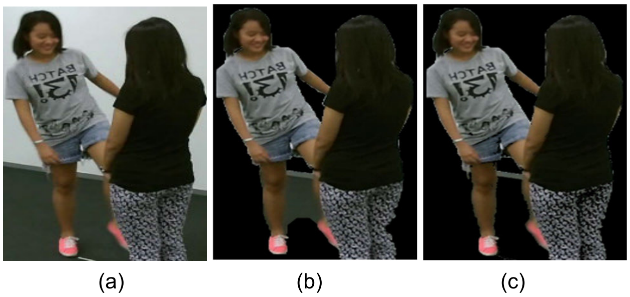


FIGURE 2. Segmentation results on RGB images showing (a) the original RGB image, (b) GrabCut output, and (c) segmented humans after applying floor mask.

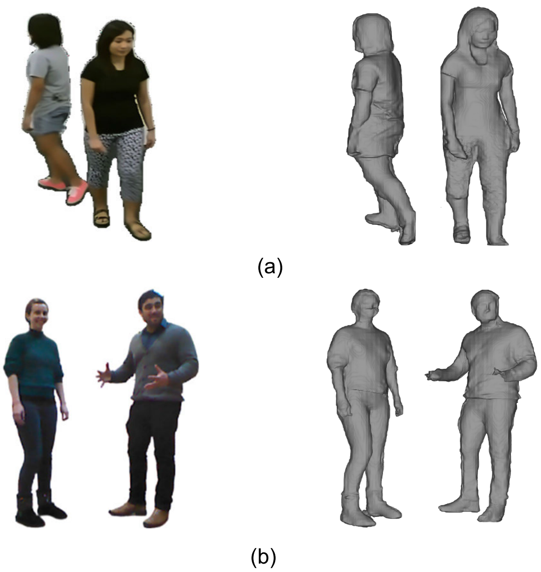


FIGURE 3. 2D images and the respective 3D meshes of both humans involved in interactions: (a) walking apart; (b) talk.

fundamental solution to the heat equation. The heat equation describes the variation of heat distribution with time. HKS is one of the many recent shape descriptors which are based on the Laplace–Beltrami operator associated with the shape [37]. The thermal diffusion process can be described by the heat equation as given in (5).

$$(\Delta - \partial/\partial t)u\,(x,t) = 0 \tag{5}$$

where Δ is the Laplace–Beltrami operator and $u\,(x,t)$ is the heat distribution at any point x at a given time t. The solution to this heat equation can be expressed in (6).

$$u\,(x,t) = \int h_t(x,y)u_o(y)dy \tag{6}$$

where $h_t(x,y)$ is called heat kernel function. The heat kernel equation is the fundamental solution to the heat equation. The Eigenvalue decomposition of the heat kernel is expressed in (7).

$$h_t u\,(x,t) = \sum\nolimits_{i=0}^{\infty} \exp(\lambda_i t)\emptyset_i(x)\emptyset_i(y) \tag{7}$$

Algorithm 2 Key Body Points Extraction

```
Input: segmented silhouettes
Output: key body points (p1,p2,p3...pn)
mesh ← Get3Dmesh(segmentedsilhoutte)
HKS ← GetHeatKernelSignature(mesh)
Clusters ← GetIntensityBasedClusters(HKS)
for cluster in Clusters:
KeyPoint ← GetClusterCentroid(Cluster)
KeyBodyPoints.append(KeyPoint)
end
returnKeyBodyPoints
```

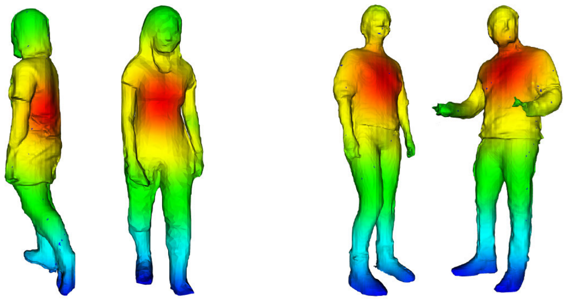


FIGURE 4. Heat kernel signatures of humans involved in interactions: (a) walking apart; (b) talk.

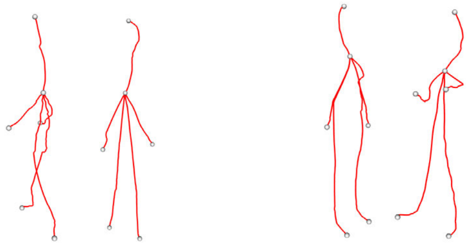


FIGURE 5. 2D leaf skeleton models using key body points of humans involved in interactions: (a) walking apart; (b) talk.

where λ_i and $\emptyset_i$ are the i^{th} eigenvalue and Eigen function of Δ. For a concise feature descriptor, HKS restricts the heat kernel only to the temporal domain.

$$h_t\,(x,x) = \exp(-\lambda_i t)\emptyset_i^2(x) \tag{8}$$

After obtaining the heat kernel signatures, all vertices in a mesh are grouped into multiple clusters based on their color or intensity value. Moreover, the centroid of each cluster is detected and is stored as a key body point. In this way, six key body points are obtained for each silhouette. When a geodesic path is drawn from the source vertex to the other five target vertices, a 2D leaf skeleton model is obtained as shown in Fig. 5. Algorithm 2 explains the process of extraction of key body points from full-body silhouettes.

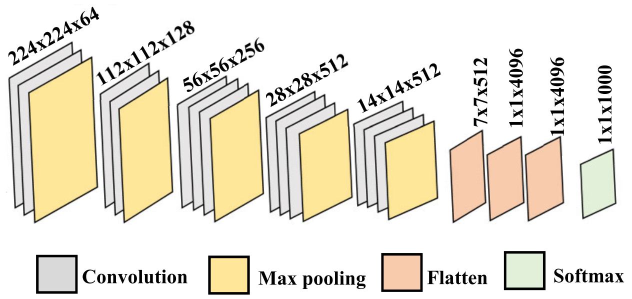


FIGURE 6. Different layers of the VGG16 architecture with configurations.

D. FEATURE EXTRACTION

This section can be divided into two phases. In the first phase, a Convolutional Neural Network (CNN) has been used to extract features from full-body silhouettes. Full-body silhouettes have been extracted from the segmented images by removing the black background from the images and making them transparent. In the second phase, topological and geometric features have been extracted using key body points. Both these phases are described in the following sub-sections.

1) FULL BODY SILHOUETTES: CNN-BASED FEATURES

For extraction of features from images, a convolutional neural network has been used. The transfer learning approach has been employed, which includes using VGG16 as the base model and then fine-tuning its weights according to the used datasets. Visual Geometry Group-16 layers deep (VGG16) [38] is a CNN model that achieved 92.7% on the ImageNet dataset which has 1000 classes. Fig. 6 shows all the layers in the VGG16 model. All images have been reshaped to 224 × 224 × 3 to match the desired input size of the VGG16 model. After training on the VGG16 base model, input images are resized to 7 × 7 × 512. These are then trained on the proposed CNN model.

There are three convolutional layers in the proposed model with 128, 64, and 32 filters respectively. The size of each filter is 3 × 3. The convolutional layers compute the output of neurons that are connected to local regions in the input. Convolution is similar to sliding a filter over an image, computing the dot product of filter weights and image pixels. Rectified Linear Unit (RELU) is used as the activation function for all three convolutional layers. It simply rounds up all the negative values to zero as shown in (9).

$$y_k = \max(0, x_k) \tag{9}$$

The convolutional layers are followed by a batch normalization layer. The pixels x_k of input images of each batch are normalized using (10).

$$\hat{x_k} = \frac{x_k - E(x_k)}{Var(x_k)} \tag{10}$$

where $E(x_k)$ is the mean and $Var(x_k)$ is the variance of pixel values.

TABLE 1. A brief summary of the cnn model.

Layer	Output Shape	Parameters
Conv:128	(None,7,7,128)	65664
Conv:64	(None,7,7,64)	8256
Con:32	(None,7,7,32)	2080
BatchNorm	(None,7,7,32)	128
Flatten	(None,7,7,1568)	0
Dropout	(None,7,7,1568)	0

The batch normalization layer is followed by a flatten layer that remaps the output of the batch normalization layer to a column vector. Lastly, a drop-out layer of 0.2 has been used to avoid overfitting. Two such models have been trained: one for RGB images and one for depth images. The two CNN models are then concatenated. Table 1 shows a summary of the CNN model.

2) KEY BODY POINTS: TOPOLOGICAL FEATURES

Topology can be defined as the spatial relationship between adjacent or neighboring features. Topological features are the properties of a geometric object that are preserved under continuous deformations. In the proposed architecture, four types of topological features have been extracted using the key body points:

1. Geodesic distance from the source.
2. Geodesic path.
3. Connected faces.
4. Nearest neighbors.

A mesh is a collection of vertices, edges, and faces that describe the shape of a 3D object. Every single point in a mesh is a vertex, a line connecting two vertices is an edge, and a flat surface enclosed by edges is called a face. In the proposed approach, the 3d meshes have been converted into graph models and these four topological features have been extracted for each key point.

First, the geodesic distance gd_i between the source vertex and each key point or target vertex has been obtained. Geodesic distance gives the distance between two vertices in a graph along the shortest path between the vertices. Hence, unlike Euclidean distance, geodesic distance considers the shape of the object while computing the distance between two points. If any two vertices are not connected in a graph, the geodesic distance between them will be infinite. After storing the value of geodesic distance, an array of all the vertices lying on this shortest path from source to target vertex has been stored as the geodesic path gp_i. For finding the connected faces, each key point or target vertex has been compared with the three vertices in each face of the mesh. In this way, the faces containing one or more of these target vertices have been found and stored as connected faces cf_i. Finally, the distance of each target vertex from all other vertices in the graph has been computed and stored in a sorted array. Then the top 128 vertices corresponding to the shortest 128 distances have been acquired. These have been stored as the nearest 128 neighbors nn_i. These features are

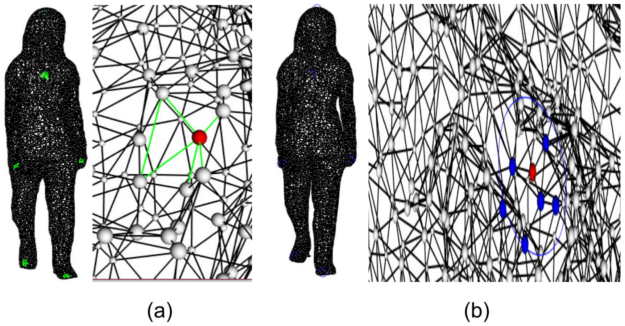


FIGURE 7. Topological features including: (a) connected edges (full mesh + zoomed in on one face) and (b) nearest neighbors (full mesh + zoomed in on vertex).

shown in Fig. 7. Hence, for each key point, a topological feature vector [gd_i, gp_i, cf_i, nn_i] has been obtained.

3) KEY BODY POINTS: GEOMETRIC FEATURES

Similar to topological features, some geometric features have also been obtained using the key body points. Ten triangular shapes have been drawn by joining different combinations of key points as shown in Fig. 8. These key points are labeled as left hand (LH), right hand (RH), left foot (LF), right foot (RF), head (H), and torso (T). Finally, the feature vector is also updated as geometric features are added to it. Algorithm 3 explains how these topological and geometric features have been extracted and concatenated in the proposed system.

Algorithm 3 Topological and Geometric Features

```
Input: key body points
Output: combined feature vectors (f1,f2,f3...fn)
                %Graph Model%
                input ← mesh
points, faces ← getpointsandcellsfrompolydata(input)
for i in range(len(points)):
actor1 ← createSphere(points[i], radius =0.003)
end
for j in range(len(faces)):
actor2 ← createLine(points[faces[j][0]], points[faces[j][1]])
actor2 ← createLine(points[faces[j][0]], points[faces[j][2]])
actor2 ← createLine(points[faces[j][1]], points[faces[j][2]])
end
                    %Feature Extraction%
for i in range(len(target)):
Distance ← GetGeodesicDistance (source,target[i])
Path ← GetGeodesicPath(source,target[i])
Connected ← GetConnectedEdges(target[i])
Neighbors ← GetNeighbors(target[i])
Geometricfeatures ← GetGeometricShape(target[i])
FeatureVector.append(Path, Distance, Connectedfaces, Neighbors,
Geometricfeatures)
end
return FeatureVector
```

E. INTERACTION RECOGNITION

At this stage of the proposed model, the interaction that has been performed in the input video has been recognized.

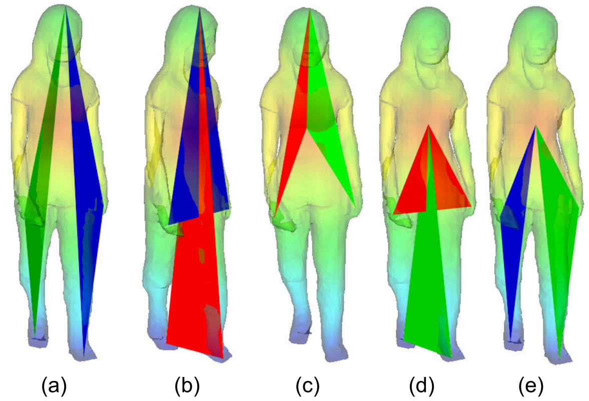


FIGURE 8. Geometric features including (a) H+LH+LF, H+RH+RF, (b) H+LH+RH, H+LF+RF, (c) H+LH+T, H+RH+T, (d) T+LH+RH, T+LF+RF, and (e) T+LH+LF, T+RH+RF.

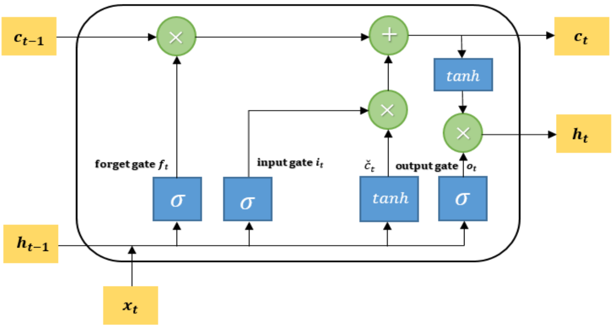


FIGURE 9. LSTM cell structure.

After concatenating the different features extracted using full-body silhouettes and key body points, the feature vector has been fed into an LSTM model which is followed by a dense layer and a Softmax classifier. Hence, this section is subdivided into two sections: LSTM and Softmax classifier.

1) LSTM

Long Short-Term Memory (LSTM) [39] is a special type of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies. The cell structure of LSTM is shown in Fig. 9. The working of LSTM has been described as follows:

1. The output value at a previous time h_{t-1} and the input value at the current time x_t are entered into the forget gate, and the output value of the forget gate f_t is obtained using (11).

$$f_t = \sigma(W_f[h_{t-1}, x_t]) \tag{11}$$

2. The output value at a previous time h_{t-1} and the input value at the current time x_t are also entered into the input gate. The output value i_t and the candidate cell state $\check{c}_t$ of the input gate are obtained using (12) and (13).

$$i_t = \sigma(W_i.[h_{t-1}, x_t]) \tag{12}$$

$$\check{c}_t = tanh(W_c.[h_{t-1}, x_t]) \tag{13}$$

TABLE 2. A brief summary of the datasets.

Dataset	# videos	# classes	Modality
SBU Kinect Interaction	231	8	RGB, depth, skeletal
NTU RGB+D	528/2880	11/60	RGB, depth, skeletal
ISR UOL 3D Social Activity	80	8	RGB, depth, skeletal

3. The current cell state c_t is updated using (14).

$$c_t = f_t * c_{t-1} + i_t * \check{c}_t \tag{14}$$

4. The output and input are received as input values at the output gate at time t, and the output of the output gate o_t is obtained using (15).

$$o_t = \sigma(W_o.[h_{t-1}, x_t]) \tag{15}$$

5. Finally, the output value of LSTM is calculated by using the output of the output gate o_t and the state of the cell c_t, as shown;

$$h_t = o_t * \tanh(c_t) \tag{16}$$

2) SOFTMAX

The Softmax classifier has been used to recognize human interactions. The Softmax function is a popular choice for multiclass classification [40]. It is an activation function that computes the probabilities of all the classes based on the output of the fully connected layer. The probabilities are between the values of 0 and 1 and the normalized sum of these probabilities is always equal to 1. It uses cross-entropy loss. The Softmax output for each class is computed using (17).

$$SM\left(z_j\right) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}} \tag{17}$$

where z is the probability of each class, i is a vector of the inputs to the output layer, j is the set of the output units, and n is the total number of classes.

IV. EXPERIMENTAL SETUP AND RESULTS

This section explains the details of the experiments conducted to validate the proposed system. All the processing and experiments have been performed using Python 3.8 with Tensorflow 2.5.0 and Keras 2.4.3. A hardware system with an Intel Core i5 processor and a 64-bit Windows-10 has been used. The system has an 8 GB and 5 (GHz) CPU. The proposed system has been tested on three different datasets and the recognition accuracies for each interaction class have been computed in the form of their confusion matrices along with precision, sensitivity, and F1 scores. For further validation, the accuracies have been compared with those of other State-Of-The-Art (SOTA) methods. This section is further divided into two sections: dataset description and experimental results.

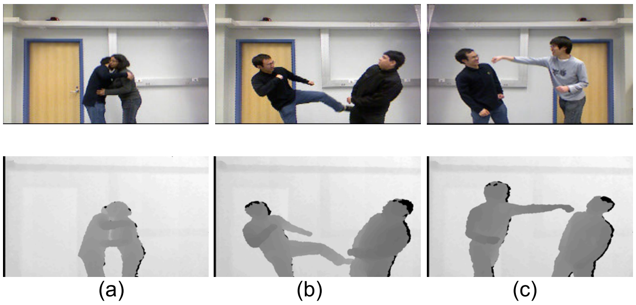


FIGURE 10. RGB and depth frames from the SBU Kinect interaction dataset. (a) hugging; (b) kicking; (c) punching.

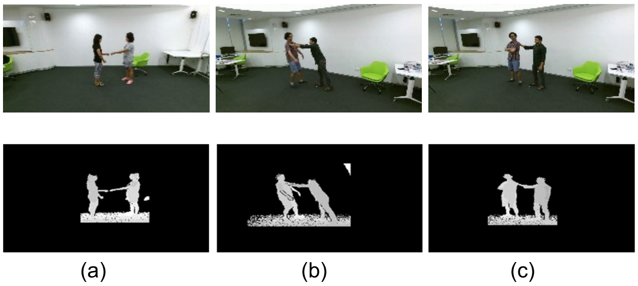


FIGURE 11. RGB and depth frames from the NTU RGB+D dataset. (a) giving object; (b) pushing; (c) pat on back.

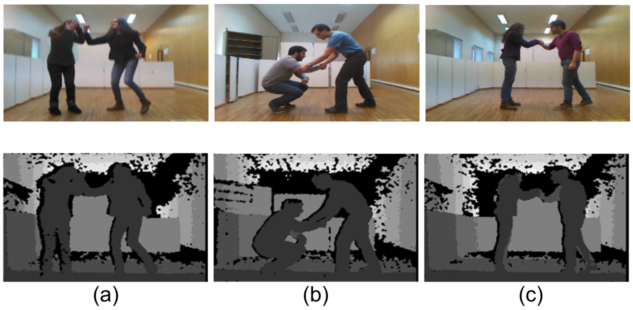


FIGURE 12. RGB and depth frames from the ISR-UOL 3D dataset. (a) fight; (b) help stand; (c) shaking hands.

A. DATASETS

The three datasets that are used for experimentation are the SBU Kinect Interaction dataset [41], the NTU RGB+D dataset [42,43], and the ISR-UoL 3D social activity dataset [44]. Details of each dataset are given in the following subsections:

1) THE SBU KINECT INTERACTION DATASET

This dataset consists of RGB, depth, and skeletal information for various interactions performed by two people. The interactions have been recorded using Microsoft Kinect sensors in an indoor environment. It consists of eight interaction classes including *approaching, departing, kicking, punching, pushing, shaking hands, exchanging an object,* and *hugging*. The dataset has a total of 21 folders with subfolders for each interaction class performed by seven different actors. For interactions in which one person is performing and the

TABLE 3. Confusion matrix of individual classes of the SBU kinect interaction dataset.

Classes	Approaching	Departing	Kicking	Pushing	SH	Hugging	EO	Punching
Approaching	**0.90**	0.06	0.00	0.00	0.02	0.02	0.00	0.00
Departing	0.05	**0.91**	0.02	0.00	0.00	0.02	0.00	0.00
Kicking	0.02	0.02	**0.92**	0.00	0.00	0.00	0.00	0.04
Pushing	0.00	0.00	0.04	**0.89**	0.00	0.00	0.03	0.04
SH	0.00	0.00	0.00	0.00	**0.92**	0.02	0.06	0.00
Hugging	0.02	0.02	0.00	0.00	0.00	**0.94**	0.02	0.00
EO	0.00	0.00	0.00	0.00	0.03	0.02	**0.92**	0.03
Punching	0.00	0.00	0.00	0.03	0.00	0.00	0.04	**0.93**
Mean recognition accuracy = 91.63%								

*SH=Shaking hands, EO=Exchanging an object.

TABLE 4. Confusion matrix of individual class of the NTU RGB+D dataset.

Classes	Kicking	Pushing	PB	PF	Hugging	GO	TP	SH	WT	WA	PG
Kicking	**0.91**	0.03	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.03	0.00
Pushing	0.02	**0.92**	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.03
PB	0.00	0.04	**0.88**	0.04	0.00	0.00	0.04	0.00	0.00	0.00	0.00
PF	0.02	0.00	0.06	**0.89**	0.00	0.00	0.03	0.00	0.00	0.00	0.00
Hugging	0.00	0.04	0.00	0.00	**0.92**	0.02	0.00	0.02	0.00	0.00	0.01
GO	0.00	0.00	0.00	0.00	0.00	**0.94**	0.04	0.02	0.00	0.00	0.00
TP	0.00	0.00	0.04	0.02	0.00	0.04	**0.90**	0.00	0.00	0.00	0.00
SH	0.00	0.00	0.00	0.02	0.02	0.03	0.00	**0.93**	0.00	0.00	0.00
WT	0.04	0.00	0.00	0.00	0.00	0.02	0.00	0.03	**0.90**	0.00	0.00
WA	0.04	0.00	0.00	0.00	0.02	0.03	0.00	0.03	0.00	**0.88**	0.05
PG	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.03	**0.89**
Mean recognition accuracy = 90.54%											

*PB=Pat on back, PF=Point finger, GO=Giving Object, TP=Touch Pocket, SH=Shaking hands, WT=Walking towards, WA=Walking apart, PG=Punching.

other is receiving the action, there are two videos. The person performing the action in one video is receiving the action in the second video and vice versa. Videos have been segmented at the rate of 15 frames per second (fps). The sizes of both RGB and depth images are 649 × 480.

2) THE NTU RGB+D DATASET

This dataset provides RGB, depth, and skeletal information. It consists of 60 classes, 11 of which are two-person interactions including *punching, kicking, pushing, pat on back, point finger, hugging, giving object, touch pocket, shaking hands, walking towards,* and *walking apart*. There are 48 videos for each interaction class. Each session has three sets of videos since each video has been recorded from three different viewpoints.

3) THE ISR-UoL 3D SOCIAL ACTIVITY DATASET

This dataset also consists of RGB, depth, and skeletal information recorded using Kinect 2 sensor. In this dataset, some interactions are everyday interactions while others are assisted living interactions. There are a total of eight interactions including *shaking hands*, *hugging, help walk, help stand-up, fight, push, talk,* and *draw attention*. The actions are performed by four males and two females. There are ten sessions and each session contains all eight interactions. For each interaction, 24-bit RGB images, 8-bit and 16-bit resolution depth images, and the skeletal information of 15 joints are available. Each interaction is repeated over a period of 40–60 repetitions in one video.

B. EXPERIMENTS AND RESULTS

For validating the performance of the proposed system, different metrics have been used. The experimentation phase has been divided into two categories: classification accuracy of each class in terms of confusion matrix, precision, sensitivity, and F1 score, and comparison of the proposed system with other state-of-the-art methods. The results for each stage are given in the following sub-sections.

1) INDIVIDUAL CLASS ACCURACY

The results of the proposed model's performance are given in the form of confusion matrices showing true positives, true

TABLE 5. Confusion matrix of individual class over the ISR-UoL 3D social activity dataset.

Classes	SH	Hugging	Help walk	HS	Fight	Push	Talk	DA
SH	**0.90**	0.02	0.03	0.03	0.02	0.00	0.00	0.00
Hugging	0.05	**0.91**	0.00	0.04	0.00	0.00	0.00	0.00
Help walk	0.05	0.04	**0.89**	0.02	0.00	0.00	0.00	0.00
HS	0.03	0.05	0.04	**0.88**	0.00	0.00	0.00	0.02
Fight	0.01	0.00	0.00	0.00	**0.92**	0.04	0.03	0.00
Push	0.00	0.00	0.00	0.03	0.04	**0.90**	0.03	0.00
Talk	0.00	0.00	0.03	0.00	0.03	0.03	**0.90**	0.02
DA	0.03	0.02	0.00	0.00	0.00	0.00	0.04	**0.91**
Mean recognition accuracy = 90.13%								

*SH=Shaking hands, HS=Help stand up, DA=Draw attention.

TABLE 6. Measurements of precision, sensitivity, and F1 scores over the SBU kinect interaction dataset.

Class	Precision	Sensitivity	F1 score
Approaching	0.88	0.90	0.90
Departing	0.90	0.91	0.90
Kicking	0.91	0.92	0.91
Pushing	0.89	0.89	0.90
SH	0.90	0.92	0.93
Hugging	0.93	0.94	0.93
EO	0.91	0.92	0.91
Punching	0.91	0.93	0.92
Mean	0.90	0.91	0.91

*SH=Shaking hands, EO=Exchanging an object.

TABLE 7. Measurements of precision, sensitivity, and F1 scores over the ISR-UoL 3D social activity dataset.

Class	Precision	Sensitivity	F1 score
SH	0.92	0.90	0.90
Hugging	0.92	0.91	0.91
Help walk	0.89	0.89	0.88
HS	0.89	0.88	0.89
Fight	0.92	0.92	0.91
Push	0.90	0.90	0.88
Talk	0.91	0.90	0.91
DA	0.92	0.91	0.92
Mean	0.91	0.90	0.90

*SH=Shaking hands, HS=Help stand up, DA=Draw attention.

negatives, false positives, and false negatives for each class individually. The confusion matrices for SBU Kinect Interaction, NTU RGB+D, and ISR-UoL 3D social activity datasets are given in Tables 3, 4, and 5 respectively. It can be observed, from the above-mentioned tables that the interaction classes of all three datasets achieved higher recognition rates with the mean accuracy rates of 91.63%, 90.54%, and 90.13% with the SBU Kinect Interaction, NTU RGB+D, and ISR-UoL 3D social activity datasets respectively.

However, there is still some confusion between interaction classes that involve similar actions such as the *departing* and *approaching* interactions in the sports dataset. Similarly, shaking hands and exchanging an object interactions of the SBU Kinect Interaction dataset are confused with each other as shown in Table 3. Table 4 shows that the *pat on back* and *point finger* interactions of the NTU RGB+D datasets are confused with each other. As seen in Table 5, there is confusion between the *hugging* and *shaking hands* interaction of the ISR-UOL 3D social activity dataset.

Tables 6, 7, and 8 show the precision, sensitivity, and F1 scores of each class in SBU Kinect Interaction, ISR-UoL 3D social activity, and NTU RGB+D datasets respectively. The precision, sensitivity, and F1 scores of all the interaction classes for each dataset have been calculated as;

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{19}$$

$$F1score = \frac{2(Precision \times Sensitivity)}{(Precision + Sensitivity)} \tag{20}$$

where TP, FP, and FN stand for True Positives, False Positives and False Negatives respectively.

2) COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, the proposed method is compared with different methodologies adopted by researchers for HIR recognition from recent years. The action recognition accuracies of each evaluated methodology are used for comparison with the proposed system. Tables 9, 10, and 11 give the comparison of the proposed system with other state-of-the-art (SOTA) systems evaluated on SBU Kinect Interaction, NTU RGB+D, and ISR-UoL 3D social activity datasets respectively. The results show that the proposed hybrid descriptors are more

TABLE 8. Measurements of precision, sensitivity, and F1 score over the NTU RGB+D dataset.

Class	Precision	Sensitivity	F1 score
Kicking	0.91	0.91	0.90
Pushing	0.91	0.92	0.91
PB	0.89	0.88	0.88
PF	0.91	0.89	0.91
Hugging	0.91	0.92	0.91
GO	0.93	0.94	0.94
TP	0.89	0.90	0.89
SH	0.95	0.93	0.93
WT	0.90	0.90	0.91
WA	0.87	0.88	0.89
PG	0.89	0.89	0.90
Mean	0.90	0.90	0.91

*PB=Pat on back, PF=Point finger, GO=Giving Object, TP=Touch Pocket, SH=Shaking hands, WT=Walking towards, WA=Walking apart, PG=Punching.

TABLE 9. Comparison with other SOTA methods over the SBU dataset.

Methods	Accuracy (%)
Joint features [41]	80.30
Body parts contrast mining [45]	86.9
Joint Features [46]	90.3
Deep LSTM [47]	90.41
STA-LSTM [48]	91.51
Hybrid descriptors (Proposed Method)	**91.63** (RGB+D) **89.53** (Depth only) **88.24** (RGB only)

TABLE 10. Comparison with other SOTA methods over the NTU RGB+D dataset.

Methods	Accuracy (%)
geometric features [49]	70.26
ensemble TS-LSTM v2 [50]	74.60
STA-LSTM [48]	81.2
multitask deep learning [51]	85.5
pair wise features [52]	88.6
Hybrid descriptors (Proposed Method)	**90.54** (RGB+D) **88.12** (Depth only) **87.63** (RGB only)

robust than the different types of features used in the recent SOTA systems.

V. DISCUSSION

The proposed system is a complete HIR solution that should be applicable to many real-world problems involving the tasks of human behavior monitoring, security, surveillance, and managing smart homes. It is designed for RGB+D datasets but can also be used with RGB only or depth only datasets using only one stream of the proposed CNN model and skipping the model concatenation stage.

Each step from the preprocessing stages to the classification stage contributes to the improved performance achieved by the system. The proposed feature extraction method successfully extracts robust features, which in turn, play a critical role in accurate classification of the interactions. Using two CNN models for training RGB and depth images separately and then concatenating the models gives better results than those obtained by concatenating both RGB and Depth images first and then training the 4-dimensional images using only one CNN model. Moreover, since all three datasets contain video sequences, the LSTM-based classification step gives accurate results.

Despite yielding good results, the proposed system is not without limitations. The proposed 2D leaf skeleton model for the detection of key body points can only extract six key points so far. However, better accuracies can be achieved if more key points are identified and their features are extracted. Moreover, the proposed system is very extensive and computationally expensive. The time complexity of the proposed system is shown in Table 12.

TABLE 11. Comparison with other SOTA methods over the ISR-UoL 3D social activity dataset.

Methods	Accuracy (%)
probabilistic merging of skeletal features [44]	85.1
multimodal feature level fusion [53]	85.12
statistical and geometrical features [54]	85.56
skeletal data [55]	87
Hybrid descriptors (Proposed Method)	**90.13** (RGB+D only) **88.14** (Depth only) **86.83** (RGB only)

TABLE 12. Time complexity of the proposed system.

Dataset	Execution time (s)
SBU Kinect Interaction	4795.71
NTU RGB+D	6560.05
ISR UOL 3D Social activity	3671.32

VI. CONCLUSION

In this paper, a novel HIR framework had been proposed that uses both machine learning and deep learning techniques for feature extraction from 2D human silhouettes and 3D meshes. Using efficiently segmented silhouettes of the humans from images, multiple features using full-body silhouettes and key body points from their corresponding 3D meshes have been extracted. The features have then been fed to an LSTM and a Softmax-based classifier. The proposed system achieved average accuracies of 91.63%, 90.54%, and 90.13% with the SBU Kinect Interaction, NTU RGB+D, and ISR-UoL 3D social activity datasets respectively.

In the future, the authors plan to shift their focus to the task of human-object interaction recognition and investigate new features and modeling techniques for better classification results.

REFERENCES

[1] O. Aran and D. Gatica-Perez, "One of a kind: Inferring personality impressions in meetings," in *Proc. ICMI*, 2013, pp. 11–18.

[2] A. Jalal, N. Sarif, J. T. Kim, and T.-S. Kim, "Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home," *Indoor Built Environ.*, vol. 22, no. 1, pp. 271–279, Feb. 2013.

[3] S. U. Khan, T. Hussain, A. Ullah, and S. W. Baik, "Deep-ReID: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance," *Multimedia Tools Appl.*, vol. 184, pp. 1–22, Jan. 2021.

[4] G.-H. Liu, J.-Y. Yang, and Z. Y. Li, "Content-based image retrieval using computational visual attention model," *Pattern Recognit.*, vol. 48, no. 8, pp. 2554–2566, 2015.

[5] S. Sempena, N. U. Maulidevi, and P. R. Aryan, "Human action recognition using dynamic time warping," in *Proc. Int. Conf. Electr. Eng. Informat. (ICEEI)*, Jul. 2011, pp. 1–5.

[6] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," *Appl. Sci.*, vol. 9, no. 22, p. 4963, Nov. 2019.

[7] A. Jalal, M. Batool, and K. Kim, "Stochastic recognition of physical activity and healthcare using tri-axial inertial wearable sensors," *Appl. Sci.*, vol. 10, no. 20, p. 7122, Oct. 2020.

[8] A. Jalal, M. A. K. Quaid, S. B. U. D. Tahir, and K. Kim, "A study of accelerometer and gyroscope measurements in physical life-log activities detection systems," *Sensors*, vol. 20, no. 22, p. 6670, Nov. 2020.

[9] A. Jalal, M. Batool, and K. Kim, "Sustainable wearable system: Human behavior modeling for life-logging activities using K-ary tree hashing classifier," *Sustainability*, vol. 12, no. 24, p. 10324, Dec. 2020.

[10] M. Javeed, A. Jalal, and K. Kim, "Wearable sensors based exertion recognition using statistical features and random forest for physical healthcare monitoring," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 512–517.

[11] S. U. Khan and R. Baik, "MPPIF-Net: Identification of plasmodium falciparum parasite mitochondrial proteins using deep features with multilayer bi-directional LSTM," *Processes*, vol. 8, no. 6, p. 725, Jun. 2020.

[12] A. Shehzad, A. Jalal, and K. Kim, "Multi-person tracking in smart surveillance system for crowd counting and normal/abnormal events detection," in *Proc. Int. Conf. Appl. Eng. Math. (ICAEM)*, 2019, pp. 163–168.

[13] M. Pervaiz, A. Jalal, and K. Kim, "Hybrid algorithm for multi people counting and tracking for smart surveillance," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 530–535.

[14] 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, p. 970, Jan. 2021.

[15] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. P. Lopez, X. Baro, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proc. Automat. Face Gesture Recognit.*, 2017, pp. 476–483.

[16] A. Jalal, J. T. Kim, and T.-S. Kim, "Human activity recognition using the labeled depth body parts information of depth silhouettes," in *Proc. Sustain. Healthy Buildings*, 2012, pp. 1–8.

[17] D. Rado, A. Sankaran, J. Plasek, D. Nuckley, and D. F. Keefe, "A real-time physical therapy visualization strategy to improve unsupervised patient rehabilitation," in *Proc. IEEE Vis.*, Oct. 2009, pp. 1–2.

[18] M. H. Khan, M. Zöller, M. S. Farid, and M. Grzegorzek, "Marker-based movement analysis of human body parts in therapeutic procedure," *Sensors*, vol. 20, no. 11, p. 3312, Jun. 2020.

[19] C.-C. Chen, C.-Y. Liu, S.-H. Ciou, S.-C. Chen, and Y.-L. Chen, "Digitized hand skateboard based on IR-camera for upper limb rehabilitation," *J. Med. Syst.*, vol. 41, no. 2, p. 36, Feb. 2017.

[20] R. E. Mayagoitia, A. V. Nene, and P. H. Veltink, "Accelerometer and rate gyroscope measurement of kinematics: An inexpensive alternative to optical motion analysis systems," *J. Biomech.*, vol. 35, no. 4, pp. 537–542, 2002.

[21] N. Ganter, A. Krüger, M. Gohla, K. Witte, and J. Edelmann-Nusser, "Applicability of a full body inertial measurement system for kinematic analysis of the discus throw," in *Proc. Int. Soc. Biomech. Sports*, 2010, pp. 1–4.

[22] F. Eckardt, A. Münz, and K. Witte, "Application of a full body inertial measurement system in dressage riding," *J. Equine Vet. Sci.*, vol. 34, nos. 11–12, pp. 1294–1299, 2014.

[23] F. A. de Magalhaes, G. Vannozzi, G. Gatta, and S. Fantozzi, "Wearable inertial sensors in swimming motion analysis: A systematic review," *J. Sports Sci.*, vol. 33, no. 7, pp. 732–745, 2015.

[24] M. M. Esfahani, O. Zobeiri, B. Moshiri, R. Narimani, M. Mehravar, E. Rashedi, and M. Parnianpour, "Trunk motion system (TMS) using printed body Worn sensor (BWS) via data fusion approach," *Sensors*, vol. 17, no. 12, p. 112, Jan. 2017.

[25] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 3, pp. 313–323, May 2012.

[26] S. A. Rizwan, A. Jalal, M. Gochoo, and K. Kim, "Robust active shape model via hierarchical feature extraction with SFS-optimized convolution neural network for invariant human age classification," *Electronics*, vol. 10, no. 4, p. 465, Feb. 2021.

[27] M. Khan, M. Schneider, M. Farid, and M. Grzegorzek, "Detection of infantile movement disorders in video data using deformable part-based model," *Sensors*, vol. 18, no. 10, p. 3202, Sep. 2018.

[28] M. H. Khan, J. Helsper, M. S. Farid, and M. Grzegorzek, "A computer vision-based system for monitoring vojta therapy," *Int. J. Med. Informat.*, vol. 113, pp. 85–95, May 2018.

[29] M. Javeed, M. Gochoo, A. Jalal, and K. Kim, "HF-SPHR: Hybrid features for sustainable physical healthcare pattern recognition using deep belief networks," *Sustainability*, vol. 13, no. 4, p. 1699, Feb. 2021.

[30] M. Gochoo, I. Akhter, A. Jalal, and K. Kim, "Stochastic remote sensing event classification over adaptive posture estimation via multifused data and deep belief network," *Remote Sens.*, vol. 13, no. 5, p. 912, Feb. 2021.

[31] A. Jalal, A. Ahmed, A. A. Rafique, and K. Kim, "Scene semantic recognition based on modified fuzzy C-mean and maximum entropy using object-to-object relations," *IEEE Access*, vol. 9, pp. 27758–27772, 2021.

[32] A. Jalal, I. Akhtar, and K. Kim, "Human posture estimation and sustainable events classification via pseudo-2D stick model and K-ary tree hashing," *Sustainability*, vol. 12, no. 23, p. 9814, Nov. 2020.

[33] 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, p. 817, Jul. 2020.

[34] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[35] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. ICCV*, 2019, pp. 2304–2314.

[36] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, Jul. 2009.

[37] R. Litman and A. M. Bronstein, "Learning spectral descriptors for deformable shape correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 171–180, Jan. 2014.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] K. Banerjee, V. Prasad, R. Gupta, K. Vyas, H. Anushree, and B. Mishra, "Exploring alternatives to softmax function," *CoRR*, vol. abs/2011.11538, pp. 1–8, Nov. 2020.

[41] 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. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 28–35.

[42] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[43] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, and A. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[44] 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. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 5055–5061.

[45] Y. Ji, G. Ye, and H. Cheng, "Interactive body part contrast mining for human interaction recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.

[46] T. Huynh-The, O. Banos, B.-V. Le, D.-M. Bui, S. Lee, Y. Yoon, and T. Le-Tien, "PAM-based flexible generative topic model for 3D interactive activity recognition," in *Proc. Int. Conf. Adv. Technol. Commun. (ATC)*, Oct. 2015, pp. 117–122.

[47] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. Artif. Intell.*, 2016, pp. 3697–3703.

[48] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. Artif. Intell.*, 2016, pp. 1–8.

[49] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 148–157.

[50] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1012–1020.

[51] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5137–5146.

[52] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 601–604.

[53] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019.

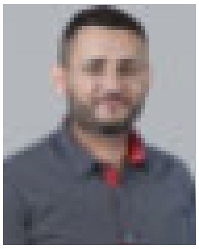
[54] C. Coppola, S. Cosar, D. R. Faria, and N. Bellotto, "Automatic detection of human interactions from RGB-D data for social activity classification," in *Proc. 26th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 871–876.

[55] A. Manzi, L. Fiorini, R. Limosani, P. Dario, and F. Cavallo, "Two-person activity recognition using skeleton data," *IET Comput. Vis.*, vol. 12, no. 1, pp. 27–35, Feb. 2018.

MANAHIL WAHEED received the B.S. degree in electronics engineering from the International Islamic University, in 2018. She is currently pursuing the M.S. degree in data science with Air University, Islamabad. Her research interests include digital image processing, data science, and artificial intelligence.

AHMAD JALAL received the Ph.D. degree from the Department of Biomedical Engineering, Kyung Hee University, Republic of Korea. He was working as a Postdoctoral Research Fellow at POSTECH. He is currently an Associate Professor with the Department of Computer Science and Engineering, Air University, Pakistan. His research interests include multimedia contents and artificial intelligence.

MOHAMMED ALARFAJ received the B.S., M.Eng., and Ph.D. degrees in electrical and computer engineering from Oregon State University, in 2011, 2014, and 2019, respectively. He is currently an Assistant Professor of electrical engineering at King Faisal University. He is the Head of the Electrical Engineering Department, King Faisal University. His current research interests include MIMO, mmWave and wireless communications, signal processing, and applications in wireless communication and sensor networks.



YAZEED YASIN GHADI received the Ph.D. degree in electrical and computer engineering from Queensland University. He is currently an Assistant Professor of software engineering at Al Ain University. He was a Postdoctoral Researcher at Queensland University before joining Al Ain. He has published more than 25 peer-reviewed journals and conference papers. He holds three pending patents. His current research interests include developing novel electro-acoustic-optic neural interfaces for large-scale high-resolution electrophysiology and distributed optogenetic stimulation. He was a recipient of several awards. His dissertation on developing novel hybrid plasmonic photonic on-chip biochemical sensors received the Sigma Xi Best Ph.D. Thesis Award.

TAMARA AL SHLOUL is currently an Assistant Professor (humanities) at Al Ain University. She has vast experience of teaching education and humanities courses, along with experience in school supervision, thinking skills, and higher education improvement ability. Her research interests include teacher socialization and professional development.

SHAHARYAR KAMAL received the M.S. degree in computer engineering from Mid Sweden University, Sweden, and the Ph.D. degree from the Department of Radio and Electronics Engineering, Kyung Hee University, Republic of Korea. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Air University, Pakistan. His research interests include advanced wireless communication and image and signal processing.

DONG-SEONG KIM (Senior Member, IEEE) received the Ph.D. degree in EECS from Seoul National University, Seoul, South Korea. From 1994 to 1998, he worked as a full-time Researcher at ERC-ACI, Seoul National University. From September 2000 to December 2001, he worked as a part-time Lecturer with the Department of Information and Communication, Dongguk University. He is currently a Professor with the Department of IT Convergence Engineering, School of Electronic Engineering, Kumoh National Institute of Technology, Gumi, South Korea. He has other responsibilities as the Director and the Head of the Convergence Technology Institute, ICT-CRC, South Korea. His main research interests include industrial networked control systems, fieldbus and real-time systems, and wired/wireless military networks.