

Article

# Automatic Recognition of Human Interaction via Hybrid Descriptors and Maximum Entropy Markov Model Using Depth Sensors

Ahmad Jalal <sup>1</sup>, Nida Khalid <sup>1</sup> and Kibum Kim <sup>2,\*</sup> 

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

<sup>2</sup> Department of Human-Computer Interaction, Hanyang University, Ansan 15588, Korea

\* Correspondence: kikum@hanyang.ac.kr

Received: 27 June 2020; Accepted: 24 July 2020; Published: 26 July 2020



**Abstract:** Automatic identification of human interaction is a challenging task especially in dynamic environments with cluttered backgrounds from video sequences. Advancements in computer vision sensor technologies provide powerful effects in human interaction recognition (HIR) during routine daily life. In this paper, we propose a novel features extraction method which incorporates robust entropy optimization and an efficient Maximum Entropy Markov Model (MEMM) for HIR via multiple vision sensors. The main objectives of proposed methodology are: (1) to propose a hybrid of four novel features—i.e., spatio-temporal features, energy-based features, shape based angular and geometric features—and a motion-orthogonal histogram of oriented gradient (MO-HOG); (2) to encode hybrid feature descriptors using a codebook, a Gaussian mixture model (GMM) and fisher encoding; (3) to optimize the encoded feature using a cross entropy optimization function; (4) to apply a MEMM classification algorithm to examine empirical expectations and highest entropy, which measure pattern variances to achieve outperformed HIR accuracy results. Our system is tested over three well-known datasets: SBU Kinect interaction; UoL 3D social activity; UT-interaction datasets. Through wide experimentations, the proposed features extraction algorithm, along with cross entropy optimization, has achieved the average accuracy rate of 91.25% with SBU, 90.4% with UoL and 87.4% with UT-Interaction datasets. The proposed HIR system will be applicable to a wide variety of man–machine interfaces, such as public-place surveillance, future medical applications, virtual reality, fitness exercises and 3D interactive gaming.

**Keywords:** cross entropy; depth sensors; Gaussian mixture model; maximum entropy Markov model

## 1. Introduction

Human interaction recognition (HIR) deals with the understanding of communication taking place between a human and an object or other persons [1]. HIR includes an understanding of various actions, such as social interaction, person to person talking, meeting or greeting in the form of a handshake or a hug and the performance of inappropriate actions, such as fighting, kicking or punching each other. There are many different kinds of interactions that can easily be identified by human observations. However, in many situations, personal human observation of some actions is impractical due to the cost of resources and also to hazardous environments. For example, in the case of smart rehabilitation, it is more suitable for a machine to monitor a patient's daily routine rather than for a human to constantly observe a patient (24/7) [2]. Similarly, in the case of video surveillance, it is more appropriate to monitor human actions via sensor devices, especially in places where risk factors and suspicious activities are involved.

Due to a wide variety of applications, HIR has gained much attention in recent years. These applications include public security surveillance, e-health care, smart homes, assistive robots and sports assistance [3] each of which requires an efficient understanding and identification of discrete human movements [4–9]. Many HIR systems have been proposed to tackle problems faced in activity monitoring in healthcare, rehabilitation, surveillance and many other situations. Reliable and accurate monitoring is essential in order to monitor the progress of patients in physical therapy and rehabilitation centers, to detect potential and actual dangers, such as falls and thefts, and to prevent mishaps and losses due to lack of attention [10]. HIR systems are also proposed for security reasons [11], such as a Fuzzy logic based human activity recognition system, which was proposed in [12]. A Hidden Markov Model [HMM] based HIR system was proposed for surveillance [13]. A random forest based HIR system for smart home and elderly care was proposed by H. Xu et al. [14]. A neural network-based HIR system was presented by S. Chernbumroong for assisted living [15]. Clearly, HIR systems are in demand and highly applicable in many daily life domains.

Motivated by the applications of HIR systems in daily life, we proposed a robust system which is able to track human interactions and which is easy to deploy in real world applications [16]. Challenges, such as complex and cluttered background, intra-class variations and interclass similarity make it difficult to accurately recognize and distinguish between human interactions. Therefore, we aim to increase the recognition rate of human–human interactions and tackle challenges faced by recent HIR systems by incorporating depth sensors. The recognition rate of human interactions is being boosted with a recent low-cost depth sensors technology [17,18]. Depth imaging technology is getting more attention in recent years because it is providing promising results without the attachment of marker sensors [2,19,20]. HIR systems based on depth sensors are easy to deploy in daily life applications compared to systems based on wearable or marker sensors [21]. This is because wearable sensors need to be attached to the body of an individual in order to give better performance, but this creates usability and mobility problems for the wearer. The main purpose of this research work is to propose a multi-vision sensor based HIR system which consists of a hybrid of four unique features in order to achieve a better performance rate. Our system aims at giving computers sensitivity to automatically monitor, recognize and distinguish between human actions happening in respective surroundings.

Basically, HIR can be categorized into four types—i.e., human–object interaction (HOI), human–robot interaction (HRI), human–human interaction (HHI) and human–group of humans interaction (HGI). In the case of HOI, humans act, communicate and interact with various objects to perform different daily actions [22,23] such as picking-up a glass for drinking, holding a ball for throwing and taking food for eating. During HRI, a robot may be able to perform different postures, such as hand shaking, serving food and waving hands, etc. Robots in HRI can precisely predict a human’s future actions and analyze the gestures of the persons that interact with them [24]. Similarly, in HHI and HGI, a system can estimate the trajectory information of the human–human or capture the movement patterns of groups of people [25] in crowded or public areas. However, our research work is focused on human to human interaction.

In this paper, we propose a novel hybrid HIR system and entropy Markov model that examines the daily interactions of humans. The proposed model measures the spatiotemporal properties of the body’s posture and estimates an empirical expectation of pattern changes using depth sensors. For the vision (RGB or depth) filtered data, we have adopted mean filter, pixel connectivity analysis and Otsu’s thresholding method. For hybrid descriptor features, we proposed four types of features characteristics as follows:

- Space and time based—i.e., spatio-temporal features—in which displacement measurements between key human body points are recognized as temporal features. Intensity changes along the curved body points of silhouettes are taken as spatial features.
- Motion-orthogonal histograms of oriented gradient (MO-HOG) features are based on three different views of human silhouette. These views are projected in the form of orthogonal shape and then HOG is applied.

- Shape based angular and geometric features include angular measurements over two types of shapes—i.e., inter-silhouettes and intra-silhouette.
- Energy based features examine distinct body parts energy distribution within a silhouette.

These hybrid descriptors are fed into a Gaussian mixture model (GMM) and into fisher encoding for codebook generation and for proper discrimination among various activity classes. Then, we applied cross entropy algorithm which resulted in the optimized distribution of matrixes. Finally, the maximum entropy Markov model (MEMM) is embodied in the proposed HIR system to estimate empirical expectation and the highest entropy of different human interactions to achieve significant accuracy. Four experiments were performed using a leave-one-out cross validation method on three well-known datasets. Our proposed method acquired significant performance compared to well-known statistical state-of-the-art methods. The major contributions of this paper can be highlighted as follows:

1. We proposed to apply hybrid descriptor features of spatiotemporal characteristics, invariant properties, view-orientation as well as displacement and intra/inter angular values to distinct human interactions.
2. We introduced a combination of GMM with fisher encoding for codebook generation and optimal discrimination of features.
3. We designed cross entropy optimization and MEMM to analyze contextual information as well as to classify complex human interactions in a better way.
4. We performed experiments using three publicly available datasets and the proposed method was fully validated for the efficacy, outperforming other state-of-the-art methods, including deep learning.

The rest of the paper is organized as follows: Section 2 consists of related work in the field of HIR. Section 3 presents details of our proposed methodology. Section 4 reports the experimentation, dataset description results generation. Section 5 presents a discussion of the overall paper. Finally, Section 6 concludes the proposed research work with some future directions.

## 2. Related Work

Recently, a lot of works have been done by researchers for the development of HIR using multiple types of sensors. On the basis of methods used to capture human interactions, we categorize these sensors in our related work into three major types: (1) wearable sensor-based HIR; (2) vision sensor-based HIR and (3) Marker sensor-based HIR.

### 2.1. Wearable Sensor-Based HIR Systems

In wearable sensor-based technology, many sensors (e.g., accelerometers, gyroscopes and magnetometers) are attached to the subject's limb and body in order to examine interactions with the surroundings [26–28]. In [29], A. Howedi et al. proposed a unique HIR methodology based on different entropy measures, such as Fuzzy, sample and approximate entropy. They achieved significant accuracy in entropy measurements for the detection and identification during human interactions. In [30], M. Ehatisham et al. designed an action recognition system based on K-nearest neighbors and SVM via multiple sensors, including RGB cameras, depth sensors and wearable sensors for accurate recognition of human behaviors. H. Xu et al. [31] developed a wearable sensor based HIR that extracted various feature values via Hilbert–Huang transform (HHT). HHT spectrum features include frequency, amplitude, means and energy values that are tested over PAMAP2 wearable sensor datasets. Experimental results showed that multi features approach achieved better performance for HIR.

Motivated by the application of wearable sensors in health departments, a human motion detection system based on accelerometer sensor measurements is proposed by A. Jalal et al. [32]. In order to extract features of each activity class axial components of the accelerometer are taken. After extracting features, Random forest is applied to classify interactions that result in good performance in human

motion detection. In order to recognize the physical activities of humans, wearable sensors are used by M. Batool et al. in [33]. They used both the gyroscope and accelerometer sensor data.

They extracted statistical and Mel-frequency cepstral coefficient data. A combination of particle swarm optimization (PSO) and support vector machine (SVM) resulted in a better recognition rate. In order to solve the problem of feature selection and classification of sensor data, a genetic algorithm-based approach is used by M.A. Quaid et al. [34]. Statistical and acoustic features are extracted and then features are reweighted. After reweighting the features, biological operations for crossover and mutation are applied. One self-annotated dataset is proposed in this work. Experiments on three-mark datasets proved the efficiency of proposed human behavior analysis system. Motivated by applications using wearable sensors for elderly care, S. Badar et al. proposed a wearable sensor-based activity monitoring system [35]. This system consists of inertial and motion node sensors. Three types of features, such as binary, wavelet and statistical are extracted. In order to optimize features, adaptive moment estimation (Adam) and Ada delta are applied. Experiments on two datasets are used for system evaluation. The results showed a better performance compared to other state of the art systems.

In order to recognize daily activity, a smartphone with built in accelerometer was used by A.M. Khan et al. [36]. Two types of features, such as autoregressive coefficients and signal magnitude area were extracted. Kernel discriminant analysis and Artificial Neural Network (ANN) were then used for accurately identifying the activity class. Inspired by the applications of sensors embedded in smartphones, N.A. Capela et al. [37] proposed a human activity recognition system. In this research work, sensor data were taken from patients and elderly people. Seventy-six signal features were extracted and then selected on the basis of feature selection methods. Three classifiers were used to evaluate the proposed methodology and results reveal a better rate of accuracy. Motivated by healthcare and rehabilitation-based applications for human activity recognition systems, W. Jiang et al. proposed a wearable sensor-based method [38]. They collect signals from sensors in the form of activity images. They applied deep CNN for feature extraction. Evaluation on three benchmark datasets validated the performance of their system. However, these technologies face several limitations in HIR, such as discomfort and restricted movement for subjects, due to many wires and wearable sensors that are attached to their bodies [39]. Similarly, in order to capture full-body movements, the multiple sensors that are attached to the human body, cause computational complexity in the system. Background noises picked up by wearable sensors during measurements are also incorporated in the data and these result in numerous false predictions which affect decision making [40]. Therefore, instead of relying on wearable devices, vision-based sensor technologies have started gaining global attentions as a solution in HIR studies.

## 2.2. Vision-Based HIR Systems

In vision-based HIR systems, video cameras are mounted for automated inspection of human interactions in various public areas (i.e., shopping malls, parks and roads). In [41], M. Sharif et al. proposed a human activity monitoring system. They used a fused feature algorithm technique that consists of HOG, Harlick and binary patterns. Then, a novel joint entropy-based feature selection algorithm is used along with a recognizer engine (i.e., multi-class SVM) to examine HIR behavior. In [42], O. 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 using Set I of UT-Interaction dataset. In [43], X. Ji et al. presented a vision based HIR system using multiple stage probability fusion. They divided interaction between two persons into the start state, execution state and the end state. Through weighted fusion, better recognition accuracy rates were obtained.

S. Bibi et al. [44] proposed a multi-feature model along with median compound local binary patterns for HIR system. They monitored individual action through multi-view cameras and showed better human-human interaction recognition rates. N. Cho et al. [45] described a novel system in order to identify complex human interaction identification. Their feature descriptors contained movements

at global, local and individual levels. They detected points of significance in order to identify human motion. Experiments on two publicly available datasets with a SVM classifier showed that their system produced a better accuracy rate. A human activity recognition system based on depth sensors is proposed by O.F. Ince et al. [46]. Their system, which is based on joint-angle features, can detect activities in 3D space. The Haar-wavelet transform and dimension reduction algorithm is also applied. K-nearest neighbor (KNN) is applied to recognize human actions. In order to track human interaction recognition, a wise human interaction and tracking model was proposed by M. Mahmood et al. [47]. They extracted data from spatio-temporal and angular-geometric features. They evaluated their system on three benchmark datasets and, as a result, the performance of the system was better than many state-of-the-art systems.

In order to recognize human interactions in both indoor and outdoor environments, an RGB-based HIR system was introduced by Jalal [48]. Multiple features are proposed in this research work and Convolution Neural Networks (CNN) was applied. CNN proved to be better than other state-of-the-art classifiers. N. Nguyen et al. proposed an HIR system motivated by the performance of deep learning methods [49]. Hierarchical invariant features are extracted using Independent Subspace Analysis (ISA) via three-layer CNN. Through experimentation they showed that their three-layer approach is better at recognizing human interaction in complex environments than other approaches. Motivated by the success of bag-of-words, an automated recognition system was proposed by K.N. Slimani et al. [50]. They extracted a 3D volume of spatio-temporal features. Each interaction between two persons is represented by the co-occurrence of words through their frequency. Inspired by the applications of information technology (IT) in the education sector, Jalal et al. proposed a student behavior recognition (SBR) system [51]. They extracted spatio-temporal features for identifying student-student interaction. They tested their system against one self-annotated and one RGB dataset. In [52] depth map-based person-person interaction is recognized. Interaction is divided into body part interactions. Regression based learning is used to process each camera view then features from multiple views are combined. The efficacy of the system was evaluated on three public depth-based datasets.

These methods mentioned above are either implemented on single RGB data or have used a very small set of features. On the other hand, we propose a vision based HIR system that consists of hybrid features having generic properties for RGB as well as depth images. For experimental validation, we use two depth datasets and one RGB dataset that consist of complex interactions over indoor-outdoor environments.

### 2.3. Marker Sensor-Based HIR Systems

In marker-based HIR systems, different markers, such as light emitting diodes, infrared or reflective spheres, are attached to the human body in order to capture motion information [53]. These sensors are attached to targeted body regions, such as joints or limbs of the human body. Many researchers used marker sensors for human activity analysis, clinical diagnosis and in rehabilitation centers. For example, M.H. Khan proposed a marker sensor-based system in order to provide home based therapy for patients [54]. Markers of different colors are attached to the individual's joints and motion information is recorded. Experiments were conducted on 10 patients which validated the performance of proposed systems. In [55], color markers are used to track foot positions. The motions of different body parts are tracked with the help of marker sensors and then interaction information between person and virtual surroundings is achieved. In order to analyze upper limb function of patients with abnormal limbs, a combination of a hand skateboard device, an IR camera and an infrared emitter is used [56]. Experiments showed that this system is easy to use and that it delivers results immediately.

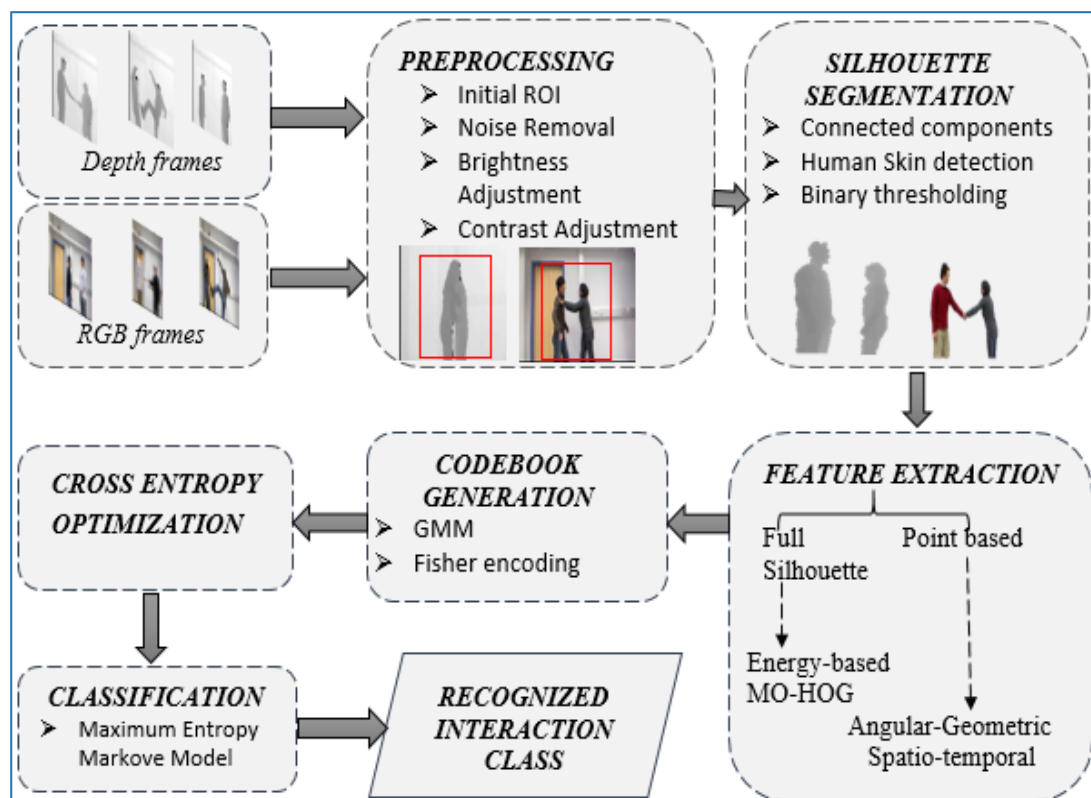
In order to perform biomechanical examinations and to capture motion in sports activities, marker based optical sensors are used [57]. For system evaluation, collegiate and elite baseball pitchers are used. A Trunk Motion System (TMS) was developed by M.I. Esfahani [58]. They used Body-Worn Sensors (BWS) in their system. They measured the 3D motions of the trunk. Their system is very lightweight. They attached 12 Body-Worn Sensors on stretchable clothing. Seven actions were

performed wearing BWS and the results were evaluated based on these actions. Motivated by a wide variety of applications using motion sensing in the healthcare department, a BWS based monitoring system was proposed by J. Michael et al. [59]. In order to measure the physical activities of humans, an innovative wireless system is proposed by N. Golestani et al. [60]. They proposed a magnetic induction system to track human actions. In this system, markers are attached to the joints. Successful evaluations were performed by applying laboratory measurements and deep recurrent neural network monitoring.

These sensors provide very accurate information regarding position, but they lack effectiveness in high speed motion because they cannot read and produce data on factors such as acceleration, velocity and torque. More precise results are generated via marker sensors, which provide better results in many clinical studies [61]. However, their performance was affected by surroundings such as dust, temperature changes and vibrations [62].

### 3. Proposed System Methodology

In this section, we describe details of each process involved in the proposed HIR system. Firstly, raw image (i.e., RGB and depth) sequences are preprocessed to remove noise. Then the segmentation algorithms are applied to extract the foregrounds from the backgrounds. Secondly, after segmentation, four different types of features are extracted as hybrid descriptor features. These feature descriptors are then fed into a codebook generation algorithm. Thirdly, cross entropy algorithms are applied to optimize the quantized codebook. Finally, experiments are performed and MEMM is used to determine each interaction class. Figure 1 shows the complete system architecture of the proposed HIR methodology.



**Figure 1.** System architecture of proposed human interaction recognition model.

#### 3.1. Image Acquisition

During image acquisition, we start with video normalization to extract human silhouette representations by applying various techniques for noise removal, handling varying scales and

contrasting distribution. For these purposes, all image sequences are first cropped to a fixed dimension to remove unnecessary areas. In order to enhance image quality, brightness and contrast, the distribution of both RGB and depth images are adjusted to make the images clearer via histogram equalization. Then, a smoothing filter is applied as mean filter [63], which calculates all mean values between a current pixel and its neighboring pixels. The mean filter of input signal  $x$  is given through Equation (1):

$$y[i] = \frac{1}{M} \cdot \sum_{i=0}^{M-1} x(i + j) \quad (1)$$

where  $y$  is the smoothened image,  $i$  and  $j$  are pixel values, and  $M$  is the window size, having a number of neighboring pixels.

### 3.2. Silhouette Representation

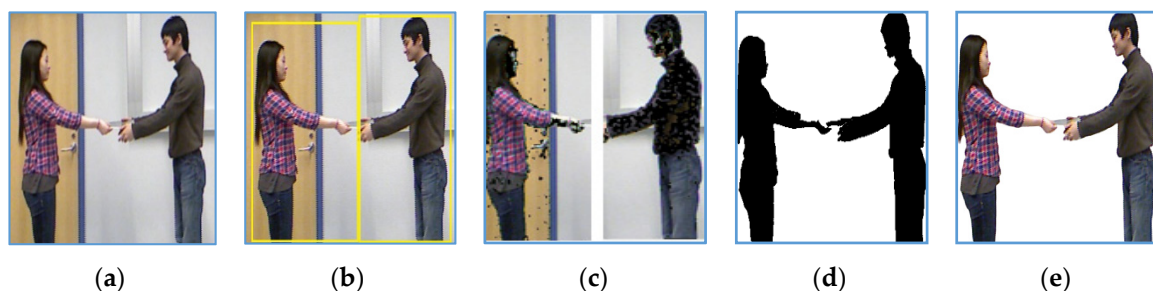
For robust identification of HIR, actual human interaction areas need to be extracted and to distinguish target images from clutter [64]. To extract efficient silhouette representation, we depend mainly on connected components, skin tone, region growing and color spacing [65]. Various algorithms are used for both RGB and RGB-D silhouette segmentation to improve the performance of the proposed system. We discuss this below.

#### 3.2.1. Silhouette Segmentation of RGB Images

RGB silhouette segmentation is performed on the basis of pixel connectivity analysis and skin detection [66]. Initially, we detect human silhouettes where connected components in an image are found using 8-connected pixel analysis. This technique seeks to identify horizontal, vertical and diagonal connections between pixels. Human silhouettes are then defined by auto-bounding on the boxes on the basis of height and width parameters. Next, to segment silhouettes from a noisy background, we apply coloring algorithms to identify all light intensity colors, such as yellow, skin color and white. These light intensity colors are then converted from RGB to luminance, chrominance (yCbCr) color space, which is formulated as:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112 \\ 112 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (2)$$

where  $Y$  is luminance,  $Cb$  and  $Cr$  represent blue difference and red difference chrominance. After the identification of light intensity colors, they are converted to black color. Then we apply threshold-based segmentation, which works as growing regions to segment humans from the background. Full procedure of RGB silhouette identification and segmentation is shown in Figure 2.



**Figure 2.** Example of RGB silhouette segmentation for an Exchanging object interaction of SBU dataset: (a) original image; (b) detected silhouettes; (c) skin coloring on cropped left and right silhouette; (d) binary thresholding over silhouettes and (e) segmented RGB silhouettes.

### 3.2.2. Silhouette Segmentation of Depth Images

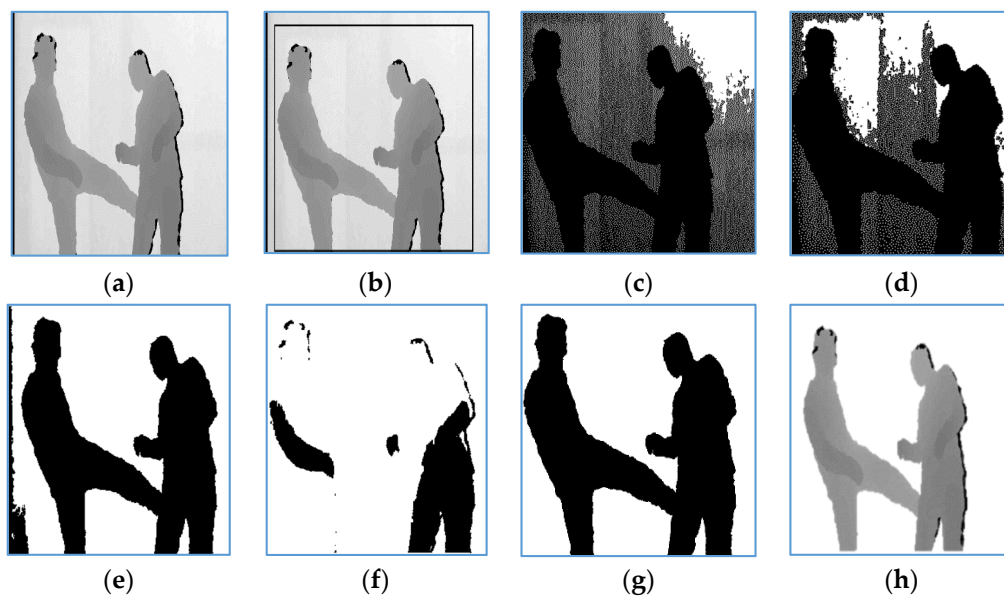
For silhouette segmentation of depth images, we used Otsu's thresholding method in which an image is divided into two classes—i.e., background class and foreground class [67]. In this method, multiple iterations with possible threshold values  $T$  are performed and one unique value of  $T$  is chosen that best separates foreground and background pixels. In order to calculate  $T$ , inter-class and intra-class variance are needed for analysis. In the case of intra-class analysis, variance should be as low as possible so, it is minimized through Equation (3):

$$\sigma_w^2(T) = w_0(T)\sigma_0^2(T) + w_1(T)\sigma_1^2(T) \quad (3)$$

where probabilities of both classes that are divided by  $T$ , is given by  $w_0$  and  $w_1$ . Variances of both classes are shown by  $\sigma_0^2$  and  $\sigma_1^2$ . On the other hand, variance between two classes—i.e., inter-class variance should be as high as possible, as shown through Equation (4):

$$\sigma_b^2(T) = \sigma^2 - \sigma_w^2(T) \quad (4)$$

In this way, depth silhouettes are separated from their background. Figure 3 demonstrates an example of the depth silhouette segmentation of kicking interaction from the SBU dataset.



**Figure 3.** Depth silhouette segmentation of kicking interaction from SBU dataset: (a) original image; (b) initial ROI; (c) binary image at  $T = 0.25$ , (d) binary image at  $T = 0.22$  (e) binary image at  $T = 0.20$ , (f) binary image at  $T = 0.13$ , (g) segmented binary silhouette at  $T = 0.19$ ; (h) segmented depth silhouette.

### 3.3. Hybrid Feature Extraction

After the extraction of silhouettes from complex backgrounds, we proposed a novel hybrid feature extraction method. This method is a fusion of key-body point features and full silhouette features. Spatio-temporal and angular-geometric features are based on key-body points while motion-orthogonal HOG and energy-based features are based on full silhouettes. These four novel features are extracted and discussed in sub-sections below.

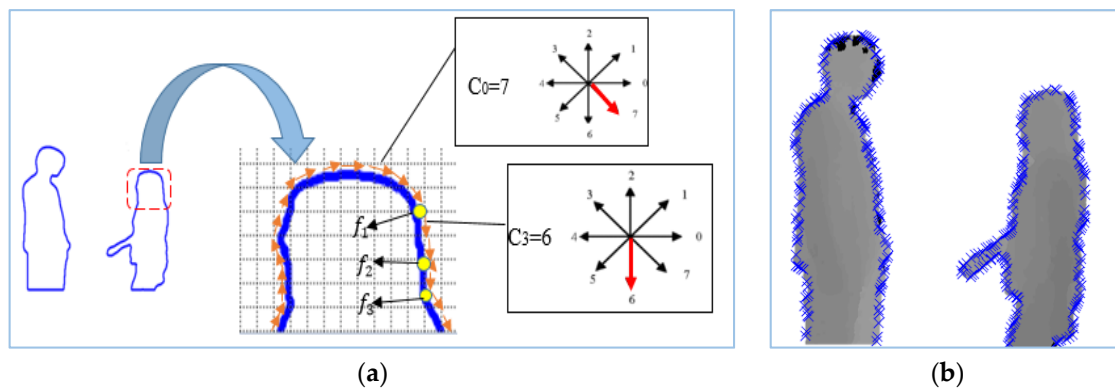
#### 3.3.1. Spatio-Temporal Feature

Spatial features give information regarding changes with respect to space, location or position [68]. For spatial features, we measured intensity changes along the curve points of the body using the 8 Freeman chain code algorithm. These features are extracted along the boundary of the human



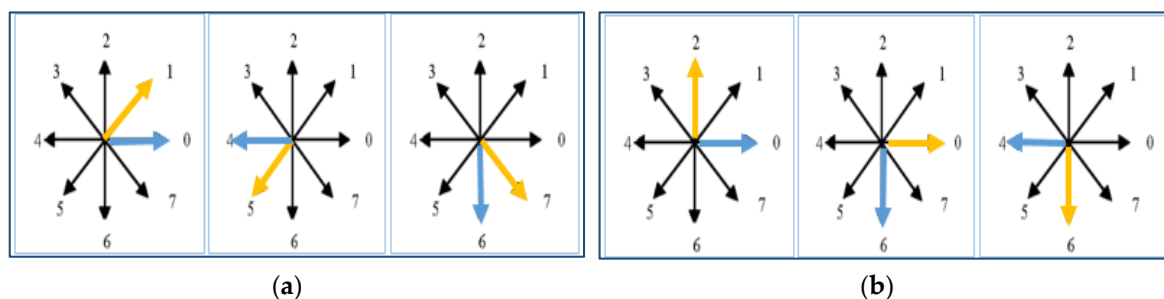
silhouette because a small change in the position of a human silhouette results in changes in the curves of silhouette. So, in order to extract spatial features, we first identified the boundaries of the two human silhouettes involved in the interaction. Then, all the curve points along the human contour of both silhouettes were identified and represented using the 8 Freeman chain code. If we suppose that all the points along the boundary  $b$  are represented by  $n$  points, then curve points  $C_b$  along the boundary are found from starting point  $C_0$  to  $n - 1$  as  $C_b = \{C_0, C_1, \dots, C_{n-1}\}$ .

Moreover, we start to find a feature point from curve  $C_0$  and move in a clockwise direction along with the boundary until we observe a change in direction from  $C_0$ . Suppose that  $C_0$  is the current curve point and  $C_1$  is the next point and if the direction of both  $C_0$  and  $C_1$  is the same then we will move to next curve point  $C_2$ . If the directions of both  $C_0$  and  $C_1$  are not the same, then we will consider  $C_1$  as a feature point  $f$  (see Figure 4a). So, we will consider a curve point to be a feature point  $f$  if the difference between current curve point and the next curve point is not equal to 0. In this way, spatial feature finds almost all the parts of body of a human silhouette (see Figure 4b). Figure 4 demonstrates the overall procedure to find the feature points using the 8 Freeman chain code.



**Figure 4.** Spatial feature extraction: (a) method to find a spatial feature point; (b) depth silhouette of approaching interaction with marked feature points.

In order to find a feature point, we have taken eight cases of  $45^\circ$  and four cases of  $90^\circ$  to find changes in the direction of each curve point. Figure 5 describes a few cases of  $45^\circ$  and  $90^\circ$  change in direction in which yellow arrows show the current curve point direction while the blue arrows show the subsequent curve point direction.



**Figure 5.** Cases of spatial feature point extraction: (a) three cases of  $45^\circ$  change in direction; (b) three cases of  $90^\circ$  change in direction.

Temporal features give information about changes with respect to time. In order to extract temporal features, critical displacement measurements between eight key-body points [69,70] are considered. Initially, our system tracked eight key-body points (head, left shoulder, right shoulder, left arm, right arm, left foot, right foot and torso) on detected RGB and depth silhouettes. These silhouettes were converted to binary and then the outer boundaries of silhouettes were identified. Then, different positions,

such as the topmost, right most, left most, bottom left most, bottom right most and center point of a human silhouette, are identified. Algorithm 1 presents the overall procedure used for the key-body point detection of human silhouettes.

---

**Algorithm 1** Detection of key-body points human silhouette

---

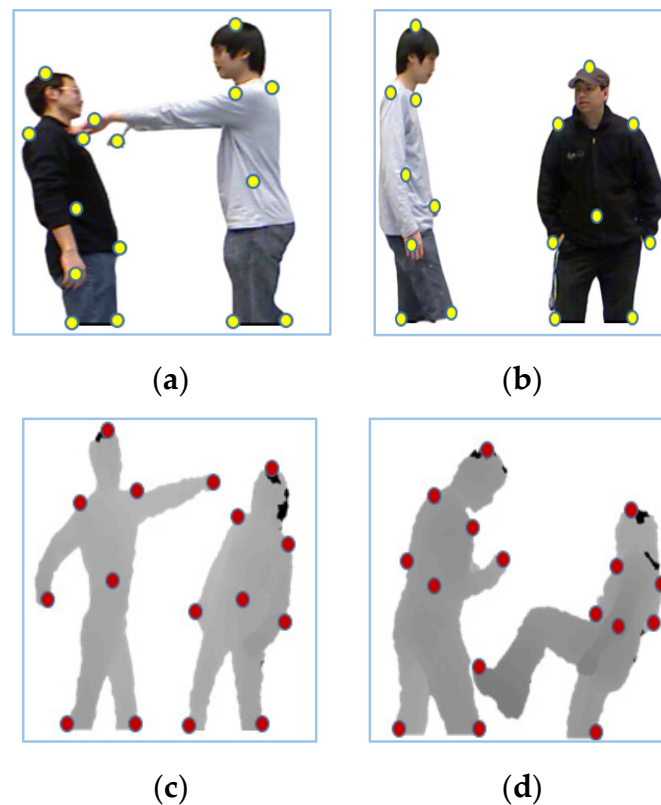
**Input:** S: Segmented Human Silhouettes  
**Output:** 8 key-body points as Head, Left-shoulder, Right-Shoulder, Left-arm, Right-arm, Left-foot, Right-foot, Torso.  
 B1 = boundary of left silhouette, B2 = boundary of right silhouette, H = height of silhouette, W = width of silhouette  
 % Extract boundaries of both silhouettes %  
 B = binarize(S);  
 BW = bwboundaries(B);  
 Object = detectobject(B,BoundingBox, Area)  
 % search boundaries of both silhouettes for outermost pixels %  
**for** i = 1 to B1  
**for** j = 1 to B2  
 Search (B1;B2)  
 Top\_Pixel = [x, y\_max];  
 Left\_pixel = [x\_min,y];  
 Rightl\_pixel = [x\_max,y];  
 Bottom\_left\_pixel = [x\_min,y\_min];  
 Bottom\_right\_pixel = [x\_max,y\_min];  
**end**  
**end**  
 % detect top, bottom, left and right region of both silhouettes %  
 % Repeat for both silhouettes %  
 [rows, cols] = size(object)  
 Top\_half = floor(rows/2)  
 bottom\_half = rows-Head\_region  
 Head\_region = floor(Top\_half/3)  
 Torso\_region = floor(Top\_half-Head\_region)  
 % identifying head region in top half of silhouette%  
 Head = Top\_Pixel(Head\_Region)  
 Left\_Shoulder = Bottom\_left\_pixel(Head\_Region)  
 Right\_Shoulder = Bottom\_right\_pixel(Head\_Region)  
 % identifying Torse, left arm and right arm%  
 X<sub>t</sub> = (W/2)  
 Y<sub>t</sub> = (H/2)  
 Torso = (X<sub>t</sub>, Y<sub>t</sub>)  
 Left\_arm = Left\_Pixel(Torso\_region)  
 Right\_arm = Right\_Pixel(Torso\_region)  
 % identifying left foot and right arm %  
 Left\_foot = Bottom\_left\_pixel(bottom\_half)  
 Right\_foot = Bottom\_right\_pixel(bottom\_half)  
**return** Head, Left-shoulder, Right-Shoulder, Left-arm, Right-arm, Left-foot, Right-foot, Torso.

---

After identifying key-body points, position displacement measurement between all key-body points of the first person's silhouette (silhouette of person on left side) and all key-body points of the second person's silhouette (silhouette of person on right side) are measured as shown in Equation (5):

$$D(p, q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \quad (5)$$

where  $D(p, q)$  is Euclidian distance,  $p_x$  and  $p_y$  are  $x, y$  coordinates of the key body points of the first person's silhouette and  $q_x$  and  $q_y$  are  $x, y$  coordinates of the second person's silhouette. As a person moves or performs any interaction, the distance between these key-body points may increase or decrease in values. Key-body points for both RGB and depth images are shown in Figure 6.

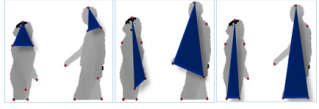
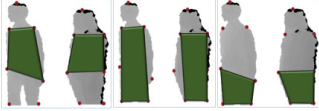

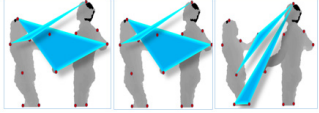
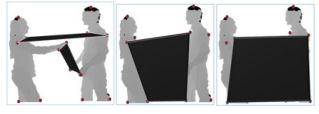
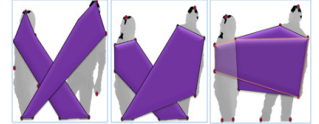


**Figure 6.** Key-body points on RGB and depth frames of: (a) pushing; (b) approaching; (c) punching; (d) kicking interaction of SBU dataset.

### 3.3.2. Angular–Geometric Features

An angular–geometric feature is a shape-based entity defined as a key-body point feature. In order to extract angular and geometric features, seven extreme body points (head, left shoulder, right shoulder, left arm, right arm, left foot, right foot) are first identified. Then, three geometric shapes—i.e., pentagon, quadrilateral and triangle—are made by joining these extreme points. In angular features, we measure changes in angular values between extreme point positions in consecutive frames. Two types of geometric shapes are made by joining these extreme points, such as: inter-silhouette shapes and intra-silhouette shapes. Table 1 shows a detailed overview of a number of inter-silhouette and intra-silhouette shapes and angles that are made by joining the extreme body points of silhouettes.

**Table 1.** Properties of inter-silhouette and intra silhouette geometrical shapes.

Type of Geometrical Shape	Connected Extreme Points	No. of Angles	Diagrammatical Representation
Inter-Silhouette Triangle	H1 <sup>1</sup> + RS1 <sup>2</sup> + LS1 <sup>3</sup> H1 + RA1 <sup>4</sup> + LA1 <sup>5</sup> H1 + RF1 <sup>6</sup> + LF1 <sup>7</sup> H2 <sup>8</sup> + RS2 <sup>9</sup> + LS2 <sup>10</sup> H2 + RA2 <sup>11</sup> + LA2 <sup>12</sup> H2 + RF2 <sup>13</sup> + LF2 <sup>14</sup>	18	
Inter-Silhouette Quadrangular	RS1 + LS1 + RA1 + LA1 RA1 + LA1 + RF1 + LF1 RS1 + LS1 + RF1 + LF1 RS2 + LS2 + RA2 + LA2 RA2 + LA2 + RF2 + LF2 RS1 + LS1 + RF1 + LF1	24	
Inter-Silhouette Pentagon	H1 + RS1 + LS1 + RA1 + LA1 H1 + RA1 + LA1 + RF1 + LF1 H1 + RS1 + LS1 + RF1 + LF1 H2 + RS2 + LS2 + RA2 + LA2 H2 + RA2 + LA2 + RF2 + LF2 H2 + RS1 + LS1 + RF1 + LF1	30	
Intra-Silhouette Triangle	H1 + RS2 + LS2 H1 + RA2 + LA2 H1 + RF2 + LF2 H2 + RS1 + LS1 H2 + RA1 + LA1 H2 + RF1 + LF1	18	
Intra-silhouette Quadrangular	RS1 + LS1 + RS2 + LS2 RA1 + LA1 + RA2 + LA2 RS1 + RS2 + RF1 + RF2 LS1 + LS2 + LF1 + LF2	16	
Intra-silhouette Pentagon	H1 + LS1 + RS1 + LF2 + RF2 H2 + LS2 + RS2 + LF1 + RF1 LS1 + RS1 + RA1 + LF2 + RF2 LS2 + RS2 + LA2 + LF1 + RF1 LS1 + RS1 + RA1 + LS2 + LA2 LS2 + RS2 + LA2 + LS1 + LA1	30	
Total 6 types	32 Geometrical Shapes	136 angles	

<sup>1</sup> Head of first (left) silhouette, <sup>2</sup> Right Shoulder of first silhouette, <sup>3</sup> Left Shoulder of first silhouette, <sup>4</sup> Right Arm of first silhouette, <sup>5</sup> Left Arm of first silhouette, <sup>6</sup> Right Foot of first silhouette, <sup>7</sup> Left Foot of first silhouette, <sup>8</sup> Head of second (right) silhouette, <sup>9</sup> Right Shoulder of second silhouette, <sup>10</sup> Left Shoulder of second silhouette, <sup>11</sup> Right Arm of second silhouette, <sup>12</sup> Left Arm of second silhouette, <sup>13</sup> Right Foot of second silhouette and <sup>14</sup> Left Foot of second silhouette.

Inter-silhouette shapes are made within each silhouette. These are geometric shapes made by connecting the extreme points of each silhouette individually. Intra-silhouette shapes are made between two silhouettes by connecting the extreme points of one silhouette with the extreme points of the second silhouette within each frame. After the completion of both types of geometric shapes, the inverse cosine angle is measured between all these shapes, as shown in Equation (6):

$$\theta = \cos^{-1} \frac{u \cdot v}{|u||v|} \tag{6}$$

where  $u$  and  $v$  are the two vectors in which the angle is measured. After measuring the angles, the shape's areas of all the inter-silhouette and intra-silhouette triangles are calculated. The area of the triangle is measured through Equation (7):

$$A_t = \sqrt{S(S - a)(S - b)(S - c)} \tag{7}$$

where  $a$ ,  $b$  and  $c$  are three sides of the triangle in which vectors are joined together—i.e., three extreme points to make a triangle—and  $S$  is the semi-perimeter of a triangle—i.e., half the length of the triangle's perimeter.

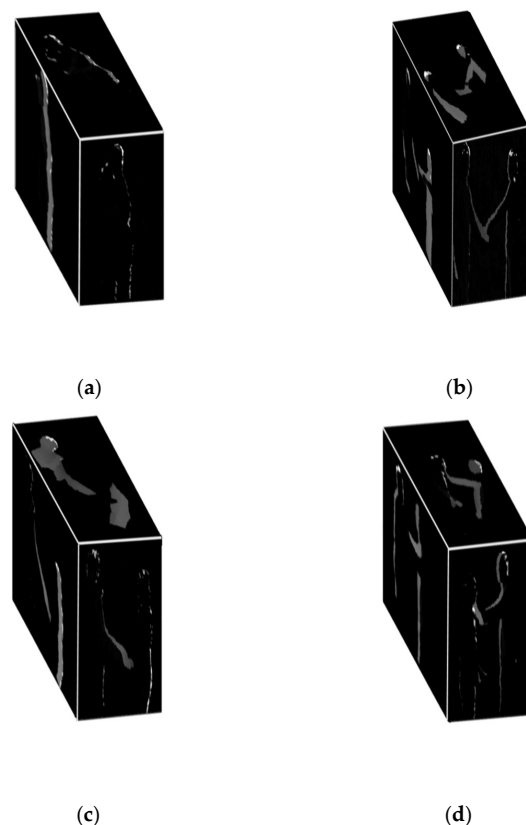
With the movement of each extreme point during interaction, the area of each geometric shape may increase or decrease. So, angular and geometric features measure changes in the angles as well as changes in the area of each shape between consecutive frames. The rate of change for the angles and the area are more evident in interactions like fight and kick, because they involve rapid movements of the extreme points during interaction as compared to approaching and departing interactions that include less pronounced movements of the extreme points.

### 3.3.3. Motion-Orthogonal Histogram of Oriented Gradient (MO-HOG)

MO-HOG is a motion-based feature applied over full silhouettes. It was observed that, in most of the interactions, the postures of both humans' silhouettes remain the same. For example, in approaching, departing, pushing and talking, the front views of both humans look like they are standing with only slight movements. Interactions like exchanging object and shaking hands are hardly distinguishable from each other. Punching and pushing interactions also have similar body movements. Therefore, our system proposed a novel multi-views approach including front, side and top views of both RGB as well as depth silhouettes by using a 3D Cartesian planes approach [71]. In order to incorporate motion data, we created RGB and depth differential silhouettes (DS) by taking differences between top  $t$ , front  $f$  and side  $s$  views of two consecutive frames as defined by Equation (8):

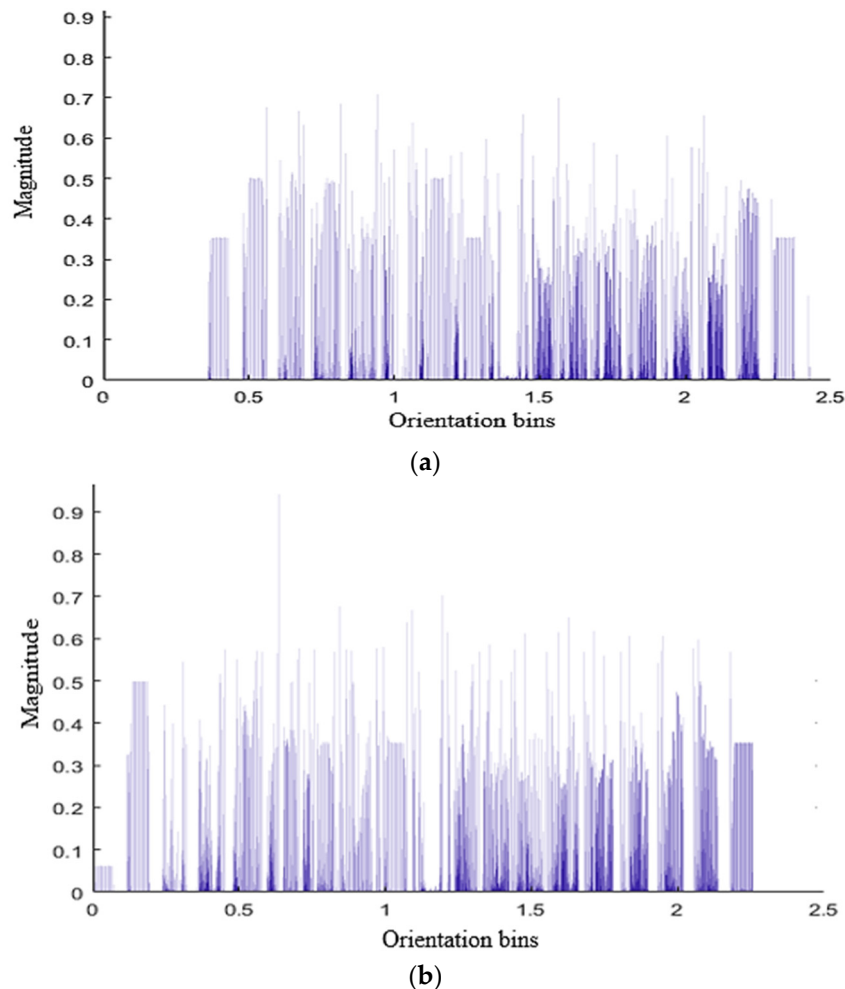
$$DS(Fc)_{f,s,t} = |Fc_{f,s,t} - Fp_{f,s,t}| \quad (8)$$

where  $Fc$  is current frame and  $Fp$  is previous frame. After taking DS of multi-views of silhouettes, they are projected as 3-dimensional (3D) Cartesian planes in the form of orthogonal shapes, as shown in Figure 7.



**Figure 7.** Orthogonal projection of 3D views of DS for: (a) hugging, (b) shaking hands, (c) kicking and (d) pushing interactions.

These multi-view DS are fed into HOG to extract orientation features [72]. It calculates magnitude and gradient by dividing the image into  $8 \times 8$  cells which are stored in a 9-bin histogram. A bar graph shows the magnitude and orientation bins of different interactions in Figure 8.



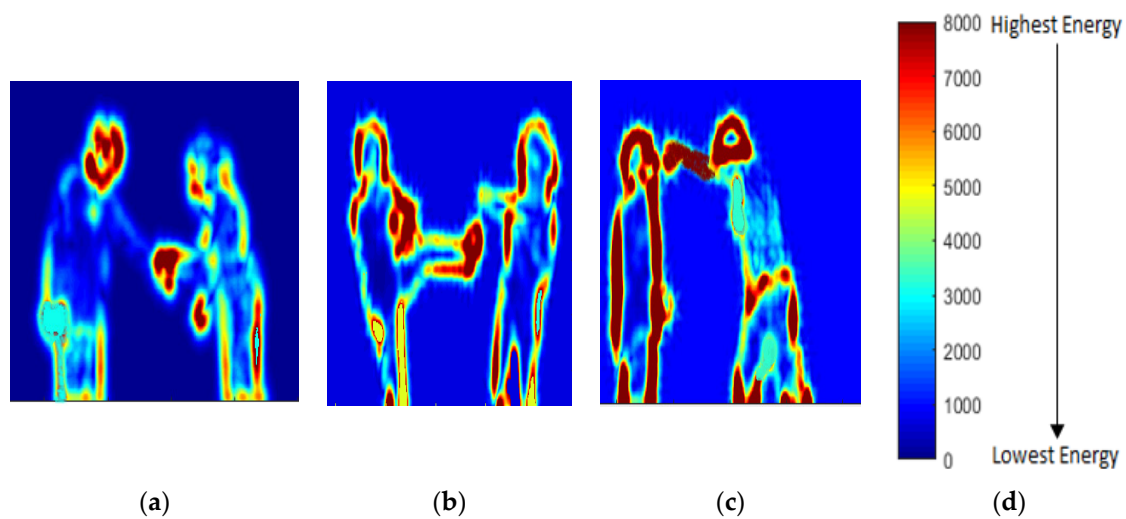
**Figure 8.** A bar graph showing HOG feature vector of (a) hugging and (b) kicking interactions.

### 3.3.4. Energy-Based Features

In energy-based features, the movements of human body parts are captured in the form of Energy Maps (EMs). EMs distribute the energy matrix between a set of  $[0-8000]$  indexes values over the detected silhouette. After energy distribution, a threshold-based technique is used in which only higher energy index values that are greater than a specified threshold are extracted into a 1D vector. Energy distribution is represented by Equation (9):

$$ER(v) = \sum_0^N \ln R(N) \quad (9)$$

where  $ER(v)$  is 1D energy vector,  $N$  is the index number, and  $\ln R$  is the RGB values of  $N$ . Energy distributions over some interactions of the SBU dataset are shown in Figure 9.



**Figure 9.** Energy features applied over (a) shaking hands; (b) kicking; (c) punching interactions of SBU dataset; (d) color bar showing energy range.

In Figure 9a, most of the energy is distributed in the region of hands. In Figure 9b, most of the energy is distributed in the left foot and the left shoulder because, when a person kicks, the upper body moves a little backward. Lastly, in Figure 9c, when the right silhouette starts punching, it moves forward, while the left silhouette moves backward as a reaction. Thus, energy distribution occurs at hands and the head of the right silhouette and around the whole body of the left silhouette. These energy maps show the energy of the body parts that are involved in the interaction in a red or a darker color. Meanwhile, those parts of the human body that are not involved during the interaction are in blue or lighter colors. Algorithm 2 explains the hybrid feature extraction algorithm.

---

**Algorithm 2** Hybrid feature extraction

---

**Input:** N: Segmented Silhouettes frames of RGB and RGB-D images

**Output:** Hybrid feature vectors( $k_1, k_2, k_3, \dots, k_n$ )

% initiating feature vector matrix %

Hybrid Feature-vectors  $\leftarrow$  []

Vectorsize  $\leftarrow$  GetVectorsize ()

% for loop on segmented silhouettes frames of all interaction classes %

**for**  $i = 1:N$

vectors\_interactions  $\leftarrow$  Getvectors(interactios)

% extracting spatio-tempora, MO-HOG, angular-geometric and energy features %

Spatio-temporal  $\leftarrow$  ExtractSpatioTemporalFeatures(vectors\_interactions)

MO-HOG  $\leftarrow$  ExtractOrthogonalHOGFeatures(vectors\_interactions)

Angular-Geometric  $\leftarrow$  ExtractAngularGeometricFeatures(vectors\_interactions)

Energy  $\leftarrow$  ExtractEnergyFeatures (vectors\_interactions)

Feature-vectors  $\leftarrow$  GetFeatureVectors (spatio-tempora, MO-HOG,

Angular-Geometric, Energy)

Hybrid Feature-vectors.append (Feature-vectors)

**end**

Hybrid Feature-vectors  $\leftarrow$  Normalize (Hybrid Feature-vectors)

**return** Hybrid Feature-vectors( $k_1, k_2, k_3 \dots \dots \dots k_n$ )

---

### 3.4. Codebook Generation

After extracting the hybrid features of both RGB and depth images, feature descriptors of all image sequences are combined to form a matrix representation. Such a matrix representation is so assorted

and complex that there is a need to represent it in a sorted and simpler way. Therefore, we applied Fisher vector encoding (FVC), based on GMM for codebook generation [73]. Initially, we applied GMM to compute the mean and covariance of each class, separately [74]. Based on computed values, clusters of each class are generated. Thus, the probability density function (pdf) of the cluster of the  $d$  dimensional vector  $X$  is defined by through Equation (10):

$$p(X; \theta) = \sum_{k=1}^K w_k N(X | \mu_k, \Sigma_k) \tag{10}$$

where  $\theta = \{w_k, \mu_k, \Sigma_k \mid k = 1, 2, \dots, K\}$ ,  $w_k$  is the weight of  $k$ th Gaussian component,  $K$  is the total number of clusters, the mean value is represented through  $\mu_k$ , the covariance matrix is given by  $\Sigma_k$  and  $N$  represents the distribution of  $d$  dimensional Gaussian. In addition, Expectation Maximization estimates the maximum likelihood of parameters of GMM. During expectation maximization soft assignment of vectors  $x_t$  to their belonging Gaussian cluster  $k$  is learned through Equation (11):

$$q_t(k) = \frac{w_k N(x_t; \mu_k, \Sigma_k)}{\sum_{j=1}^K w_j N(x_t; \mu_j, \Sigma_j)} \tag{11}$$

After applying GMM, FVC is performed on feature descriptors.  $X = \{x_t, t = 1, 2, \dots, T\}$  is a given feature set, while Gradient of log likelihood  $\nabla_{\theta}$  of  $X$  having GMM parameters  $\theta$  is given through Equation (12):

$$F_X = \frac{1}{T} \nabla_{\theta} \log p(X; \theta) \tag{12}$$

where  $F_X$  is feature vector. Now, the gradient vector is computed with respect to each mean  $\mu_k$  and covariance  $\sigma_k$  defined by Equations (13) and (14), respectively.

$$\mu_k = \frac{1}{T \sqrt{w_k}} \sum_{t=1}^T q_t(k) \frac{x_t - \mu_k}{\sigma_k} \tag{13}$$

Finally, all computed gradient vectors  $\mu_k$  and  $v_k$  for  $K$  components are combined to form  $D$ -dimensional final encoded feature vectors of dimensions 2KD. Hence, the Fisher vector reduces the intra-cluster gap and increases the inter-cluster gap, which gives a more precise discrimination of each cluster. Figure 10 demonstrates clusters formed by each interaction as a result of FVC over SBU and UoL 3D datasets.

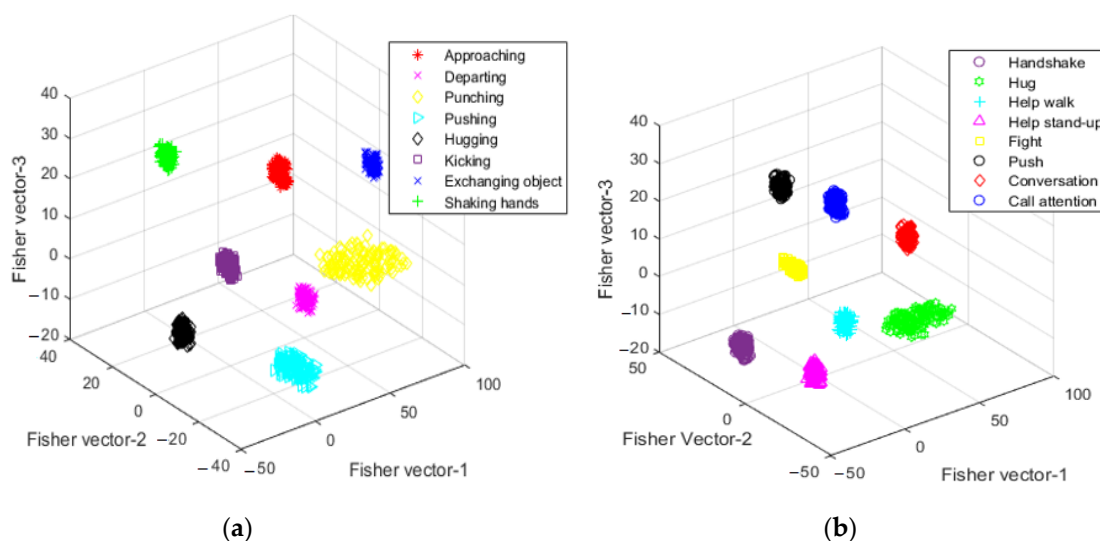


Figure 10. Three-dimensional clusters of FVC over: (a) SBU dataset and (b) UoL dataset.



### 3.5. Cross Entropy Optimization

In order to reduce the complexity of fisher encoded vectors, a cross entropy technique is implemented [75]. In cross entropy, initially, a sample of a specified size is generated from the fisher encoded vectors of each interaction class and an objective function is applied to that sample [76]. Then, more samples are extracted from encoded vectors and their objective functions are compared. This process continues until maximum numbers of iterations are reached or the best sample is obtained. The best sample of descriptors would represent an interaction class with the optimal set of descriptors. So, several iterations are performed until an optimal sample is generated. Cross entropy is measured among two probability distributions with samples  $p$  and  $q$ , and this is represented through Equation (15):

$$H(p, q) = -\sum_x p_x \log q_x \tag{14}$$

where  $p_x$  and  $q_x$  are probabilities of event  $x$  (i.e.,  $p_x$  is an actual or true value of probability and  $q_x$  is the predicted value). Meanwhile, Kullback–Leibler divergence  $D$  is calculated between true probability and predicted probability by Equation (16):

$$D_{KL}(p|q) = \sum_{i=1}^N p(x_i) \cdot (\log p(x_i) - \log q(x_i)) \tag{15}$$

In this way, the difference between the true and the predicted probability of a given sample is calculated. Cross entropy between the predicted and true probability distribution of each class of SBU dataset is shown in Figure 11.

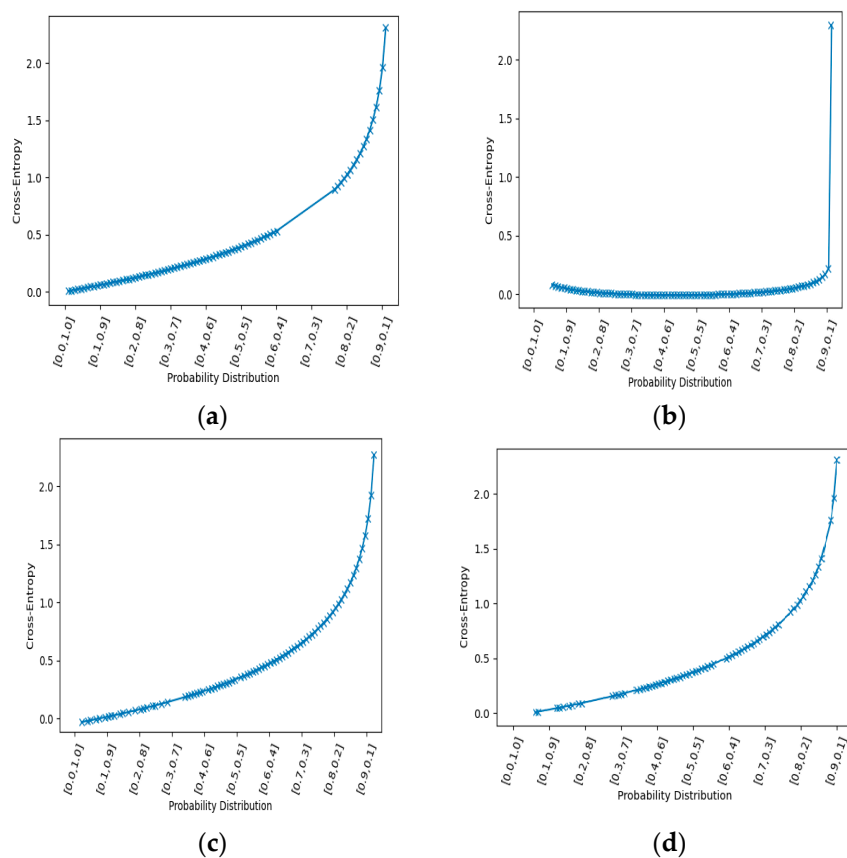
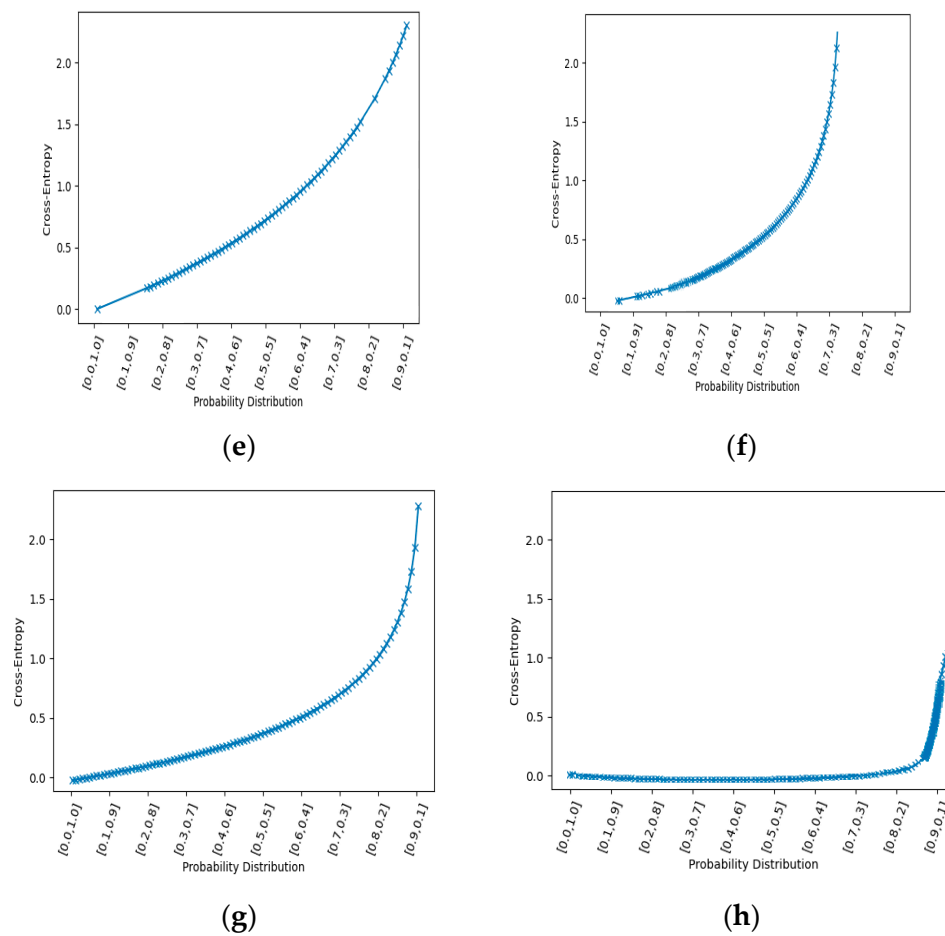


Figure 11. Cont.



**Figure 11.** Cross entropy between probability distributions of interaction classes of SBU dataset: (a) approaching; (b) departing; (c) exchanging object; (d) punching; (f) kicking; (g) hugging; (h) shaking hands.

### 3.6. Classification Via MEMM

After getting an optimized representation of the vectors, they are fed into a maximum entropy-based classifier in order to determine the different interaction classes (see Algorithm 3). MEMM is a combination of both the HMM and the maximum entropy model [77]. It is a discriminative model where a conditional probability is used to predict the interaction class. Such conditional probability is represented as  $P(S|S', X)$ . Each transition between state and observation in the MEMM is given through a log linear model, which is represented in Equation (16):

$$P_{S',}(S|X) = \frac{1}{Z(X,S')} \exp\left(\sum_k \lambda_k f_k(X,S)\right) \tag{16}$$

where  $S$  is the current state,  $S'$  is the next state,  $X$  is an observation,  $f_k$  is a feature function of  $X$  and possible  $S'$ ,  $Z(X,S')$  is a normalization factor that ensures the matrix sum and  $\lambda_k$  is the weight to be learned and is associated with feature  $f_k$ . According to the above observations, it is clear that MEMM is not only dependent on current observations but also on the previously predicted interaction. Figure 12 shows the overall procedure of the MEMM over different interaction classes of SBU Kinect interaction dataset.

**Algorithm 3** Vector Optimization and Interactions Classification

---

```

Input: Hybrid feature vectors  $(V_1, V_2, \dots, V_N)$ 
GMM parameters  $\theta = \{\pi_k, \mu_k, \Sigma_k \text{ where } k = 1, 2, \dots, K\}$ 
Output: Recognized Interaction  $I = \{I_1, I_2, I_3, \dots, I_n\}$ 
          % Fisher Vector Encoding %

FisherVector  $\leftarrow []$ 
for  $V = 1:V_N$  where  $V_N$  is total no. of vectors
deviance_mean  $\leftarrow$  ComputeGradientVector( $\pi_k$ )
deviance_covariance  $\leftarrow$  ComputeGradientVector( $\Sigma_k$ )
          %concatenate deviance w.r.t mean and covariance matrix of all vectors in N%
FisherVectors  $\leftarrow$  Concatenate(deviance_mean, deviance_stand_dev)
FisherVectors  $\leftarrow$  FisherVector.append(FisherVector)
end

          % Cross Entropy Optimization %

Best_Sample  $\leftarrow []$ 
while  $t < T$  where  $t$  is current iteration and  $T$  is total number of iterations
for  $i = 0:S_N$  where  $S_N$  is maximum no. of Samples
Sam  $\leftarrow$  ExtractSamples(FisherVectors)
end
ComputeSamplePerformance (Sam)
ComparePerformance (Sam, Best_Sample)
SortSamplebyPerformance (Sam)
Selected_Sample  $\leftarrow$  ChooseBestSample (Sam)
Best_Sample  $\leftarrow$  (Selected_Sample)
end
return optimized vector

          % Classifying interactions via MEMM %

Recognized interaction  $\leftarrow []$ 
Initialize State  $S = \{X_1, X_2, \dots, X_T\}$  where  $T =$  total no. of states
Observations =  $\{O_1, O_2, \dots, O_N\}$  where  $O_N$  is total no. of observations
          % Suppose a random state to be a current state %

 $X_t \leftarrow$  CurrentState
while state
          % suppose  $S_f$  state to be determining state %

 $S_f \leftarrow$  StatetoFind

          % ComputeCotditionalprobability %

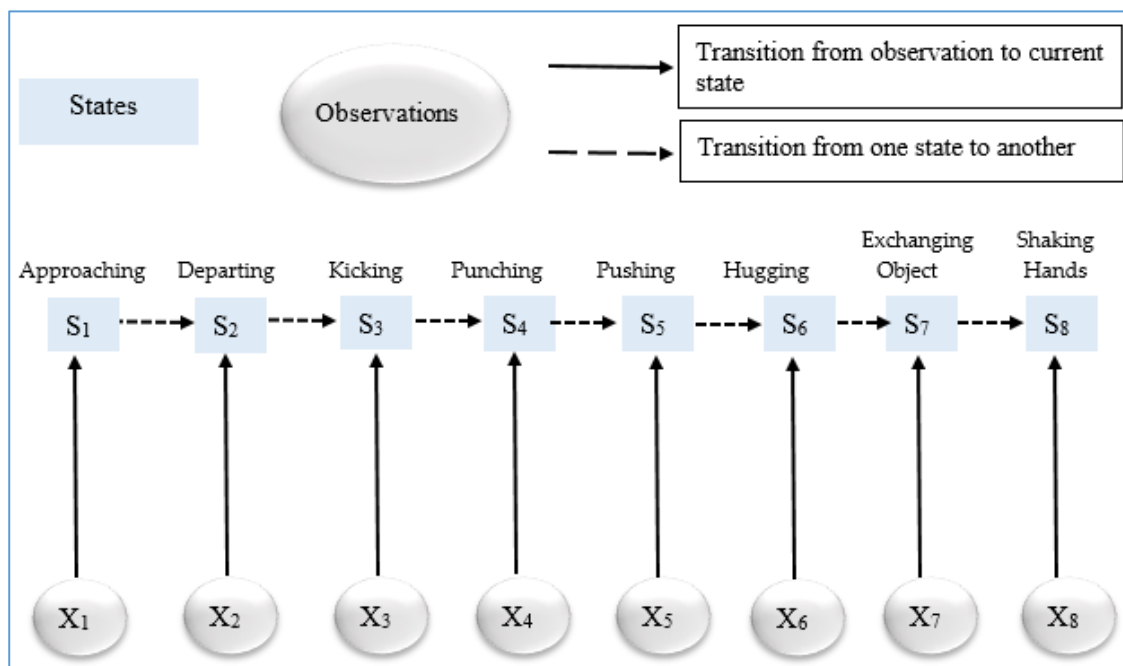
 $S_f \leftarrow$  ComputeStatetoFind( $S_f | S_{f-1}, O_t$ )
state  $\leftarrow S_f$ 
 $X_t \leftarrow$  state
end
return state as Recognized interaction  $\{I_1, I_2, I_3, \dots, I_n\}$ 

```

---

**4. Experimental Setting and Results**

In this section, we report training/testing experimentation results using the  $n$ -fold cross validation method over three publicly available benchmarks datasets. SBU Kinect interaction and UoL 3D datasets include both RGB and depth image sequences while UT interaction dataset consists of RGB data only. Furthermore, complete descriptions of each dataset are given in this section. The proposed model is evaluated on the basis of various performance parameters—i.e., computation time, recognition accuracy, precision, recall, F1 Score, number of states and number of observations. Discussion about various well-known classifiers and comparison of the proposed HIR with other statistically known state-of-the-art HIR methods is also given in this section.

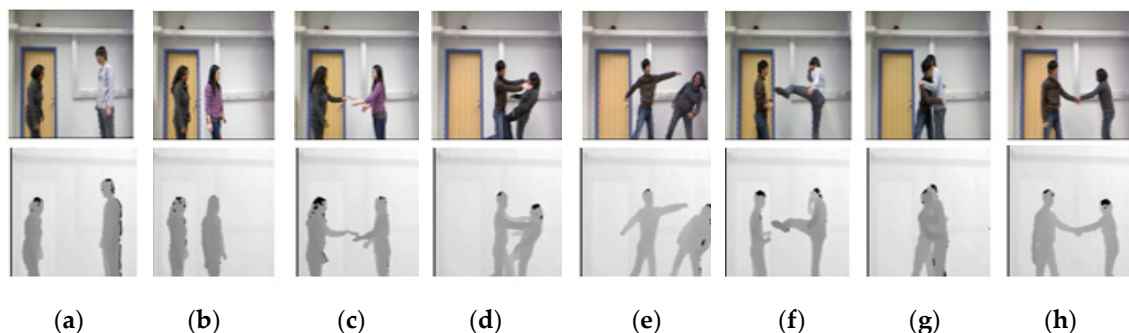


**Figure 12.** Overall flow of MEMM recognizer engine at different interaction classes of SBU Kinect interaction dataset.

#### 4.1. Datasets Description

##### 4.1.1. SBU Kinect Interaction Dataset

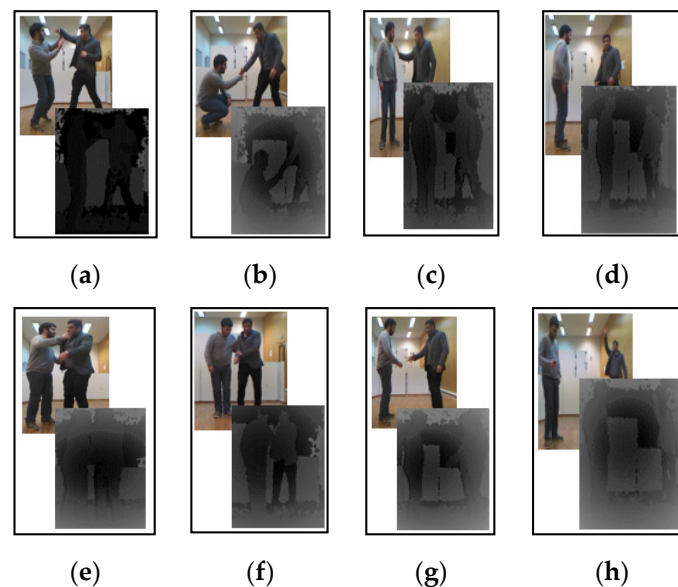
The SBU Kinect interaction dataset [78] consists of RGB, depth and skeletal information for the two-person performing interactions collected by Microsoft Kinect sensors in an indoor environment. Eight types of interactions including Approaching, Departing, Kicking, Punching, Pushing, Shaking Hands, Exchanging Object and Hugging are performed. The overall dataset is really challenging to interpret due to the similarity or closer proximity of movements in the different interaction classes. The sizes of both RGB and depth images are  $649 \times 480$ . Additionally, the dataset has a total of 21 folders, where each folder consists of all eight interaction classes performed by a different combination of seven actors. The ground truth labels of each interaction class are also provided. Videos are segmented at the rate of 15 frames per second (fps). Figure 13 shows some examples of human interaction classes of the SBU dataset.



**Figure 13.** RGB and depth snapshots of interaction classes of SBU dataset. (a) Approaching; (b) departing; (c) exchanging object; (d) pushing; (e) punching; (f) kicking; (g) hugging; (h) shaking hands.

#### 4.1.2. UoL 3D Dataset

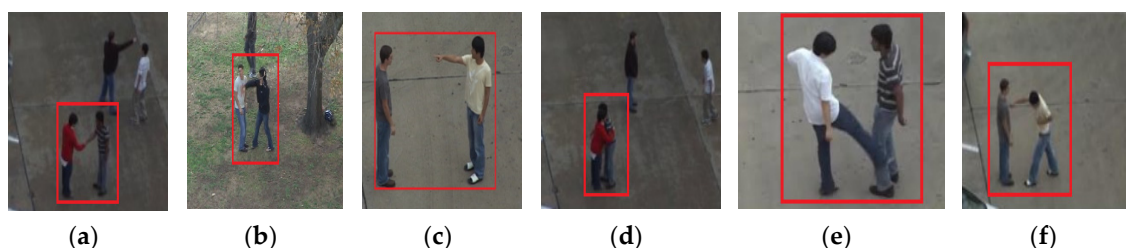
In the UoL 3D dataset, there is a combination of three types of interaction, such as casual daily life, harmful and assisted living interactions [79]. Included are interactions, such as handshake, hug, help walk, help stand-up, fight, push, conversation and call attention, performed by four males and two females. In addition, RGB, depth and skeletal information for each interaction is captured through the Kinect 2 sensor. Each folder has 24-bit RGB images, 8-bit and 16-bit resolution depth images of both 8-bit and 16-bit resolution and the skeletal information has 15 joints. There are ten different sessions of eight interactions performed by two subjects (in pairs), which are recorded in an indoor environment for period of 40–60 repetitions. This is a very challenging dataset and consists of over 120,000 data frames. Some snapshots of interactions of this dataset are shown in Figure 14.



**Figure 14.** RGB and depth snapshots of interaction classes of UoL 3D dataset: (a) fight; (b) help stand-up; (c) push; (d) conversation; (e) hug; (f) help walk; (g) handshake; (h) call attention.

#### 4.1.3. UT Interaction Dataset

The UT interaction dataset [80] consists of only RGB data. It has six interaction classes: point, push, shake hands, hug, kick and punch performed, by several participants with different appearances. This dataset is divided into two sets, named as: UT-Interaction Set 1 and UT-Interaction Set 2. The environment of Set 1 is a parking lot and the environment of Set 2 is a windy lawn. Video is captured with a resolution of  $720 \times 480$  at 30 fps. There are 20 videos per interaction providing a total of 120 videos of six interactions. Figure 15 demonstrates some examples of interaction classes for UT-Interaction dataset.



**Figure 15.** Few examples of interaction classes of UT-Interaction dataset. (a) Shake hands; (b) push; (c) point; (d) hug; (e) kick; (f) punch.

#### 4.2. Performance Parameters and Evaluation

In order to validate the methodology of the proposed HIR system, four different types of experiments with various performance parameters—i.e., recognition accuracy, precision, recall, F-score, computational time and comparison with state-of-the-art methods—were performed. Details and observations for each experiment are discussed in the sub-sections.

##### 4.2.1. First Experiment

In the first experiment, optimized feature vectors are subjected to MEMM in order to evaluate the average accuracy of the proposed system. We used the  $n$ -fold cross validation method for training/testing over three benchmark datasets. Tables 2 and 3 show the accuracy of the interactions of SBU and UoL datasets in the form of a confusion matrix. Similarly, recognition accuracies of UT-Interaction Set 1 and Set 2 are shown in Tables 4 and 5, respectively. While, the mean accuracy of the SBU dataset is 91.25%, the accuracy of UoL is 90.4% and the combined accuracy of the UT-Interaction Set 1 and Set 2 is 87.4%.

**Table 2.** Confusion matrix showing accuracies over interaction classes of SBU dataset.

Interaction Classes	Approaching	Departing	Kicking	Punching	Pushing	Hugging	Exchanging Object	Shaking Hands
Approaching	<b>0.92</b>	0.04	0	0	0	0	0.02	0.02
Departing	0	<b>0.95</b>	0	0	0.03	0.01	0.01	0
Kicking	0	0	<b>0.98</b>	0.02	0	0	0	0
Punching	0.02	0	0.03	<b>0.89</b>	0.05	0.01	0	0
Pushing	0.01	0.02	0	0.05	<b>0.88</b>	0.03	0	0.01
Hugging	0.01	0	0	0.02	0.02	<b>0.95</b>	0	0
Exchanging Object	0.04	0	0	0	0	0.02	<b>0.86</b>	0.08
Shaking Hands	0.05	0	0	0	0	0.01	0.07	<b>0.87</b>
Mean Recognition Accuracy rate = 91.25%								

**Table 3.** Confusion matrix showing accuracies over interaction classes of UoL dataset.

Interaction Classes	Handshake	Hug	Help Walk	Help Stand-up	Fight	Push	Conversation	Call Attention
Handshake	<b>0.85</b>	0	0.05	0.07	0	0	0.03	0
Hug	0	<b>0.93</b>	0.06	0	0	0.01	0	0
Help walk	0.02	0.05	<b>0.89</b>	0.04	0	0	0	0
Help stand-up	0.09	0	0	<b>0.87</b>	0	0.04	0	0
Fight	0	0	0	0	<b>0.95</b>	0.03	0.02	0
Push	0	0	0.01	0	0.07	<b>0.91</b>	0.01	0
Conversation	0.02	0	0	0	0	0	<b>0.91</b>	0.07
Call Attention	0	0	0	0	0	0	0.08	<b>0.92</b>
Mean Recognition Accuracy rate = 90.4%								

**Table 4.** Confusion matrix showing accuracies over interaction classes of UT-Interaction Set 1.

Interaction Classes	Shake Hands	Point	Hug	Push	Kick	Punch
Shake Hands	<b>0.93</b>	0.05	0	0.02	0	0
Point	0.04	<b>0.88</b>	0	0.05	0	0.03
Hug	0.03	0	<b>0.89</b>	0.05	0	0.03
Push	0	0.03	0.03	<b>0.84</b>	0	0.10
Kick	0	0.02	0	0	<b>0.90</b>	0.08
Punch	0.02	0.03	0	0.08	0.03	<b>0.84</b>
Mean Recognition Accuracy rate = 88.0%						

From the experimental results, it is observed that our hybrid features methodology, along with cross entropy optimization and the MEMM, can clearly recognize human interactions better. However, some confusion is observed between pairs of similar interactions, such as shaking hands and exchanging object, and punching and pushing interactions, in the SBU dataset. In the UoL dataset, confusion is

observed between handshake and help stand-up interactions. Such confusion is due to the similarity in body movements involved in these interactions. In the UT-interaction dataset, there is confusion between shaking hands and point, and push and punch interactions due to similarities of these interactions. In addition, it is also observed that when combinations of RGB and depth vectors were fed into the MEMM, we achieved better recognition rates compared to RGB alone. The recognition rate of the RGB dataset i.e., UT interaction (87.4%) is less than those of the SBU and the UoL datasets, which are 91.25% and 90.4%, respectively. Thus, incorporating depth information results causes improvements in accuracy rate.

**Table 5.** Confusion matrix showing accuracies over interaction classes of UT-Interaction Set 2.

Interaction Classes	Shake Hands	Point	Hug	Push	Kick	Punch
Shake Hands	<b>0.90</b>	0.05	0	0.03	0	0.02
Point	0.05	<b>0.87</b>	0	0.04	0	0.04
Hug	0.02	0	<b>0.88</b>	0.06	0	0.04
Push	0	0.01	0.03	<b>0.85</b>	0	0.11
Kick	0	0	0.01	0.04	<b>0.89</b>	0.06
Punch	0.04	0.05	0	0.09	0	<b>0.82</b>
<b>Mean Recognition Accuracy rate = 86.8%</b>						

#### 4.2.2. Second Experiment

In the second experiment, precision, recall and F1 Score for each interaction class of three datasets are evaluated, as shown in Table 6.

**Table 6.** Comparison of precision, recall and F1 score over three benchmark datasets.

Datasets	Interactions	Precision	Recall	F1 Score
SBU	Approaching	0.88	0.92	0.90
	Departing	0.94	0.95	0.95
	Kicking	0.97	0.98	0.98
	Punching	0.91	0.89	0.90
	Pushing	0.90	0.88	0.89
	Hugging	0.92	0.95	0.94
	Exchanging Object	0.90	0.86	0.88
	Shaking Hands	0.89	0.87	0.88
<b>Average</b>		<b>91.037</b>	<b>91.25</b>	<b>91.50</b>
UoL	Handshake	0.87	0.85	0.86
	Hug	0.95	0.93	0.94
	Help walk	0.88	0.89	0.89
	Help Stand-up	0.89	0.87	0.88
	Fight	0.93	0.95	0.94
	Push	0.92	0.91	0.91
	Conversation	0.87	0.91	0.89
	Call Attention	0.93	0.92	0.92
<b>Average</b>		<b>90.50</b>	<b>90.37</b>	<b>90.38</b>
UT-Interactions Set 1	Shake Hands	0.91	0.93	0.92
	Point	0.87	0.88	0.88
	Hug	0.97	0.89	0.93
	Push	0.81	0.84	0.82
	Kick	0.97	0.90	0.93
	Punch	0.78	0.84	0.81
<b>Average</b>		<b>88.50</b>	<b>88.0</b>	<b>88.16</b>
UT-Interactions Set 2	Shake Hands	0.89	0.9	0.9
	Point	0.89	0.87	0.88
	Hug	0.96	0.88	0.92
	Push	0.77	0.85	0.81
	Kick	1.0	0.89	0.94
	Punch	0.75	0.82	0.78
<b>Average</b>		<b>87.66</b>	<b>86.83</b>	<b>87.16</b>

It is observed that, in the SBU dataset, the Approaching interaction has the least precise rate of 88% and it also has a highest rate of false positive. This is because many periodic actions of many interactions such as departing, shaking hands and exchanging object are similar to the approaching interaction. On the other hand, the kicking interaction gives the most precise results with a less false positive ratio of 3%. In the UoL dataset, Hug interaction gives the most precise result of 95% because the periodic actions performed during the Hug interaction are different from the other interactions of this dataset. Handshake and conversation interactions have the highest false positive ratios of 13% and 14%, respectively, because body movements of silhouettes during these two interactions are similar to many other interactions. Overall, if we compare three datasets, the precision recall and F1 score ratios of both sets of the UT Interaction dataset are less as compared to the SBU and UoL datasets.

#### 4.2.3. Third Experiment

In the third experiment, nine sub-experiments for each dataset were performed. In this experiment, different combinations of the two parameters (i.e., number of states and observations) were used to evaluate the performance of MEMM. As a result, comparisons are made in terms of time complexity and recognition accuracy. During MEMM, each transition not only depends on the current state but also on the previous state. Therefore, increasing the number of states and observations affects the performance rate of HIR. Tables 7–9 show a comparison of number of states and observations for time complexity and recognition accuracy over the SBU, UoL 3D and UT-Interaction datasets.

**Table 7.** Comparison of number of states and observations of MEMM over SBU dataset.

Parameters		Performance	
Number of States	Observations	Computational Time (sec)	Accuracy (%)
4	X = 10	15.5	76.8
	X = 20	18.3	78.5
	X = 30	23	79.2
5	X = 10	21.5	82.8
	X = 20	26.4	83
	X = 30	30.8	86
6	X = 10	22.2	88.6
	X = 20	30.6	89
	X = 30	37.9	91.20

**Table 8.** Comparison of number of states and observations of MEMM over UoL dataset.

Parameters		Performance	
Number of states	Observations	Computational time (sec)	Accuracy (%)
4	X = 15	17.5	78
	X = 25	25.3	80.5
	X = 35	28.5	81.9
5	X = 15	25.2	83.8
	X = 25	32.9	86
	X = 35	40.0	88
6	X = 15	45.2	88.9.6
	X = 25	48.2	90
	X = 35	49	90.8



**Table 9.** Comparison of number of states and observations of MEMM over UT-Interaction dataset.

Parameters		Performance	
Number of States	Observations	Computational Time (sec)	Accuracy (%)
3	X = 10	10.2	69.8
	X = 20	12.1	70.2
	X = 30	15	70.9
4	X = 10	14	74.2
	X = 20	16.4	76.3
	X = 30	18.8	78.0
5	X = 10	22.4	82.5
	X = 20	25.7	85.2
	X = 30	38.0	88.5

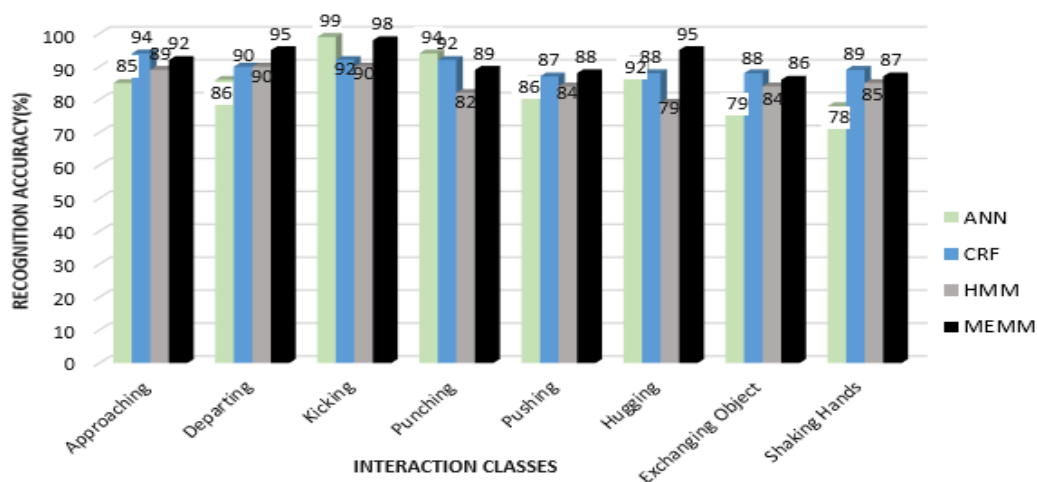
In Table 7, by using four states and changing the number of observations from 10 to 30, computational time and recognition accuracy were gradually increased. These experiments are repeated for five and six states. Similarly, Table 8 used 4–6 states and received significant results for computational time and recognition accuracy at 15 to 35 numbers of observations. Table 9 presents the results of these experiments on Set 1 of the UT-Interaction dataset, respectively.

It is concluded from the third experiment that reducing the number of states to two reduces recognition accuracy and computational time. On the other hand, increasing the number of states to six results in increased computational time with no change in accuracy. However, similar patterns of observations are noticed in Tables 7–9 (i.e., increasing the number of states and observations results in increased computational time and accuracy as well).

#### 4.2.4. Fourth Experiment

In the fourth experiment, we compared our proposed system in two parts. In the first part, a hybrid descriptor-based MEMM classifier is compared with other commonly used classifiers. In the second part, the proposed system is compared with other statistically well-known state-of-the-art HIR systems.

In the first part, quantized features vectors are given to most commonly used classifiers—i.e., ANN, HMM and Conditional Random Field (CRF)—and compared with MEMM to find the HIR accuracy rates for the interactions of each dataset. Figure 16 shows a comparison of recognition accuracies for each interaction class of the SBU dataset using all four classifiers.

**Figure 16.** Comparison of other classifiers with MEMM over interaction classes of SBU dataset.

From Figure 16, it can be seen that the mean recognition accuracy for ANN is 87.3%, CRF is 90%, HMM is 85.3% and MEMM is 91.25%. It is observed that, in some interactions, such as exchanging object and shaking hands, CRF performed better than MEMM. Additionally, ANN performed better in a few interactions, such as kicking and punching. Overall accuracy using the MEMM was higher than for other classifiers. Figure 17 shows the comparison of recognition accuracies for each interaction class using the UoL dataset.

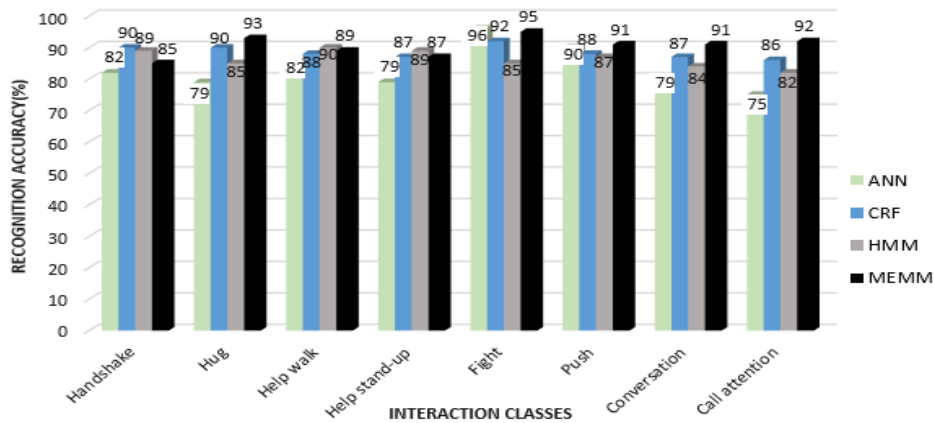
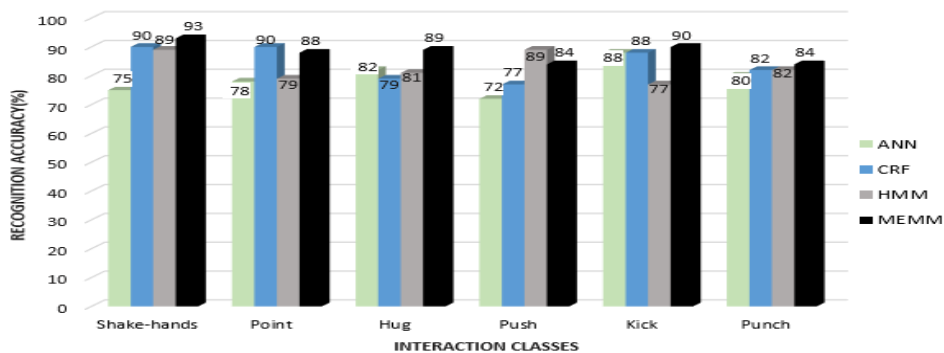
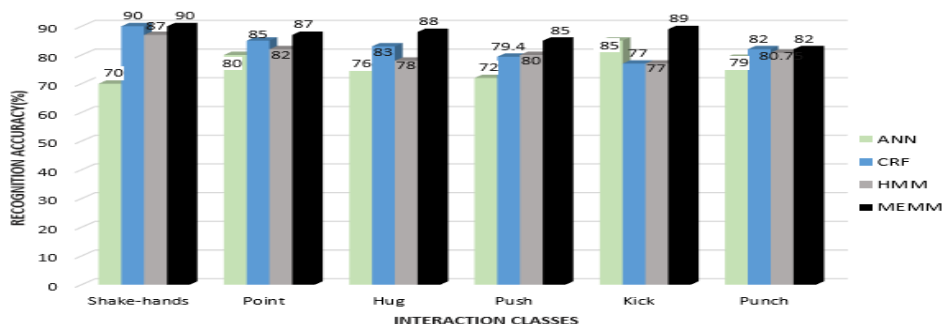


Figure 17. Comparison of other classifiers with MEMM over interaction classes of UoL dataset.

From Figure 17, it is shown that the mean recognition accuracy of ANN is 82.75%, CRF is 88.5%, HMM is 86.37% and MEMM is 90.4%, using the UoL dataset. It is observed that some interactions, such as fight in the case of ANN, handshake in the case of CRF and help walk in the case of HMM, achieved better recognition accuracy than the MEMM. However, the overall recognition rate was still higher with the MEMM. Figure 18 shows the comparison of four classifiers over interaction classes of the UT-Interaction Set 1 and Set 2.



(a)



(b)

Figure 18. Comparison of other classifiers with MEMM over interaction classes of UT-Interaction: (a) Set 1 and (b) Set 2.

From Figure 18a,b, it is observed that the mean recognition accuracy rates of Set 1 and of Set 2 for the UT-interaction dataset are less than the depth datasets. The mean accuracies of Set 1 of the UT Interaction dataset are 79.16% with ANN, 84.3% with CRF, 82.8% with HMM and 88% with the MEMM classifier. Mean accuracies are further reduced with Set 2 for the UT interaction dataset due to the cluttered background of a windy lawn. The mean accuracy for ANN is 77%, CRF is 82.7%, HMM is 80.7% and MEMM is 86.8%. Meanwhile, it is observed that patterns of recognition accuracies for Set 1 and for Set 2 are similar to those of the UoL and of the SBU datasets and that the MEMM has the highest accuracy rate while ANN has the lowest. Accuracy rates for the MEMM and CRF are comparable. Moreover, CRF, HMM and MEMM performed better in most of the interactions classes except for the fight interaction, where ANN has better or nearly similar recognition rate. However, overall MEMM has best recognition rates. Thus, it is concluded that MEMM based performance is best for HIR.

In second part of this experiment, the proposed HIR system is compared with other statistically well-known state-of-the-art systems. Table 10 presents a comparison of results for the SBU, UoL and UT interaction datasets, respectively.

**Table 10.** Comparison of proposed hybrid features HIR system with other state-of-the-art systems over SBU interaction dataset.

Methods	HIR Accuracy (%)			
	SBU Dataset	UoL Dataset	UT Interaction Set 1	UT Interaction Set 2
CFDM [81]	89.4			
CWDTW [82]	90.8			
CHARM [83]	84			
Joint Features [84]	90.3			
Body parts contrast mining [85]	86.9			
Deep LSTM [86]	90.41			
Skeletal data [87]	88	87		
Skeletal and Geometrical features [88]	-	85.56		
Spatio-temporal+ social features [30]	-	85.12		
Spatio-temporal features [89]			83.5	72.5
SPN Graph [90]			82.4	85.3
Discriminative model [91]			85	85
<b>Proposed hybrid feature</b>	<b>91.25</b>	<b>90.4</b>	<b>88.0</b>	<b>86.8</b>

## 5. Discussion

A unique HIR system is proposed in this research work. Four unique features are extracted from both RGB and depth silhouettes. The efficiency of the proposed model is proved through four types of experiments. However, certain challenges were faced during this research work. In the silhouette detection phase of the RGB frames, a connected components algorithm was used to identify connected objects. However, this algorithm does not give the best results as it confuses white or light color clothes of individual with the white wall background. Therefore, in order to tackle this problem, we applied a human skin detection algorithm as well as a pre-specified measurements ratio (i.e., the height and width) of the human performers. This ratio is compared with the height and width ratio of bounding boxes of connected components. Again, the specific height and width ratio of human causes the failure of silhouette detection due to frequent changes in scaling values of human posture.

## 6. Conclusions and Future Work

In this paper, we have proposed a novel HIR system to recognize human interactions using both RGB and depth environmental settings. The main accomplishments of this research work are: (1) we achieved adequate silhouette segmentation; (2) identification of key human body parts; (3) extraction of four novel features—i.e., spatio-temporal, MO-HOG, angular-geometric and energy based features; (4) cross-entropy optimization and recognition of each interaction via MEMM. In the first phase, both RGB and depth silhouettes are identified separately. For RGB silhouette segmentation, various skin colors, connected components and binary thresholding methods are applied to separate

humans from their background. After the extraction of silhouettes, all spatio-temporal features are extracted. In these features the displacement between key body points is identified via Euclidean distance. In angular-geometric features, various geometrical shapes are made by connecting the extreme points of silhouettes. The angles of these shapes are then measured in each interaction class. After that, MO-HOG features are extracted, in which differential silhouettes are projected from three different views and then HOG is applied. Finally, unique energy features are extracted from each interaction class. A Hybrid of these feature descriptors results in very complex vector representation. In order to reduce the complexity of the feature descriptors, a GMM based FVC is applied and then cross entropy optimization is performed.

During experimental testing, four different types of experiments were conducted on three benchmark datasets in order to validate the performance of the proposed system. In the first experiment, recognition accuracies for the interaction classes of each dataset were measured. In the second experiment, F1 scores, precision and the recall of each interaction class were measured and compared. In the third experiment, computation time and accuracy were measured by changing the number of states and observations of the MEMM classifier. Finally, in the fourth experiment, recognition accuracies for the interaction classes of each dataset were measured via the most commonly used classifiers—i.e., ANN, HMM and CRF—and compared with MEMM. Results showed better performance, with an average recognition rate of 87.4% for UT-Interaction, 90.4% for UoL and 91.25% for SBU datasets. Results of these experiments validated the efficacy of the proposed system. The proposed system is applicable to various real-life scenarios, such as security monitoring, smart home, healthcare and content-based video indexing and retrieval, etc.

In the future, we plan to implement the proposed method in a group of human interactions as well as human-object interactions. We will also use entropy-based features. We will also work on more challenging datasets.

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

**Funding:** This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2018R1D1A1A02085645).

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

## References

1. Li, J.; Tian, L.; Wang, H.; An, Y.; Wang, K.; Yu, L. Segmentation and Recognition of Basic and Transitional Activities for Continuous Physical Human Activity. *IEEE Access* **2019**, *7*, 42565–42576. [[CrossRef](#)]
2. Jalal, A.; Kamal, S.; Kim, D. A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments. *Sensors* **2014**, *14*, 11735–11759. [[CrossRef](#)] [[PubMed](#)]
3. Ajmal, M.; Ahmad, F.; Naseer, M.; Jamjoom, M. Recognizing Human Activities from Video Using Weakly Supervised Contextual Features. *IEEE Access* **2019**, *7*, 98420–98435. [[CrossRef](#)]
4. Susan, S.; Agrawal, P.; Mittal, M.; Bansal, S. New shape descriptor in the context of edge continuity. *CAAI Trans. Intell. Technol.* **2019**, *4*, 101–109. [[CrossRef](#)]
5. Shokri, M.; Tavakoli, K. A review on the artificial neural network approach to analysis and prediction of seismic damage in infrastructure. *Int. J. Hydromechatron.* **2019**, *4*, 178–196. [[CrossRef](#)]
6. Tingting, Y.; Junqian, W.; Lintai, W.; Yong, X. Three-stage network for age estimation. *CAAI Trans. Intell. Technol.* **2019**, *4*, 122–126. [[CrossRef](#)]
7. Zhu, C.; Miao, D. Influence of kernel clustering on an RBFN. *CAAI Trans. Intell. Technol.* **2019**, *4*, 255–260. [[CrossRef](#)]
8. Wiens, T. Engine speed reduction for hydraulic machinery using predictive algorithms. *Int. J. Hydromechatron.* **2019**, *1*, 16–31. [[CrossRef](#)]
9. Osterland, S.; Weber, J. Analytical analysis of single-stage pressure relief valves. *Int. J. Hydromechatron.* **2019**, *2*, 32–53. [[CrossRef](#)]

10. Zhao, W.; Lun, R.; Espy, D.D.; Reinthal, M.A. Rule Based Real Time Motion Assessment for Rehabilitation Exercises. In Proceedings of the IEEE Symposium Computational Intelligence in Healthcare and E-Health, Orlando, FL, USA, 9–12 December 2014. [[CrossRef](#)]
11. Al-Nawashi, M.; Al-Hazaimeh, O.M.; Saracee, M. A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. *Neural Comput. Appl.* **2017**, *28*, 565–572. [[CrossRef](#)]
12. Abdelhedi, S.; Wali, A.; Alimi, A.M. Fuzzy Logic Based Human Activity Recognition in Video Surveillance Applications. In Proceedings of the International Afro-European Conference for Industrial Advancement AECIA, Paris, France, 29 January 2016. [[CrossRef](#)]
13. Taha, A.; Zayed, H.; Khalifa, M.E.; El-Horbarty, M. Human Activity Recognition for Surveillance Applications. In Proceedings of the International Conference on Information Technology, Amman, Jordan, 12–15 May 2015. [[CrossRef](#)]
14. Xu, H.; Pan, Y.; Li, J.; Nie, L.; Xu, X. Activity Recognition Method for Home-Based Elderly Care Service Based on Random Forest and Activity Similarity. *IEEE Access* **2019**, *7*, 16217–16225. [[CrossRef](#)]
15. Chernbumroong, S.; Cang, S.; Atkins, A.; Yu, H. Elderly activities recognition and classification for applications in assisted living. *Expert Syst. Appl.* **2013**, *40*, 1662–1674. [[CrossRef](#)]
16. Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z. A Review on Human Activity Recognition Using Vision-Based Method. *J. Healthc. Eng.* **2017**, *2017*, 3090343. [[CrossRef](#)] [[PubMed](#)]
17. Jalal, A.; Kamal, S.; Kim, D. Human Depth Sensors-Based Activity Recognition Using Spatiotemporal Features and Hidden Markov Model for Smart Environments. *J. Comput. Netw. Commun.* **2016**, *1026*, 2090–7141. [[CrossRef](#)]
18. Ye, M.; Zhang, Q.; Wang, L.; Zhu, J.; Yang, R.; Gall, J. A Survey on Human Motion Analysis from Depth Data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 149–187. [[CrossRef](#)]
19. Aggarwal, J.K.; Xia, L. Human activity recognition from 3d data: A review. *Pattern Recognit. Lett.* **2014**, *48*, 70–80. [[CrossRef](#)]
20. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 4405–4425. [[CrossRef](#)]
21. Park, S.U.; Park, J.H.; Al-masni, M.A.; Al-antari, M.A.; Uddin, M.Z.; Kim, T.S. A Depth Camera-based Human Activity Recognition via Deep Learning Recurrent Neural Network for Health and Social Care Services. *Proced. Comput. Sci.* **2016**, *100*, 78–84. [[CrossRef](#)]
22. Nadeem, A.; Jalal, A.; Kim, K. Human Actions Tracking and Recognition Based on Body Parts Detection via Artificial Neural Network. In Proceedings of the International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020. [[CrossRef](#)]
23. Ahmed, A.; Jalal, A.; Kim, K. Region and Decision Tree-Based Segmentations for Multi-Objects Detection and Classification in Outdoor Scenes. In Proceedings of the International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 16–18 December 2019. [[CrossRef](#)]
24. Schadenberg, B.R. Predictability in Human-Robot Interactions for Autistic Children. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Korea, 11–14 March 2019. [[CrossRef](#)]
25. Cho, N.G.; Kim, Y.J.; Park, U.; Park, J.S.; Lee, S.W. Group Activity Recognition with Group Interaction Zone Based on Relative Distance Between Human Objects. *Int. J. Pattern Recognit. Artif. Intell.* **2015**, *29*, 1555007. [[CrossRef](#)]
26. Tang, Y.; Li, Z.; Tian, H.; Ding, J.; Lin, B. Detecting Toe-Off Events Utilizing a Vision-Based Method. *Entropy* **2019**, *21*, 329. [[CrossRef](#)]
27. Jalal, A.; Quaid, M.A.K.; Hasan, A.S. Wearable Sensor-Based Human Behavior Understanding and Recognition in Daily Life for Smart Environments. In Proceedings of the International Conference on FIT, Islamabad, Pakistan, 17–19 December 2018. [[CrossRef](#)]
28. Tahir, S.B.; Jalal, A.; Batool, M. Wearable Sensors for Activity Analysis Using SMO-based Random Forest over Smart home and Sports Datasets. In Proceedings of the ICACS, Lahore, Pakistan, 17–19 February 2020. [[CrossRef](#)]
29. Howedi, A.; Lotfi, A.; Pourabdollah, A. Exploring Entropy Measurements to Identify Multi-Occupancy in Activities of Daily Living. *Entropy* **2019**, *21*, 416. [[CrossRef](#)]

30. Ehatisham-Ul-Haq, M.; Javed, A.; Awais, M.A.; Hafiz, M.A.M.; Irtaza, A.; Hyun, I.L.; Tariq, M.M. Robust Human Activity Recognition Using Multimodal Feature-Level Fusion. *IEEE Access* **2019**, *7*, 60736–60751. [[CrossRef](#)]
31. Xu, H.; Liu, J.; Hu, H.; Zhang, Y. Wearable Sensor-Based Human Activity Recognition Method with Multi-Features Extracted from Hilbert-Huang Transform. *Sensors* **2016**, *16*, 2048. [[CrossRef](#)] [[PubMed](#)]
32. Jalal, A.; Quaid, M.A.K.; Sidduqi, M.A. A Triaxial Acceleration-Based Human Motion Detection for Ambient Smart Home System. In Proceedings of the IEEE International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 8–12 January 2019. [[CrossRef](#)]
33. Batool, M.; Jalal, A.; Kim, K. Sensors Technologies for Human Activity Analysis Based on SVM Optimized by PSO Algorithm. In Proceedings of the IEEE International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 27–29 August 2019. [[CrossRef](#)]
34. Quaid, M.A.K.; Jalal, A. Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm. *Multimed. Tools Appl.* **2020**, *79*, 6061–6083. [[CrossRef](#)]
35. Tahir, S.B.; Jalal, A.; Kim, K. Wearable Inertial Sensors for Daily Activity Analysis Based on Adam Optimization and the Maximum Entropy Markov Model. *Entropy* **2020**, *22*, 579. [[CrossRef](#)]
36. Khan, A.M.; Lee, Y.; Lee, S.Y.; Kim, T. Human Activity Recognition via an Accelerometer-Enabled-Smartphone Using Kernel Discriminant Analysis. In Proceedings of the International Conference on Future Information Technology, Busan, Korea, 21–23 May 2010. [[CrossRef](#)]
37. Capela, N.A.; Lemaire, E.D.; Baddour, N. Feature Selection for Wearable Smartphone-Based Human Activity Recognition with Able bodied, Elderly, and Stroke Patients. *PLoS ONE* **2015**, *10*, e0124414. [[CrossRef](#)] [[PubMed](#)]
38. Wenchao, J.; Zhaozheng, Y. Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks. In Proceedings of the ACM International Conference on Multimedia, New York, NY, USA, 26 October 2015. [[CrossRef](#)]
39. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [[CrossRef](#)]
40. Ahmed, A.; Jalal, A.; Kim, K. A Novel Statistical Method for Scene Classification Based on Multi-Object Categorization and Logistic Regression. *Sensors* **2020**, *20*, 3871. [[CrossRef](#)]
41. Sharif, M.; Khan, M.A.; Akram, T.; Younus, M.J.; Saba, T.; Rehman, A. A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. *EURASIP J. Image Video Process.* **2017**, *2017*, 89. [[CrossRef](#)]
42. Ouyed, O.; Said, M.A. Group-of-features relevance in multinomial kernel logistic regression and application to human interaction recognition. *Expert Syst. Appl.* **2020**, *148*, 113247. [[CrossRef](#)]
43. Ji, X.; Wang, C.; Ju, Z. A New Framework of Human Interaction Recognition Based on Multiple Stage Probability Fusion. *Appl. Sci.* **2017**, *7*, 567. [[CrossRef](#)]
44. Bibi, S.; Anjum, N.; Sher, M. Automated multi-feature human interaction recognition in complex environment. *Comput. Ind.* **2018**, *99*, 282–293. [[CrossRef](#)]
45. Cho, N.; Park, S.; Park, J.; Park, U.; Lee, S. Compositional interaction descriptor for human interaction recognition. *Neurocomputing* **2017**, *267*, 169–181. [[CrossRef](#)]
46. İnce, Ö.F.; Ince, I.F.; Yıldırım, M.E.; Park, J.S.; Song, J.K.; Yoon, B.W. Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor. *ETRI J.* **2020**, *42*, 78–89. [[CrossRef](#)]
47. Mahmood, M.; Jalal, A.; Kim, K. WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors. *Multimed. Tools Appl.* **2020**, *79*, 6919–6950. [[CrossRef](#)]
48. Jalal, A.; Mahmood, M.; Hasan, A.S. Multi-Features Descriptors for Human Activity Tracking and Recognition in Indoor-Outdoor Environments. In Proceedings of the IEEE International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 8–12 January 2019. [[CrossRef](#)]
49. Nguyen, N.; Yoshitaka, A. Human Interaction Recognition Using Hierarchical Invariant Features. *Int. J. Semant. Comput.* **2015**, *9*, 169–191. [[CrossRef](#)]
50. Slimani, K.N.H.; Benezeth, Y.; Souami, F. Human Interaction Recognition Based on the Co-occurrence of Visual Words. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014. [[CrossRef](#)]

51. Jalal, A.; Mahmood, M. Students' Behavior Mining in E-learning Environment Using Cognitive Processes with Information Technologies. *Educ. Inf. Technol.* **2019**, *24*, 2797–2821. [[CrossRef](#)]
52. Li, M.; Leung, H. Multi-view depth-based pairwise feature learning for person-person interaction recognition. *Multimed. Tools Appl.* **2019**, *78*, 5731–5749. [[CrossRef](#)]
53. Rado, D.; Sankaran, A.; Plasek, J.; Nuckley, D.; Keefe, D.F. A Real-Time Physical Therapy Visualization Strategy to Improve Unsupervised Patient Rehabilitation. *IEEE Trans. Vis. Comput. Graph.* **2009**.
54. Khan, M.H.; Zöller, M.; Farid, M.S.; Grzegorzec, M. Marker-Based Movement Analysis of Human Body Parts in Therapeutic Procedure. *Sensors* **2020**, *20*, 3312. [[CrossRef](#)]
55. Paolini, G.; Peruzzi, A.; Mirelman, A.; Cereatti, A.; Gaukrodger, S.; Hausdorff, J.M.; Della Croce, U. Validation of a method for real time foot position and orientation tracking with Microsoft Kinect technology for use in virtual reality and treadmill based gait training programs. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2014**, *22*, 997–1002. [[CrossRef](#)]
56. Chen, C.C.; Liu, C.Y.; Ciou, S.H.; Chen, S.C.; Chen, Y.L. Digitized Hand Skateboard Based on IR-Camera for Upper Limb Rehabilitation. *J. Med. Syst.* **2017**, *41*, 36. [[CrossRef](#)] [[PubMed](#)]
57. Lapinski, M.; Brum Medeiros, C.; Moxley Scarborough, D.; Berkson, E.; Gill, T.J.; Kepple, T.; Paradiso, J.A. A Wide-Range, Wireless Wearable Inertial Motion Sensing System for Capturing Fast Athletic Biomechanics in Overhead Pitching. *Sensors* **2019**, *19*, 3637. [[CrossRef](#)] [[PubMed](#)]
58. Mokhlespour Esfahani, M.I.; Zobeiri, O.; Moshiri, B.; Narimani, R.; Mehravar, M.; Rashedi, E.; Parnianpour, M. Trunk Motion System (TMS) Using Printed Body Worn Sensor (BWS) via Data Fusion Approach. *Sensors* **2017**, *17*, 112. [[CrossRef](#)] [[PubMed](#)]
59. McGrath, M.J.; Scanaill, C.N. Body-Worn, Ambient, and Consumer Sensing for Health Applications. *Sens. Technol.* **2013**, 181–216. [[CrossRef](#)]
60. Golestani, N.; Moghaddam, M. Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. *Nat. Commun.* **2020**, *11*, 1551. [[CrossRef](#)] [[PubMed](#)]
61. Schlagenhaut, F.; Sahoo, P.P.; Singhose, W. A Comparison of Dual-Kinect and Vicon Tracking of Human Motion for Use in Robotic Motion Programming. *Robot. Autom. Eng. J.* **2017**, *1*, 555558. [[CrossRef](#)]
62. Reining, C.; Niemann, F.; Moya Rueda, F.; Fink, G.A.; Ten Hompel, M. Human Activity Recognition for Production and Logistics—A Systematic Literature Review. *Information* **2019**, *10*, 245. [[CrossRef](#)]
63. Mahmood, M.; Jalal, A.; Evans, H.A. Facial Expression Recognition in Image Sequences Using 1D Transform and Gabor Wavelet Transform. In Proceedings of the International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 4–5 September 2018. [[CrossRef](#)]
64. Jalal, A.; Kamal, S.; Kim, D. Depth Silhouettes Context: A New Robust Feature for Human Tracking and Activity Recognition Based on Embedded HMMs. In Proceedings of the International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Goyang, Korea, 28–30 October 2015. [[CrossRef](#)]
65. Ahmed, A.; Jalal, A.; Kim, K. RGB-D Images for Object Segmentation, Localization and Recognition in Indoor Scenes using Feature Descriptor and Hough Voting. In Proceedings of the International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 14–18 January 2020. [[CrossRef](#)]
66. Rizwan, S.A.; Jalal, A.; Kim, K. An Accurate Facial Expression Detector using Multi-Landmarks Selection and Local Transform Features. In Proceedings of the International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020. [[CrossRef](#)]
67. Ahmed, A.; Jalal, A.; Rafique, A.A. Salient Segmentation Based Object Detection and Recognition Using Hybrid Genetic Transform. In Proceedings of the International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 27–29 August 2019. [[CrossRef](#)]
68. Jalal, A.; Kamal, S.; Kim, D. Shape and Motion Features Approach for Activity Tracking and Recognition from Kinect Video Camera. In Proceedings of the IEEE International Conference on Advanced Information Networking and Applications Workshops, Gwangju, Korea, 24–27 March 2015. [[CrossRef](#)]
69. Hong, F.; Lu, C.; Liu, C.; Liu, R.; Jiang, W.; Ju, W.; Wang, T. PGNet: Pipeline Guidance for Human Key-Point Detection. *Entropy* **2020**, *22*, 369. [[CrossRef](#)]
70. Jalal, A.; Nadeem, A.; Bobasu, S. Human Body Parts Estimation and Detection for Physical Sports Movements. In Proceedings of the International Conference on Communication, Computing and Digital Systems (C-CODE), Islamabad, Pakistan, 6–7 March 2019. [[CrossRef](#)]

71. Jalal, A.; Zia, M.U.; Kim, T. Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Trans. Consum. Electron.* **2012**, *58*, 863–871. [[CrossRef](#)]
72. Firuzi, K.; Vakilian, M.; Phung, B.T.; Blackburn, T.R. Partial Discharges Pattern Recognition of Transformer Defect Model by LBP & HOG Features. *IEEE Trans. Power Deliv.* **2019**, *34*, 542–550. [[CrossRef](#)]
73. Khan, M.H.; Farid, M.S.; Grzegorzec, M.A. Generic codebook based approach for gait recognition. *Multimed. Tools Appl.* **2019**, *78*, 35689–35712. [[CrossRef](#)]
74. Dutta, A.; Ma, O.; Toledo, M.; Buman, M.P.; Bliss, D.W. Comparing Gaussian Mixture Model and Hidden Markov Model to Classify Unique Physical Activities from Accelerometer Sensor Data. In Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016. [[CrossRef](#)]
75. Pan, L.; Wang, D. A Cross-Entropy-Based Admission Control Optimization Approach for Heterogeneous Virtual Machine Placement in Public Clouds. *Entropy* **2016**, *18*, 95. [[CrossRef](#)]
76. Ghohani Arab, H.; Rashki, M.; Rostamian, M.; Ghavidel, A.; Shahraki, H.; Keshtegar, B. Refined first-order reliability method using cross-entropy optimization method. *Eng. Comput.* **2019**, *35*, 1507–1519. [[CrossRef](#)]
77. Wang, H.; Fei, H.; Yu, Q.; Zhao, W.; Yan, J.; Hong, T. A motifs-based Maximum Entropy Markov Model for realtime reliability prediction in System of Systems. *J. Syst. Softw.* **2019**, *151*, 180–193. [[CrossRef](#)]
78. Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T.L.; Samaras, D. Two-Person Interaction Detection Using Body-Pose Features and Multiple Instance Learning. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012. [[CrossRef](#)]
79. Coppola, C.; Faria, D.R.; Nunes, U.; Bellotto, N. Social Activity Recognition Based on Probabilistic Merging of Skeleton Features with Proximity Priors from RGB-D Data. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016. [[CrossRef](#)]
80. Ryoo, M.S.; Aggarwal, J.K. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009. [[CrossRef](#)]
81. Ji, Y.; Cheng, H.; Zheng, Y.; Li, H. Learning contrastive feature distribution model for interaction recognition. *J. Vis. Commun. Image Represent.* **2015**, *33*, 340–349. [[CrossRef](#)]
82. Subetha, T.; Chitrakala, S. Recognition of Human-Human interaction Using CWDTW. In Proceedings of the International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, India, 18–19 March 2016. [[CrossRef](#)]
83. Li, W.; Wen, L.; Chuah, M.C.; Lyu, S. Category-Blind Human Action Recognition: A Practical Recognition System. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [[CrossRef](#)]
84. Huynh-The, T.; Banos, O.; Le, B.-V.; Bui, D.-M.; Lee, S.; Yoon, Y.; Le-Tien, T. PAM-Based Flexible Generative Topic Model for 3D Interactive Activity Recognition. In Proceedings of the International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh, Vietnam, 14–16 October 2015. [[CrossRef](#)]
85. Ji, Y.; Ye, G.; Cheng, H. Interactive Body Part Contrast Mining for Human Interaction Recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 14–18 July 2014. [[CrossRef](#)]
86. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X. Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-16), Beijing, China, 24 March 2016.
87. Manzi, A.; Fiorini, L.; Limosani, R.; Dario, P.; Cavallo, F. Two-person activity recognition using skeleton data. *IET Comput. Vis.* **2018**, *12*, 27–35. [[CrossRef](#)]
88. Coppola, C.; Cosar, S.; Faria, D.R.; Bellotto, N. Automatic Detection of Human Interactions from RGB-D Data for Social Activity Classification. In Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN), Lisbon, Portugal, 28 August–1 September 2017. [[CrossRef](#)]
89. Mahmood, M.; Jalal, A.; Siddiqui, M.A. Robust Spatio-Temporal Features for Human Interaction Recognition Via Artificial Neural Network. In Proceedings of the International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018. [[CrossRef](#)]



90. Amer, M.R.; Todorovic, S. Sum Product Networks for Activity Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 800–813. [[CrossRef](#)]
91. Kong, Y.; Liang, W.; Dong, Z.; Jia, Y. Recognizing human interaction from videos by a discriminative model. *IET Comput. Vis.* **2014**, *8*, 277–286. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).