International Journal of Machine Learning and Cybernetics (2019) 10:2529–2540
https://doi.org/10.1007/s13042-018-0887-5

ORIGINAL ARTICLE

Kinect sensor-based interaction monitoring system using the BLSTM neural network in healthcare

Rajkumar Saini[1] · Pradeep Kumar[1] · Barjinder Kaur[2] · Partha Pratim Roy[1] · Debi Prosad Dogra[3] · K. C. Santosh[4]

Received: 22 January 2018 / Accepted: 3 November 2018 / Published online: 14 November 2018

Abstract

Remote monitoring of patients is considered as one of the reliable alternatives to healthcare solutions for elderly and/or chronically ill patients. Further, monitoring interaction with people plays an important role in diagnosis and in managing patients that are suffering from mental illnesses, such as depression and autism spectrum disorders (ASD). In this paper, we propose the Kinect sensor-based interaction monitoring system between two persons using the Bidirectional long short-term memory neural network (BLSTM-NN). Such model can be adopted for the rehabilitation of people (who may be suffering from ASD and other psychological disorders) by analyzing their activities. Medical professionals and caregivers for diagnosing and remotely monitoring the patients suffering from such psychological disorders can use the system. In our study, ten volunteers were involved to create five interactive groups to perform continuous activities, where the Kinect sensor was used to record data. A set of continuous activities was created using random combinations of 24 isolated activities. 3D skeleton of each user was detected and tracked using the Kinect and modeled using BLSTM-NN. We have used a lexicon by analyzing the constraints while performing continuous activities to improve the performance of the system. We have achieved the maximum accuracy of 70.72%. Our results outperformed the previously reported results and therefore the proposed system can further be used in developing internet of things (IoT) Kinect sensor-based healthcare application.

Keywords Activity recognition · Depth sensors · Bidirectional long short-term memory neural network · Healthcare · Autism spectrum disorders · Internet of things

✉ K. C. Santosh
santosh.kc@usd.edu

Rajkumar Saini
rajkr.dcs2014@iitr.ac.in

Pradeep Kumar
pradeep.iitr7@gmail.com

Barjinder Kaur
kaur.barjinder@gmail.com

Partha Pratim Roy
2partharoy@gmail.com

Debi Prosad Dogra
dpdogra@iitbbs.ac.in

[1] Department of Computer Science and Engineering, IIT Roorkee, Roorkee, India

[2] Department of Computer Science and Engineering, DCRUST, Sonepat, India

[3] School of Electrical Sciences, IIT Bhubaneshwar, Bhubaneshwar, India

[4] Department of Computer Science, University of South Dakota, Vermillion, SD, USA

1 Introduction

Machine learning (ML)-based techniques are developed to automate experts-guided tedious tasks. Further, it has a variety of applications like pattern recognition, computer vision, network security and internet of things (IoT). Recently, ML has been applied to the IoT-based systems, such as healthcare [16]. internet of things (IoT) is considered as a group of interrelated computing devices, mechanical and digital machine that are equipped with unique identifiers and has the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction [3, 7]. Example of such an activity is shown in Fig. 1, where two persons are performing concurrent activities such as dancing, clapping, boxing, walking and hand waving. Besides, ML-based IoT is a novel area having a great impact on society as such techniques have the capability to automatically monitor the performance of the IoT.

Healthcare plays a pivotal role in the effective monitoring of patients requiring constant supervision. Such systems are

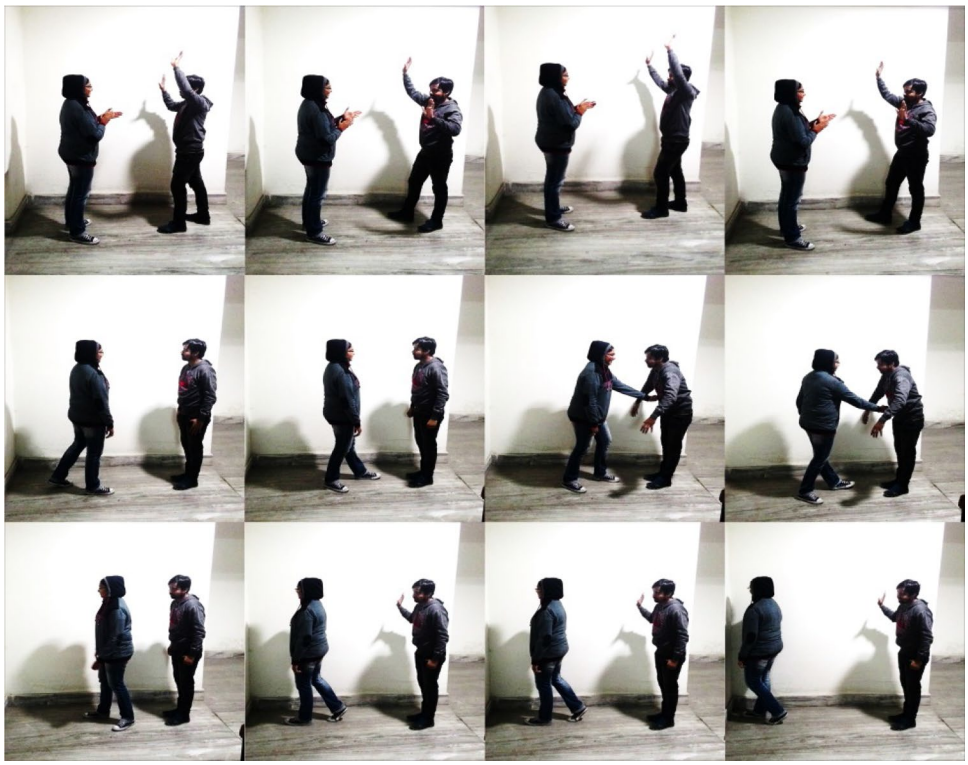

Fig. 1 A scenario of an interaction of 2 persons performing activities

also helpful to provide facilities such as emergency notifications, real-time data, data exchange, live streaming and location information etc. Individuals who suffer from autism spectrum disorder (ASD), stroke or psychological disorders, tend to avoid interactions with other peoples in the society. ASDs are early onset neuro-development disorders characterized by disturbances in social behavior and communication, such as eye contact, intonation, and facial expressions, and the presence of repetitive behaviors and restricted interests [2]. Therefore, to keep track of their progress, doctors always prefer remote monitoring of patients by keeping track of their daily activities or while they interact with others. Therefore, in the IoT, ML may facilitate the doctors in automatic monitoring of their patients. Previous researches have shown that children and adults can improve their skills with computer-based rehabilitation tools and techniques such as gaming, robotics, educational and speech therapy [23]. However, such techniques require extra hardware and installation cost, thus may not always be affordable.

Psychological disorders can severely affect human lives. It can create a significant disturbance in an individual's cognitive, emotion regulation and behavior that reflects a dysfunction in the psychological, biological, or developmental process underlying mental functioning [1]. Several different psychological disorders have been identified and classified, including eating disorders (anorexia nervosa), mood disorders (depression), personality disorders (antisocial personality), psychotic disorders (schizophrenia) and sexual disorders (sexual dysfunction) etc. One person can suffer from multiple psychological disorders. Some of them may require hospitalization and psychological treatment. However, even after discharge, they need continuous medical supervision and support to rehabilitate back to the society as well as to prevent possible recurrences. Monitoring of interaction and participation in activities provides an opportunity to evaluate symptoms and risks of developing psychological disorders. IoT solutions based on wearable sensors and smartphone have been proposed by researchers to monitor and evaluate the mental health [25, 30]. Lanata et al. [21] have proposed a T-shirt with embedded sensors to analyze the mental states of the patients suffering from various mental illness. The sensors need to be connected to a smartphone that collects electrocardiogram heart rate variability (HRV), respiration activity, and movement recognition and it can send to a centralized server for processing. However, wearable sensors are battery powered, thus they require regular maintenance for functioning. Additionally, wearable sensors require the compliance of the person wearing, to ensure that the sensors are used as intended [15].

Development of low-cost depth-sensing technology such as Microsoft Kinect has created ample opportunities for developing human-computer-interaction (HCI) applications and multimedia computing. The sensor is equipped with an RGB camera and an infrared camera as a depth sensor. It is possible to capture both RGB and depth images simultaneously at a frame rate of 30 fps [37]. The depth images are

useful in mitigating some of the inherent problems of RGB images in operations such as background segmentation, object detection, and activity recognition. In addition, the sensor is able to provide a 3D representation of the human skeleton consisting of 20 skeleton joints. In skeletal tracking [9], a human body is represented by a number of 20 joints of the human body including head, neck, shoulders, and arms. Each joint is represented by its 3D coordinates with respect to a fixed position of the sensor. The skeletal tracking of the Kinect sensor has successfully used in various applications including gaming, sign language, human activity [7] and handwriting recognition. The interactive activity recognition is possible due to the availability of such devices. The interaction of two person can be thought as the recognition of their activities e.g. two person eating/drinking together, the fight between them can also be viewed as their interaction as involves the activities like boxing and kicking between them. Yun et al. [36] have proposed a two-person activity recognition system using Kinect. The authors have recognized eight different types of interactions between two persons involving seven users and they have achieved an accuracy of 80.3% using support vector machine (SVM) classifier. However, the authors have not considered independent activities between two persons while interacting with each other.

In this work, we present the Kinect sensor-based two-person interaction monitoring system using the BLSTM neural network in healthcare. The primary motivation is to track the progress of individuals who may be suffering from ASD or other psychological disorders. The Kinect could help analyze the activities of ASD people to recover from such disease. It records the activities of such persons in its field of view. During the interaction, each person is tracked using 3D skeletal and each user is registered using a unique ID for decision making and long-run study to model the activities performed by the couple individually and in response to each other. In short, our study can be summarized as follows:

1. We present a two-person interaction monitoring framework using the Kinect. In our framework, each individual has been tracked using 3D skeletal while performing continuous activities.
2. We used bidirectional long short-term memory neural network (BLSTM-NN) classifier to recognize individual's activities at the time of interaction. In addition, a lexicon approach has been applied that improved the performance.
3. We compared our results with the previous works and observed that our performance can be compared with.

The remainder of the paper is organized as follows. In Sect. 2, we discuss the existing research work in this field of study. The proposed methodology is presented in Sect. 3. Results are presented in Sect. 4. Finally, we conclude in Sect. 5 by discussing future extensions of the presented work.

2 Related works

In this section, we discuss the work of various activity recognition and monitoring systems proposed by different researchers involving 2D images, videos, wearable sensors, smartphones and depth sensors. In [12], the authors have proposed an activity monitoring system for chronically ill patients through the fusion of data captured using Kinect, smartphone's sensors, and social media inputs. The authors have used SVM with the kernel: radial basis function (RBF) to identify individuals using Kinect data, whereas for social media analysis they have used polarity score to get the emotional information about individuals. In [8], the authors have developed an assistive tool for monitoring physical activities of older adults using Kinect. The motion analysis has been performed by analyzing relative joints position, velocity and the angular velocity of the human skeletal. The system is able to recognize seven basic activities. Garcia et al. [11] have proposed a Kinect based system to evaluate the reaction time of stepping tasks. The system was referred to as choice stepping reaction time (CSRT) and used to predict falls among older adults. Their approach is based on measuring physical abilities, such as strength and balance, and cognitive abilities like attention and speed of processing. Similarly, a health monitoring system based on the gait analysis has been proposed in [27]. The authors have used SVM with RBF kernel to perform posture analysis and gait recognition to predict fall detection among senior individuals using Kinect. A continuous activity recognition framework for single person has been proposed in [29]. Authors have used two-stage methodology for the recognition. In the first stage, the activities have been segmented into *sitting*, and *standing* postures followed by final recognition in the second stage. However, no interaction of users have been studied in their work.

Kulkarni et al. [17] have proposed a continuous activity recognition system in videos using dynamic time warping (DTW) algorithm. The authors have used the algorithm for aligning test sequences, whereas the recognition has been performed on the per-frame basis. It has been observed that different individuals perform similar activities with varying speeds resulting variations within intra-activities. An action recognition system using Kinect sensor based 3D skeletal tracking has been presented in [5]. The authors have extracted six different features including pairwise joint position difference, joint velocity, velocity magnitude, joint angle velocity in xy and xz planes, and 3D joint angles between three distinct joints. Recognition rate has been computed in

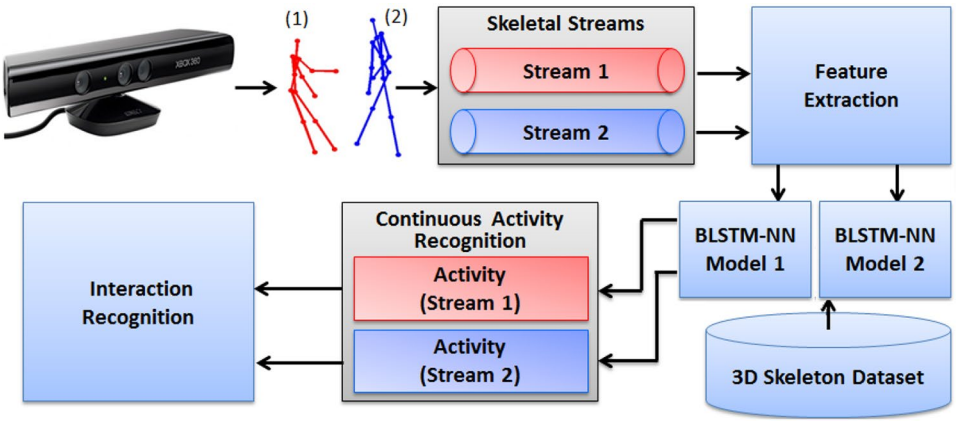

Fig. 2 Flow diagram of the proposed two-person interaction recognition system

terms of performed frame-based actions with F1 scores of 46.66% and 61.17%, respectively. In [31], the authors have utilized 3D representation of various joints of the skeletal as features to recognize activities using Kinect. However, their system has failed to recognize complex human actions when performed by multiple persons. Local occupancy patterns (LOP) based features to represent joint angles has been proposed in [33] by partitioning the 3D point cloud representing joints on a spatial grid. The approach has been tested on two 3D action datasets with recognition rates of 88.2% and 87.75%, respectively.

In addition, sensors including accelerometer, global positioning system (GPS), body sensors, microphone, and smartphone have also been used for monitoring activities of individuals. Ward et al. [35] have proposed activity recognition system using accelerometer and microphone sensors. The authors have performed segmentation of activities by analyzing the sound intensity. Linear discriminant analysis (LDA), and hidden Markov model (HMM) classifiers have been used to perform the recognition of activities on audio and acceleration data, respectively. In [34], the authors have proposed human activity recognition system using wearable sensors. Time-frequency domain features have been extracted from the sensor data and fed into a deep belief network (DBN) for the recognition purpose. In [6], authors have proposed an interaction monitoring architecture called affective interaction through wearable computing and cloud technology (AIWAC) for modeling interaction using wearable sensors and cloud computing. The architecture has been composed of three modules, namely collaborative data collection, sentiment analysis and emotional interaction using wearable sensors, facial videos, and social network analysis. Murali et al. [26] have developed a wearable device for physiological and psychological health monitoring of individuals. Physiological signals such as ECG, respiration, impedance cardiogram (ICG), blood pressure and skin conductance have been collected using the device and the Valence-Arousal model has been used to classify various emotional states of the users. The processed data has been transmitted to the users mobile phone using Bluetooth connectivity.

Authors of [16] proposed a self-diagnosis and awareness system using two circular layers. The inner layer is disease awareness layer aware of user's state giving a possible diagnosis alert. The outer layer is health awareness layer for motivating the patients so that he/she can recover from the disease. Deep networks has also been exploited for activity recognition [4, 24]. Authors of [24] has used CNN for anomaly detection in Hockey Videos. Similarly, Authors of [4] has used 3D CNNs to extract spatio temporal features to recognize human actions in videos. Such ontologies may be implemented using ML models discussed above to develop IoT-based system to monitor patients.

3 Proposed system

In this section, we present the details of the proposed Kinect based two person interaction monitoring framework.

For data acquisition, we have used Xbox 360 Software Development Kit (SDK) that provides 3D skeletal tracking of humans appearing in the device's viewing field. The skeletal view consists of twenty 3D joints computed with the help of its depth-sensing camera. Using this, we have tracked the 3D skeletal view of two persons while they interact with each other by performing different activities. The proposed work flow is depicted in Fig. 2, where two persons are interacting with each other. A unique identification (ID) number was allocated to each skeletal structure during the interaction. Using this, two different channels of data were processed, where information about the 3D joints of each user along with their unique IDs were recorded in the dataset. With unique IDs, we extracted features, such as joint positions, distance between the joints and, angle between two vectors (using joints). Using such features set, BLSTM-NN classifier is then trained and tested to recognize activities.

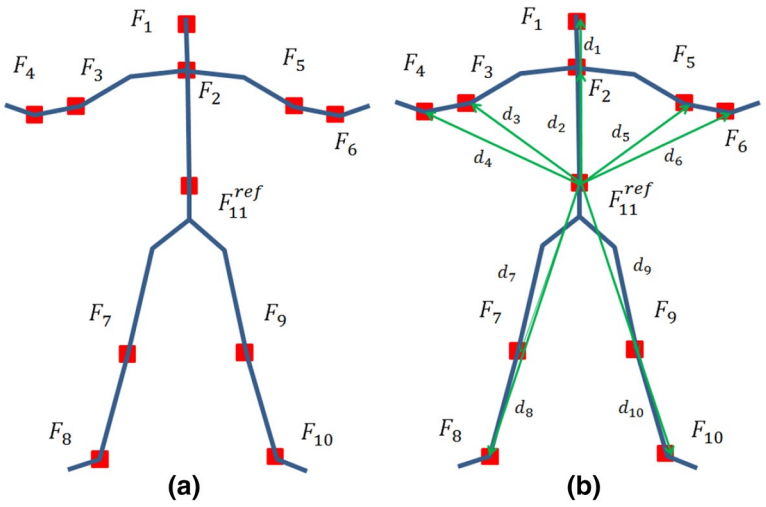

Fig. 3 Feature computation: **a** selected joints position **b** distance features with respect to the reference point F^{ref}

3.1 Feature extraction

For feature extraction, we recorded instantaneous positions of all twenty 3D joints of the human skeleton. In this work, four different features have been extracted, namely

- joint position (F_j),
- angular vector (F_a),
- velocity vector (F_v) and
- distance (F_d),

and they all form a set, $\mathcal{F}_t = \{F_j, F_d, F_a, F_v\}$.

3.1.1 Joint position feature (F_j)

We recorded all twenty joint positions of the skeletal structure. Inspired from previous work [10], since some of them repeated similar information, i.e. redundancies, we have excluded them from our set. As a consequence, we have a set of 11 joints including a reference point F^{ref} that represents the '*Spine*' location in the 3D skeletal, as shown in Fig. 3. Altogether, these joints collectively form a 33-dimensional feature vector (F_j) as mentioned before, and $F_j = \{F_1, F_2, \ldots, F_{11}\}$.

3.1.2 Distance feature (F_d)

It represents the distance of 3D joints position from the reference point F^{ref}. It has been computed as d_i ($d_i = \sqrt{(x_r^2 - x_i)^2 + (y_r^2 - y_i)^2 + (z_r^2 - z_i)^2}$) between ith joint (x_i, y_i, z_i) and the reference point (x_r, y_r, z_r). Using this, we have computed 10 distance features as shown in Fig. 3b. Altogether, these features collectively form a 10-dimensional feature vector F_d as ($F_d = \{d_1, d_2, \ldots, d_{10}\}$).

3.1.3 Angular feature (F_a)

Angular features (F_a) have been computed with the help of vectors drawn between the reference point and the selected joints. For each vector, we have computed three different angles, namely $(\theta_x^i, \theta_y^i, \theta_z^i)$ with the coordinate axes for the ith joint. A pictorial representation of angular feature for the joint F_j^5 is depicted in Fig. 4. Using this, 10 3D features have been computed that collectively represent as F_a.

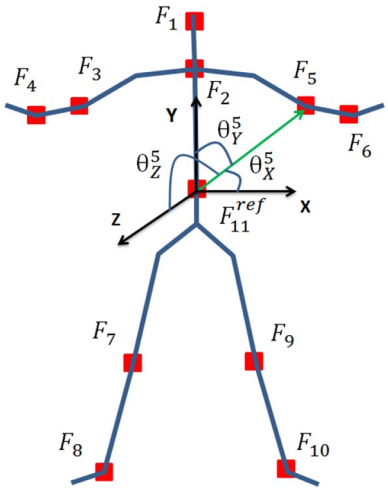

Fig. 4 Computation of angular feature for the joint F_j^5 with respect to the reference point F^{ref}

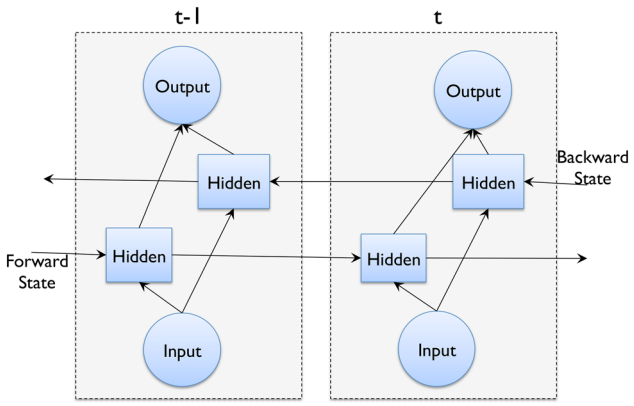

Fig. 5 Basic bidirectional RNN block: communication (both directions) using separate hidden layers at time $t-1$ and t

3.1.4 Velocity feature (F_v)

Different activities have been performed with varying speeds. For example, an activity '*walking*' is generally performed with slower speed in comparison with the activity '*running*'. Therefore, velocity features have been computed to distinguish such activities during testing. It has been done by computing the distance between two successive points of the trajectory, such as (x_i, y_i, z_i) and $(x_{i+1}, y_{i+1}, z_{i+1})$. Using this, we have computed the velocity of each joint sequence that belongs to the feature vector F_j. We get a the feature vector F_v.

3.2 Continuous activity recognition

BLSTM-NN classifier is a bidirectional recurrent neural network (RNN) that is widely used by researchers in modeling temporal sequences for hand-writing [13] and gesture recognition problems [22]. The classifier has the capability to process an input sequence in both directions, i.e. forward and backward with the help of its two hidden layers. Both the hidden layers share a common output layer that contains a node for each activity. Along with one node per activity the output layer has a special node denoted by ϵ. This special node is used to indicate '$no-activity$' to mean no decision has been made at that position. The connectionist temporal classification (CTC) objective function (O) as defined in (1), where T is the training set and the pair (p, q) denotes the input and target sequence used in our implementation,

$$O = -\ln\left(\prod_{(p,q)\in T} p(q|p)\right) = -\sum_{(p,q)\in T} \ln\left(p(q|p)\right). \tag{1}$$

The objective function O models the label sequence with the given inputs. With the help of long short-term memory (LSTM) hidden layers memory blocks, the model successfully resolves the problem of vanishing gradient, where the influence of the given input sequence decays the output exponentially over time. The memory blocks allow the units to store and access the information for a longer time [13, 14]. Figure 5 shows the basic bidirectional-RNN that processes the input in both the directions, i.e. forward and backward at time $t-1$ and t. It has separate hidden layers for both directions. Finally, Softmax layer is used to estimate the scores for each activity class which is 24 in our case.

Table 1 Description of the activities performed during data collection

Boxing	Eating	Bending
Dancing	Read sitting	Clapping
Jumping	Sit still	Hand wave
Kicking	Sitting	Phone call
Read standing	Typing	Paper toss
Running	Write sitting	Push/pull
Standing	Walking	Thinking
Stand still	Write standing	Drinking

4 Results

4.1 Dataset

In this work, 10 volunteers were involved in data collection. Twenty four basic activities were selected since these activities are usually performed by users in daily life. The dataset consists of continuous activity sequences as 3D skeleton made up of 20 body joints, where each joint itself is a 3D point in 3-dimensional euclidean space. Thus, each frame in an activity sequence is of 60-dimensions. The dataset has been made available for further research[1]. A description of all the selected activities is shown in Table 1.

Using the different combinations of 24 activities, a total of 111 activity sequences were prepared. The minimum and maximum length of any continuous activity were kept between two and six, respectively. A distribution of sequence lengths as per the number of activities involved is shown in Fig. 6, where 44.14% of the sequences consist of combinations of 4 different activities. As per the proposed framework, interactions between various pairs were recorded while performing different activities with each other. Five pairs from the 10 volunteers recorded such data. Using this, a total of 1110, i.e. 111×10, continuous activity sequences were recorded. The pictorial representation of example

[1] https://drive.google.com/open?id=1KKTMryAzI-G8cwc3v7HvueH6Ufga4qx6.

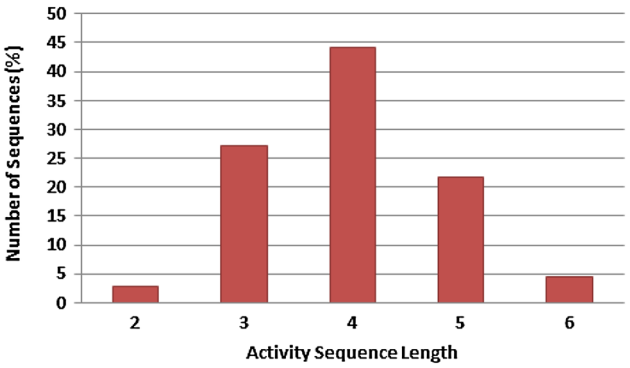

Fig. 6 Distribution of the number of activities involved in continuous activity sequences

a pair of users have been selected and tested in parallel on the classifier as per their skeletal ID. The network has been trained with a learning rate of $1e-4$, momentum of 0.9, and initialized with a Gaussian distribution of zero mean and a standard deviation of 0.1. In the 3D test sequence, activity based recognition has been performed using CTC network without providing lexicon. The learning rate of the network is shown in Fig. 9. The learning curve shows the error variation in training and validation data. After 163 epochs, the variations in the validation network becomes constant, thus marked with a dashed vertical line in the figure.

Performance of each of the five interactive groups has been computed separately by testing each pair while keeping rest of the four groups in training (Fig. 10), where an average accuracy of 58.50% has been recorded. Likewise,

Fig. 7 Pictorial representation of few example activities involved in the dataset

activities is depicted in Fig. 7, where a user is performing different activities. To show the variations among activities, two instances of each activity are shown in Fig. 7, where the first instance depicts the start posture and another one shows the end posture of the user while performing the activity. Similarly, the interactions between these two persons are shown in Fig. 8, where both users are performing a set of different activities. The experiments have been conducted in fivefolds cross validation protocol, where, training has been done using the data of 4-user groups and the other pair has been used for testing.

4.2 Two-person interaction recognition

The interaction between each pair of persons has been recognized using the BLSTM-NN classifier. The classifier has been trained on the collected dataset. In the testing phase, the activity recognition rates have also been computed for each of the involved participants. The recognition results are depicted in Fig. 11, where the performance of four users has been recorded below the set average.

4.3 Continuous action constraint

In continuous activity recognition during the interaction, the transition between different activities is often constrained by various factors. Consider an example of a person who is sitting on a chair and starts walking. One logical evaluation of such an event may result in a series of three activities, i.e. '*sit* − *still*', '*standing*' and '*walking*'. It is practically not possible to observe '*walking*' without observing '*standing*' first. Hence, in this case '*standing*' constitutes a transition constraint between '*sit* − *still*' and

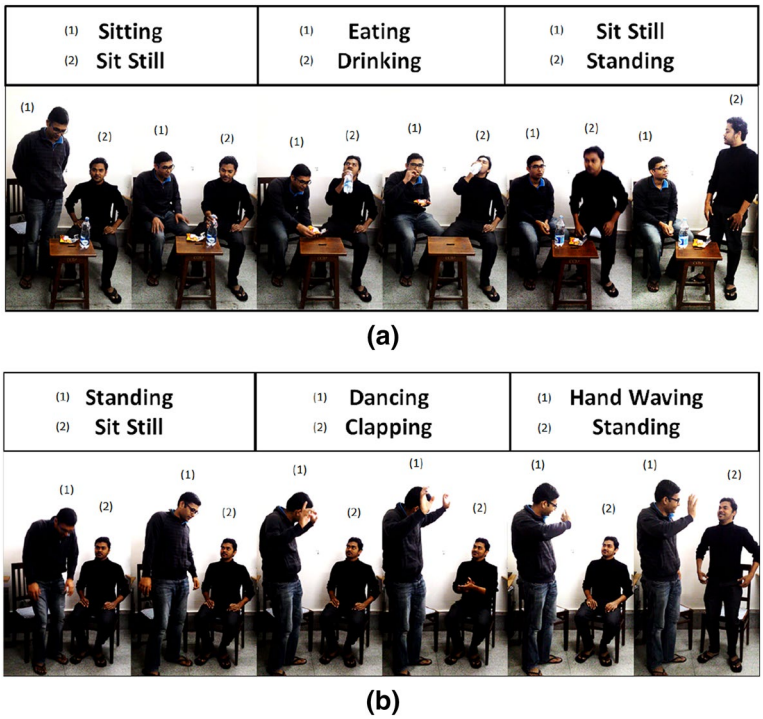

Fig. 8 Examples of two-person interaction while performing continuous activities during data collection. Each user is detected using 3D skeletal ID [marked as (1) and (2)] and the description of activities is also provided

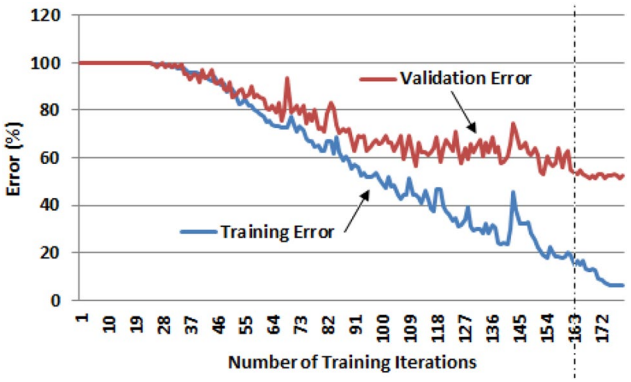

Fig. 9 Learning curve of the network during training

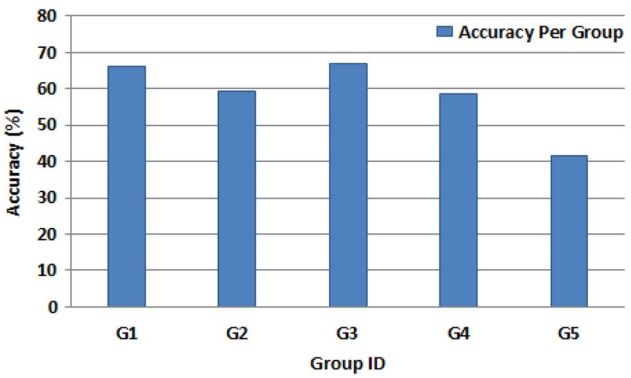

Fig. 10 Continuous activity recognition results during the interaction of each group

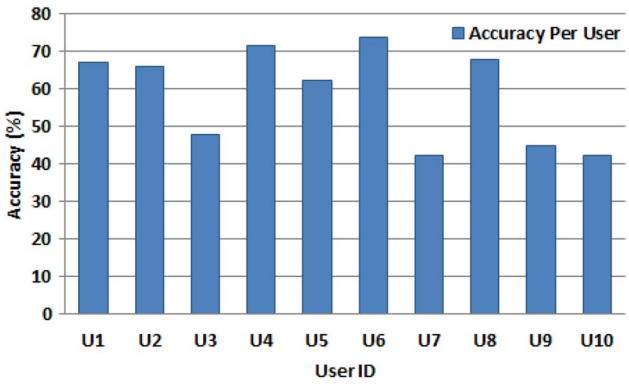

Fig. 11 Continuous activity recognition performance of each user

'*walking*'. However, sometimes the framework recognized the sequence as '*sit* – *still*', '*walking*'. A pictorial representation of this transition constraint is depicted in Fig. 12, where transitions between multiple activities are shown by double-headed arrows. In order to effectively incorporate such transitions, we have prepared a lexicon of activities that helps in improving the recognition performance. BLTSTM takes the output and the lexicon and uses the EDIT distance to predict the final recognition. The activity recognition rates have been computed for each interactive group using lexicon-based approach of BLSTM-NN as shown in Fig. 14, where the lexicon based accuracies of each group are better than without lexicon based performance.

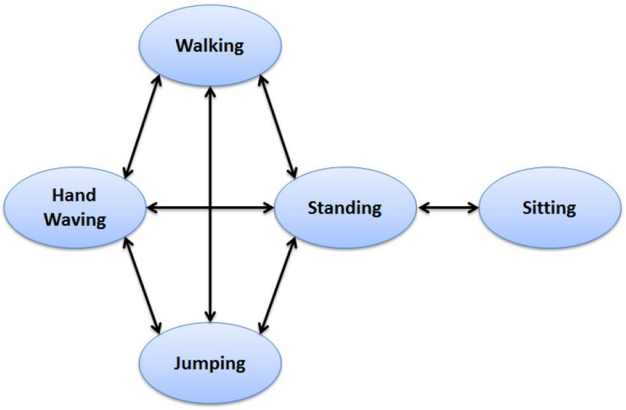

Fig. 12 An example of constraint while performing continuous activity

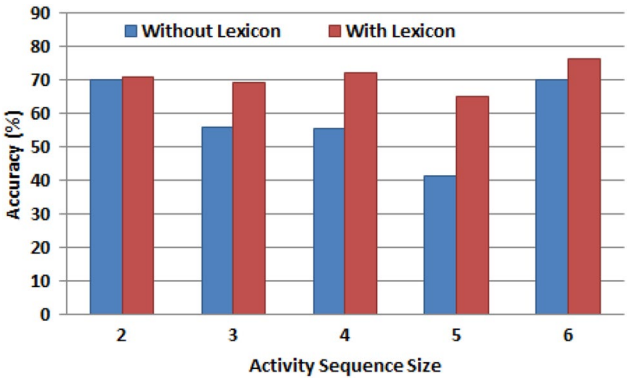

Fig. 13 Performance comparison among continuous activities having different sequence length when tested with and without lexicon-based approach

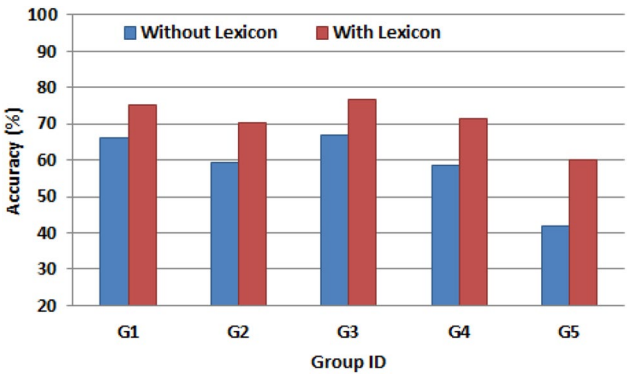

Fig. 14 Performance comparison of each interactive group when tested with and without lexicon-based approach

Distribution of activities in continuous activity sequences varies from 2 to 6 activities. Therefore, the recognition of activities has been carried out as per different sequence lengths of continuous activities using both approaches, i.e. with and without lexicon. The accuracies of both the approaches are depicted in Fig. 13, where maximum accuracies have been recorded with the sequences with two and six activities. On the other hand, sequences that consist five activities have minimum recognition rates.

The recognition rates for each interactive group have been evaluated with the lexicon as depicted in Fig. 14. As compared to without lexicon approach, recognition of each group is better using lexicon based classification.

Performance of the proposed system with and without lexicon has also been evaluated. The system has an accuracy of 58.51% without lexicon. The performance increases to 70.72% with a margin of 12.21% when tested with lexicon. The recognition rates for different top choices have also been evaluated as depicted in Fig. 15, where the maximum performance was recorded at top 5 choices with an accuracy of 89.90%, whereas the accuracy at top 1 choice was recorded as 70.72%.

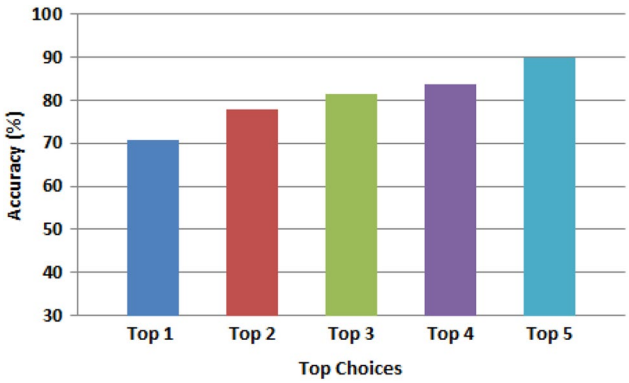

Fig. 15 Activity recognition rates for different top choices

4.4 Error analysis

In this section, we present the details of the continuous activity sequences that are not correctly recognized by the proposed framework during an interaction. The analysis is performed by referencing the Fig. 11 that depicts the recognition results as per the involved participants. It can be seen from the figure that, activity recognition rates of two users are below average accuracy. It is because these users have performed certain activities very fast which results in less number of recorded frames. For example, activities such as '*kicking*', '*boxing*' and '*toss – paper*' are performed in lesser time, therefore, results into confusion during recognition process and ultimately leads to the lower performance of the overall system.

4.5 Comparative analysis

The comparative analysis is performed to compute the two-person interaction recognition using other sequential classifier. For this, we have used left-to-right HMM classifier [28] which is widely used by researcher in activity and hand-writing recognition problems [18, 20]. The model (δ) can be described as,

$$\delta = (H, O, A, B, \pi), \tag{2}$$

where H is the number of hidden states, O are the observations sets of length m, A and B are the state transition and emission matrices. The term π is the initial state distribution matrix for class c ($c = 1, 2, 3 \ldots C$) and $\sum_{c=1}^{C} \pi = 1$. For modeling the observation sequence (O) to one of the class $c_1, c_2, \ldots, C$, the posterior probability $P(\delta_i|O)$ is estimated as,

$$C^* = \arg\max_{c} P(\delta_c|O). \tag{3}$$

We have compared the proposed method on the datasets namely, KARD [10], and CAD-60 [32]. The KARD dataset consists off 18 activities like *kicking, drinking, phone call, etc*. The authors have used k-means clustering, SVM, and HMM models for the final prediction. K-means algorithm has been used to cluster the postures during the activities. SVM has been used to validate the recognition of postures. Finally, the activity recognition has been modeled as a sequences of postures and recognized using the HMM classifier. The recognition rate as 95% has been recorded by the authors on the KARD dataset. The CAD-60 dataset consists of 12 daily life activities like *rinsing mouth, brushing teeth, etc*. The authors have used maximum entropy based models to recognize the activities. The body pose, hand position like features have been extracted in their study. The recognition rate of 76.67% have been recorded using their approach. We have compared the proposed approach with the methods in [10, 32]. First, we have extracted the features (as discussed in Sect. 3.1) from the tracked skeleton data and trained the HMM, and BLSTM classifiers. For fare comparison, we have kept the evaluation protocol same as the authors used in their work. The average accuracies as 95.56%, and 96.67% has been recorded using HMM, and BLSTM classifier, respectively using KARD dataset. Whereas, the average accuracies as 78.42%, and 79.58% has been recorded using HMM, and BLSTM classifier, respectively using the CAD-60 dataset. The confusion matrix using the BLSTM classifier is shown in the Fig. 16. It can be noticed that the proposed method performs well as compared to [10, 32]. We have performed statistical significance test (t-test) between the with and without lexicon based recognition results. The t-test [19] is performed to ensure if the mean values of two

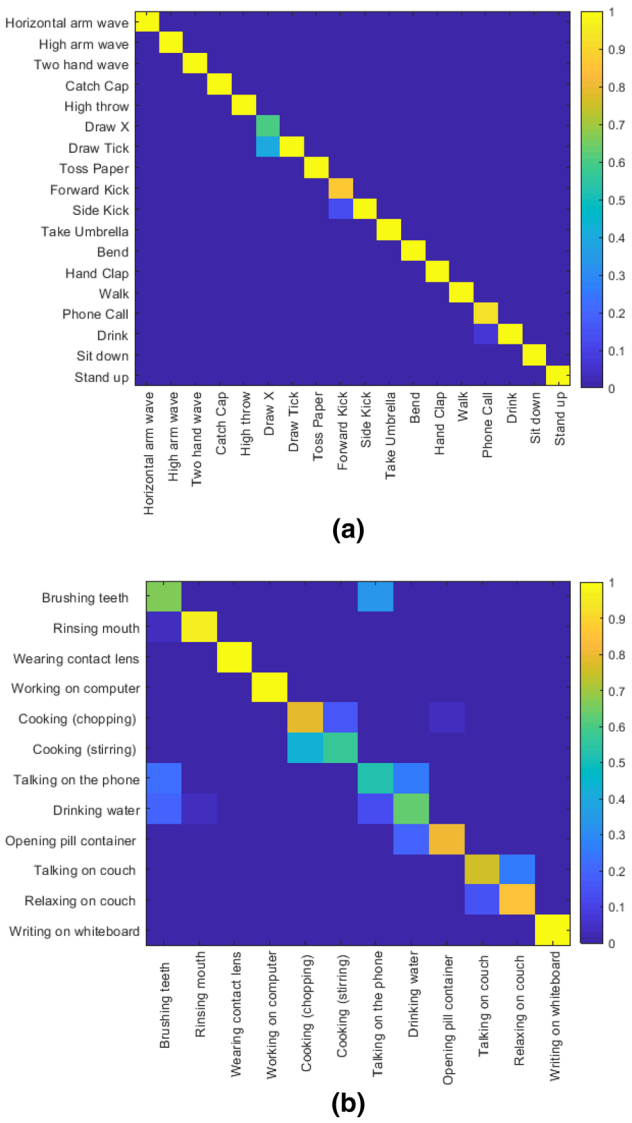

Fig. 16 Confusion matrices for **a** KARD, and **b** CAD-60 datasets using BLSTM classifier

sets significantly differ. We have noticed the p-value is less than 0.037.

5 Conclusion

The proposed system covers a wide variety of applications including healthcare, surveillance, and security. In healthcare, it allows a medical professional to monitor the interaction of their patients with others. Since, patients suffering from ASD or psychological disorders generally avoid social interaction. Therefore, patients may be shifted from doctor-centric care to patient-centric environment allowing medical professionals to keep track their patients remotely. Therefore, we have proposed a two-person interaction monitoring

system using the Kinect sensor. The system is able to recognize continuous activities performed by two persons while interacting with each other. A dataset of 10 users in 5 groups has been recorded while they have been interacting with each other by performing sequences of continuous activities. During an interaction, each individual has been detected with a unique skeletal ID. Recognition of activities has been performed in parallel using BLSTM-NN classifier. The overall performance of the system has been recorded as 70.72% when tested with the lexicon-based approach using BLSTM-NN classifier. We believe, recognition can be improved by exploring novel features and multi-classifier fusion based approaches.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institution.

References

1. A P Association et al (1994) Diagnostic and statistical manual of mental disorders (dsm). American psychiatric association, Washington, DC, pp 143–147
2. A P Association et al (2000) Diagnostic and statistical manual of mental disorders, revised, vol 943. American Psychiatric Association Washington DC, p 2000
3. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54(15):2787–2805
4. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. In: Human Behavior Understanding, pp 29–39
5. Bloom V, Makris D, Argyriou V (2012) G3d: a gaming action dataset and real time action recognition evaluation framework. In: Conference on computer vision and pattern recognition workshops, pp 7–12
6. Chen M, Zhang Y, Li Y, Hassan MM, Alamri A (2015) Aiwac: affective interaction through wearable computing and cloud technology. IEEE Wirel Commun 22(1):20–27
7. Cheok MJ, Omar Z, Jaward MH (2017) A review of hand gesture and sign language recognition techniques. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-017-0705-5
8. Dell'Acqua P, Klompstra LV, Jaarsma T, Samini A (2013) An assistive tool for monitoring physical activities in older adults. In: 2nd international conference on serious games and applications for health, pp 1–6
9. Feng S, Murray-Smith R, Ramsay A (2017) Position stabilisation and lag reduction with gaussian processes in sensor fusion system for user performance improvement. Int J Mach Learn Cybern 8(4):1167–1184
10. Gaglio S, Re GL, Morana M (2015) Human activity recognition process using 3-d posture data. IEEE Trans Hum Mach Syst 45(5):586–597
11. Garcia JA, Pisan Y, Tan CT, Navarro KF (2014) Assessing the kinects capabilities to perform a time-based clinical test for fall risk assessment in older people. In: International conference on entertainment computing, pp 100–107
12. Ghose A, Sinha P, Bhaumik C, Sinha A, Agrawal A, Dutta Choudhury A (2013) Ubiheld: ubiquitous healthcare monitoring system for elderly and chronic patients. In: Conference on pervasive and ubiquitous computing adjunct publication, pp 1255–1264
13. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J (2009) A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 31(5):855–868
14. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5):602–610
15. Holzinger A, Röcker C, Ziefle M (2015) Smart health: open problems and future challenges, vol 8700. Springer, Switzerland
16. Ji J, Scholten P, Zhao Q (2014) Support to self-diagnosis with awareness. Int J Mach Learn Cybern 5(4):647–658
17. Kulkarni K, Evangelidis G, Cech J, Horaud R (2015) Continuous action recognition based on sequence alignment. Int J Comput Vis 112(1):90–114
18. Kumar P, Gauba H, Roy PP, Dogra DP (2016) Coupled hmm-based multi-sensor data fusion for sign language recognition. Pattern Recognit Lett 86:1–8. https://doi.org/10.1016/j.patrec.2016.12.004
19. Kumar P, Saini R, Roy P, Dogra D (2017) A bio-signal based framework to secure mobile devices. J Netw Comput Appl 89:62–71
20. Kumar P, Saini R, Roy PP, Dogra DP (2016) 3d text segmentation and recognition using leap motion. In: Multimedia tools and applications, pp 1–20
21. Lanata A, Valenza G, Nardelli M, Gentili C, Scilingo EP (2015) Complexity index from a personalized wearable monitoring system for assessing remission in mental health. IEEE J Biomed Health Inf 19(1):132–139
22. Lefebvre G, Berlemont S, Mamalet F, Garcia C (2013) Blstm-rnn based 3d gesture classification. In: International conference on artificial neural networks, pp 381–388
23. Miranda JC, Sousa AA, Fernandes T, Orvalho VC (2011) Interactive technology: teaching people with autism to recognize facial emotions. In: Autism spectrum disorders—from genes to environment. InTech
24. Mukherjee S, Saini R, Kumar P, Roy PP, Dogra DP, Kim BG (2017) Fight detection in hockey videos using deep network. J Multimed Inf Syst 4(4):225–232
25. Mukhopadhyay SC (2015) Wearable sensors for human activity monitoring: a review. IEEE Sens J 15(3):1321–1330
26. Murali S, Rincon F, Atienza D (2015) A wearable device for physical and emotional health monitoring. In:Computing in Cardiology Conference, pages 121–124
27. Parajuli M, TranD, Ma W, Sharma D (2012) Senior health monitoring using kinect. In: 4th international conference on communications and electronics, pp 309–312
28. Rabiner L, Juang B (1986) An introduction to hidden markov models. IEEE ASSP Mag 3(1):4–16
29. Saini R, Kumar P, Roy PP, Dogra DP (2018) A novel framework of continuous human-activity recognition using kinect. Neurocomputing 311:99–111
30. Sebestyen G, Hangan A, Oniga S, Gál Z (2014) ehealth solutions in the context of internet of things. In: International conference automation, quality and testing, robotics, pp 261–267
31. Sempena S, Maulidevi NU, Aryan PR (2011) Human action recognition using dynamic time warping. In: International conference on electrical engineering and informatics, pp 1–5

32. Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from rgbd images. In: 2012 IEEE international conference on robotics and automation, pp 842–849
33. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Conference on computer vision and pattern recognition, pp 1290–1297
34. Wang L (2016) Recognition of human activities using continuous autoencoders with wearable sensors. Sensors 16(2):189
35. Ward JA, Lukowicz P, Troster G, Starner TE (2006) Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans Pattern Anal Mach Intell 28(10):1553–1567
36. Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D(2012) Two-person interaction detection using body-pose features and multiple instance learning. In: Conference on computer vision and pattern recognition workshops, pp 28–35
37. Zhang Z (2012) Microsoft kinect sensor and its effect. IEEE Multimed 19(2):4–10