J Ambient Intell Human Comput (2017) 8:957–976
DOI 10.1007/s12652-016-0415-y

ORIGINAL RESEARCH

Feature-level combination of skeleton joints and body parts for accurate aggressive and agitated behavior recognition

Belkacem Chikhaoui[1] · Bing Ye[1] · Alex Mihailidis[1]

Received: 7 June 2016 / Accepted: 20 September 2016 / Published online: 4 October 2016

Abstract This paper presents a novel and practical approach for aggressive and agitated behavior recognition using skeleton data. Our approach is based on feature-level combination of joint-based features and body part-based features. To characterize spatiotemporal information, our approach extracts first meaningful joint-based features by computing pairwise distances of skeleton 3D joint positions at each time frame. Then, distances between body parts as well as joint angles are computed to incorporate body part features. These features are then effectively combined using an ensemble learning method based on rotation forests. A singular value decomposition method is used for feature selection and dimensionality reduction. The proposed approach is validated using extensive experiments on variety of challenging 3D action datasets for human behavior recognition. We empirically demonstrate that our proposed approach accurately discriminates between behaviors and performs better than several state of the art algorithms.

1 Introduction

Globally we are facing a healthcare crisis related to caring for a rapidly aging population who are suffering from a variety of chronic medical conditions, such as dementia. Caring for people with dementia is more complicated given the severity of dementia they suffer from and the degree of autonomy they need for the completion of their activities of daily living (Mihailidis et al. 2008). Challenging behaviors, such as agitation and aggression, are very common in people with dementia and regarded as part of behavioral and psychological symptoms of dementia (BPSD) (Desai and Grossberg 2001). Agitation consists of an unusual state of motor or verbal activity that could be shown by some of the following symptoms such as repetitive walking, wandering, pacing or restlessness, frequent requests for attention or reassurance, frustration, anger or irritability, screaming, cursing, and refusal to allow care to be performed. Whereas aggression is when the behaviors are taken to a more physical point and can be demonstrated by behaviors such as verbal or physical threats, kicking and punching, tearing things, and violent reactions (Mallidou et al. 2013).

These challenging behaviors can cause great suffering for persons with dementia, premature institutionalization, and could result in staggering health care costs, significant loss of quality of life, and a great deal of distress and burden for caregivers (Moore et al. 2013). In addition, (Tampi et al. 2011) reported that these challenging behaviors add significantly to the direct and indirect costs of care. For example, according to (Beeri et al. 2002), the annual indirect cost of managing these challenging behaviors in a patient with Alzheimer's Disease (AD) was about $2665 US, which was 25 % over the total annual indirect cost of caring for a patient with AD ($10,350 US).

✉ Belkacem Chikhaoui
belkacem.chikhaoui@utoronto.ca

Bing Ye
bing.ye@uhn.ca

Alex Mihailidis
alex.mihailidis@utoronto.ca

[1] IATSL Laboratory, Toronto Rehab Institute, University of Toronto, Toronto, Canada

In addition, the annual direct cost of these challenging behaviors was about $1450 US, which was 35 % over the total annual direct cost of caring for a patient with AD ($3900 US) (Beeri et al. 2002). Therefore, early detection and recognition of these challenging behaviors can help effectively provide better treatment for persons with dementia, which in turn will help reduce caregiver's burden (Desai and Grossberg 2001) and reduce significantly health care costs.

Understanding aggressive and agitated behaviors of persons with dementia is usually difficult. These behaviors are usually a result of the disease and are not intentional (Gray 2004). Therefore, people with dementia would not be able to give reasons for their behaviors (Gray 2004). Direct observation from family caregivers and the care staff is usually used to identify challenging behaviors. However, this method is subjective, time consuming and could increase the workload of care staff and caregivers (Desai and Grossberg 2001). Therefore, researchers have focused on developing intelligent systems to automatically monitor and recognize aggression and agitation (Qiang et al. 2007) as not only will technology reduce the manpower and time needed to observe and detect these behaviors (Fook et al. 2007), it will also have the potential to give reliable and consistent results (Ya-Xuan et al. 2010; Mori et al. 2007; Duong et al. 2005) on predictors of these behaviors.

Much research has been conducted on human behavior recognition (Aggarwal and Cai 1999; Bouziane et al. 2013; Sheng et al. 2015; Zhu et al. 2013; Guo 2011), however, little work has been done on automatic recognition of agitated and aggressive behaviors in people with dementia. Therefore, the motivations for our current work can be summarized in the following points:

- The little work on automatic agitation and aggression recognition,
- The goal of decreasing the suffering of persons with dementia and increasing their quality of life,
- The goal of reducing caregivers' burden and related care costs.

Various types of sensors have been used for human behavior recognition such as cameras (Fook et al. 2007), the Microsoft Kinect (Osunkoya and Chern 2013), accelerometers (Benayed et al. 2014), and multimodal sensors such as motion sensors, acoustic sensors, RFID sensors and pressure sensors (Qiang et al. 2007). However, particular attention has been devoted recently to the use of Kinect sensor given the rich information it provides of a person's behaviors when compared to other sensors (van Teijlingen et al. 2012). Kinect, which is a vision sensor, allows collecting different types of data such as individual movements, physical and verbal behaviors using different data formats such as skeleton data, depth data and color data. Kinect sensor has been gaining momentum in different domains to monitor people behaviors. They have been used in video surveillance (Benayed et al. 2014), human-computer interaction (Osunkoya and Chern 2013), and health and medicine (Gantenbein 2012).

In this paper, we propose an effective approach for aggressive and agitated behavior recognition using skeleton data collected from a Kinect sensor. Our approach combines joint-based features and body part-based features using an ensemble learning method based on rotation forests. The combination of these features leads to a significant improvement in the recognition of aggressive and agitated behaviors as compared to the state of the art approaches. The novelty of our approach can be justified by the fact that, our approach combines feature selection and ensemble learning to achieve a good recognition accuracy while using only a small number of features compared to existing methods. The major contributions of this paper can be summarized as follows:

1. Combine joint-based features and body part-based features in a unified approach for aggressive and agitated behavior recognition.
2. Incorporate feature selection with ensemble learning method based on rotation forests in order to reduce dimensionality and improve the recognition accuracy.
3. Conduct extensive experiments over a variety of 3D action datasets to validate our proposed approach.

The rest of the paper is organized as follows. First, we give an overview of related work in Sect. 2. Section 3 describes the proposed approach in terms of features extraction, learning and recognition using rotation forests method. The results of our experiments on real 3D action datasets are presented in Sect. 4. Finally, Sect. 5 presents our conclusions and highlights future work directions.

2 Related work

Human action recognition was the subject of several research studies over the last two decades using different data inputs such as sensor data, images, depth data, and skeleton data (Aggarwal and Ryoo 2011; Shotton et al. 2011; Oreifej and Liu 2013; Wang et al. 2012a; Yang et al. 2012; Kläser et al. 2008; Luo et al. 2013; Lu et al. 2014; Hussein et al. 2013; Ohn-Bar and Trivedi 2013; Wang et al. 2012b; Roy et al. 2016; Nazerfard and Cook 2015; Bouchard et al. 2014; Zhan and Kuroda 2014; Maleki-Dizaji et al. 2014; Andreu and Angelov 2013; Chikhaoui et al. 2012, 2014). We refer the readers to (Aggarwal and Ryoo 2011) for a review of RGB video-based approaches, and (Chen et al. 2013; Ye et al. 2013) for a recent review of depth map-based approaches. In this section, we briefly review the research that has been done on aggressive and agitated behavior recognition.

Much work has been done on understanding and managing aggressive and agitated behaviors specifically for older adults with dementia (Ashok Krishnamoorthy 2011; Desai and Grossberg 2001). However, Only a few studies have focused on using intelligent systems to detect aggression and agitation in persons with dementia. In Bankole et al.'s work (Bankole et al. 2012), the authors investigated body sensor network technology in the detection of agitation in older adults with dementia. The authors compared observed agitated behaviors with body sensor based recorded behaviors, and found a correlation between the observed behaviors and body sensor recorded behaviors. However, in their study, participants were required different body sensors which are obtrusive. Rajasekaran et al. (2011) proposed a wearable device for early detection of anxiety and agitation in people with cognitive impairment. Thomas et al. (Plötz et al. 2012) proposed a system based on machine learning techniques to segment relevant behavioral episodes from a continuous wearable sensor stream and to classify them into distinct categories of severe behavior such as aggression, disruption, and self-injury. The system was validated using simulated data of episodes of severe behavior acted out by trained specialists, and other daily living activities available datasets. However, all these studies looked at physiological data to detect agitated and aggressive behaviors, which required specific sensors for physiological data collection. Our approach differs from the above studies in the following ways. First, in our approach participants are not required to wear any device for data collection. Second, our approach relies only on skeleton data without the need for physiological data. This makes our approach more suitable for real world applications.

In Biswas et al. (2006) work, multi-modal sensors were used to monitor agitation in people with dementia. The agitation was detected and monitored by the sensors based on the intensity of the movements such as sitting and standing. However, the authors consider only limited movements such as sitting and standing. In another agitation detection study, researchers used a video camera-based method to recognize agitated behaviors (Fook et al. 2007). The recorded video data were then annotated based on the gold standard agitation assessment tool to classify agitated behaviors and non-agitated behaviors. Skin color segmentation techniques were used in order to analyse video data and extract relevant features describing agitated behaviors. However, this technique present some limitations in terms of the difficulty in detecting the skin regions during the night and when the person is not facing the camera, which could affect the feature extraction. Nirjon et al. (2013) proposed a system to detect aggressive actions such as hitting, kicking, pushing, and throwing from streaming 3D skeleton joint coordinates obtained from Kinect sensors. The authors combined supervised and unsupervised learning for behavior classification. However, the unsupervised learning used in Nirjon et al. (2013) needs more interventions from the system's users in order to label the behaviors, which is not practical in real settings. Even though their work is similar to our work in terms of aggressive behaviors recognition using skeleton data, the main difference relies on the methodological side in terms of the features used and the classification algorithms employed. In addition, we use two more actions namely the wandering and tearing, which makes our data richer.

Some other researchers have looked at wearable devices for detection of agitation and aggression (Sakr et al. 2010; Rajasekaran et al. 2011; Plötz et al. 2012). For instance, in Sakr et al.'s work (Sakr et al. 2010) using wearable sensors to detect agitation, they used bio-physiological measures to detect agitation by monitoring the changes of the heart rate, galvanic skin response and skin temperature of the participants. In another study, Rajasekaran et al. (2011) proposed a wearable device for early detection of anxiety and agitation in people with cognitive impairment. Thomas et al. (Plötz et al. 2012) proposed a system based on machine learning techniques to segment relevant behavioral episodes from a continuous wearable sensor stream and to classify them into distinct categories of problem behavior such as aggression, disruption, and self-injury. The system was validated using simulated data of episodes of problem behavior acted out by trained specialists, and other daily living activities available datasets. The results from these studies showed accurate detection of these problem behaviors. However, all these studies required people to wear the device on the body to track their actions, which cannot be practical specifically when the sensor is taken off. In addition, the problem with the wearable device is that it could be stigmatizing to the users and not comfortable for users to wear, which could result in the abandonment of the device.

Although the aforementioned approaches showed good performance, they present some limitations. For example, (1) they fail in discriminating between similar actions, (2) most of them do not consider feature selection approaches for dimensionality reduction, and (3) most of them do not consider the combination of joint-based features and body part-based features in one unified model. These points motivate us to propose a more principled approach that combines the joint-based features and body part-based features using an ensemble learning method based on rotation forests. In addition, our approach employs singular value decomposition method for feature selection and dimensionality reduction.

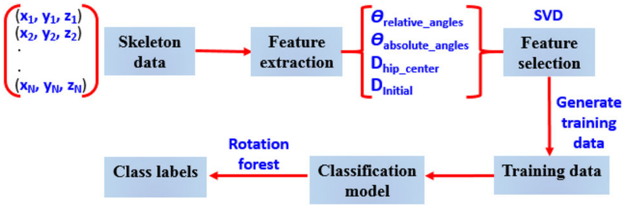


Fig. 1 Architecture of our approach

3 Proposed approach

In this section, we describe our approach of aggressive and agitated human behavior recognition in terms of feature extraction and ensemble learning classification. The general architecture of our approach is presented in Fig. 1. The details of each segment in Fig. 1 are presented in the following sections.

3.1 Feature extraction

A human skeleton can be represented by a hierarchy of joints that are connected with bones. The spatiotemporal features are local descriptions of human motions (Zhu et al. 2013). Therefore, an action can be described as a collection of time series of 3D positions. The time series of 3D positions represent 3D trajectories of the joints in the skeleton hierarchy. Figure 2 shows a graphical representation of the joints and how they are measured in the 3D space using a Microsoft Kinect sensor.

However, in order to accurately understand, recognize and differentiate between human actions, taking only 3D positions of the joints and how they evolve over time are not sufficient given the similarity between human actions. In order to obtain a better description and representation of human actions, we incorporate relative and absolute joint angles between each two connected limbs, and we represent the skeleton motion data as the changes over time of joint angles computed at each time frame. The aim of computing relative and absolute joint angles is to understand the contribution of each body part in performing actions. Moreover, we incorporate another feature which is the distance between the different joints and a fixed point of the skeleton, the hip centre, in order to give more information of the body parts involved in each movement over time. For instance, in a standing position, the hands are close to the hip center. When the hands are being raised up, the distance between the hands and hip center will increase. Figure 3 shows an example of joint angles and how they change over time, and the distance between body parts and the hip centre during the movement of rising hands. In addition, to characterize the spatial information of each joint, we compute the distance between the position of a joint at time t and its initial position at time t_0 (initial frame). This will further indicate how far the joint will be with respect to its initial position.

As we mentioned earlier, body part-based approaches represent the human body as a constellation of a set of rigid parts constrained in some fashion (Wang et al. 2012c). Angles between each consecutive two parts represent one of the important interpretable spatial features that allow to understand how body parts are related during human movements, which is important for action encoding. Note that, each body part is defined as a vector represented using two joint angle positions. For instance, the forearm is defined using the elbow joint and the wrist joint. Formally, let J be the skeleton joints in 3D space, and let $J_i(t) = (x_i(t), y_i(t), z_i(t))$ be the 3D coordinates of joint i at frame t. Therefore, the feature vector $\boldsymbol{F}_t$ structure at time t for each frame can be expressed as:

$$\boldsymbol{F}_t = [\theta_{relative-angle}, \theta_{absolute-angle}, D_{HipCenter}, D_{Initial}] \tag{1}$$

where $\theta_{relative-angle}$ consists of twelve relative angles at the following joints: shoulder, elbow, wrist, hip, knee and ankle for the left and right side, $\theta_{absolute-angle}$ consists of twelve absolute angles for the same joints computed with respect to the Kinect coordinate system as shown in Fig. 2, $D_{HipCenter}$ is the distance between each joint and the hip center, $D_{Initial}$ is the distance between a joint position at frame t and the initial frame t_0. These features are formally defined as follows:

1. The relative angle $\theta_{(P_1,P_2)}$ between two body parts P_1 and P_2 can be calculated using the triangle of joints J_i, J_j and J_k $(i \neq j \neq k)$,where P_1 is formed by joints J_i and J_j, and P_2 is formed by joints J_j, J_k as follows:

   $$\theta_{(P_1,P_2)} = \arctan(N(P_1 \times P_2, P_1.P_2)), \tag{2}$$

   where $P_1 \times P_2$ represents the cross product between the two 3D vectors P_1 and P_2, which results in a vector P, and $P_1.P_2$ is the dot product, which results in a scalar value σ. The $N(P, \sigma)$ is the normalization and is computed as follows:

   $$N(P, \sigma) = \left(\sum_{i=1}^{3} |P_i|^{\sigma} \right)^{\frac{1}{\sigma}}. \tag{3}$$

   Note that the angles are expressed in radian, therefore, all angles are multiplied by 180 and divided by π to get the values in degree. The absolute angles are computed in the same way by taking two skeleton joints with respect to the Kinect global coordinates system as origin.
2. The distance $D_{HipCenter}$ between each joint $J_i = (x_i, y_i, z_i)$ and the hip center $J_{hc} = (x_{hc}, y_{hc}, z_{hc})$ is computed as follows:

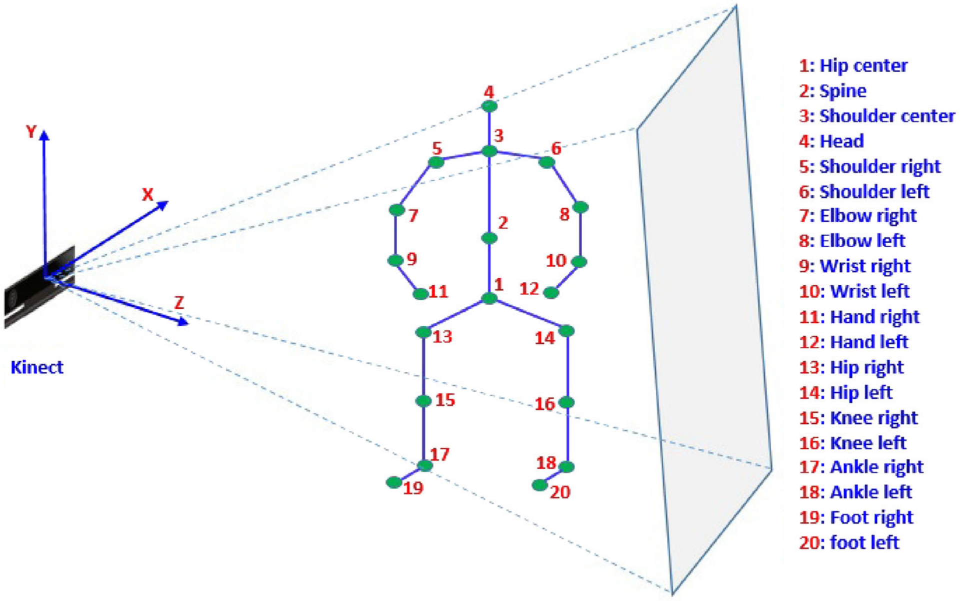


Fig. 2 Skeleton joints captured by a Kinect sensor

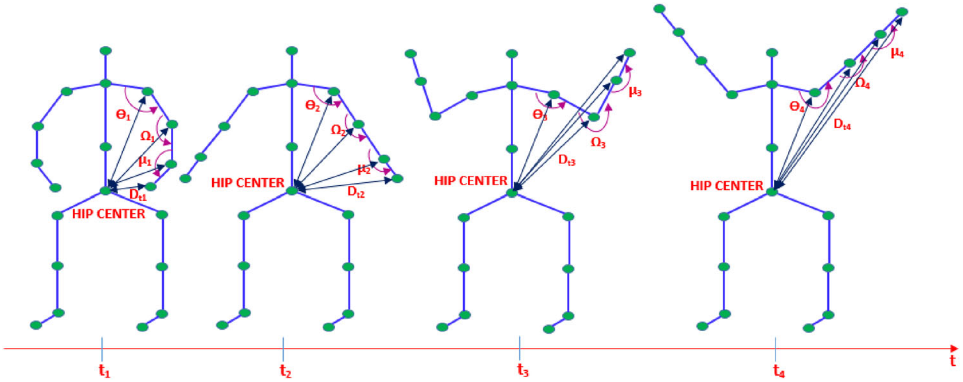


Fig. 3 Example of how joint angles and distances from the Hip center change over time during a movement

$$D_{HipCenter} = \sqrt{(x_i - x_{hc})^2 + (y_i - y_{hc})^2 + (z_i - z_{hc})^2} \tag{4}$$

3. The distance $D_{Initial}$ between a joint $J_i(t) = (x_i(t), y_i(t), z_i(t))$ at time frame (t) and the same joint at time frame (t_0) is computed as follows:

$$D_{Initial} = \sqrt{(x_i(t) - x_i(t_0))^2 + (y_i(t) - y_i(t_0))^2 + (z_i(t) - z_i(t_0))^2} \tag{5}$$

Note that, the new version of the Kinect sensor (i.e. v2) can track 25 skeleton joints instead of 20 joints. The five new joints are: Spine shoulder, Hand tip left, Thumb left, Hand tip right, and Thumb right. Therefore, more angles can be computed using this new version. Note that we combined all these features to make our approach robust in real environments. Overall, we have 75 features extracted for each frame. Once these features are computed, we can then combine them in order to build a classification model.

3.2 Feature selection

One of the key issues in classification algorithms is the large number of features used for the classification. To overcome this issue, we resort to feature selection algorithms in order to select the most discriminative features that will help us distinguish between the different classes, and reduce the classification space. In this paper, we use the singular value decomposition (SVD) method to select the most relevant features describing human behaviors. SVD has been widely used in information retrieval for reducing the dimension of the document vector space (Deerwester et al. 1990). Given a generic rectangular $m \times n$ matrix X, its singular value decomposition is:

$$X = U\Sigma V^T, \tag{6}$$

where U is a matrix $m \times r$, V^T is a matrix $r \times n$ and Σ is a diagonal matrix $r \times r$ where r is the rank of the matrix X. The diagonal elements of the Σ are the singular values such that $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_r \geq 0$. The two matrices

U and V are unitary, i.e., $U^T U = I$ and $V^T V = I$. It exists a direct relation between the informativeness of the dimension and the value of the singular value. High singular values correspond to dimensions of the new space where data have more variability, whereas low singular values determine dimensions where data have a smaller variability (Liu 2006). These dimensions can not be used as discriminative features in learning algorithms (Fallucchi and Massimo 2009).

In order to use SVD as feature selection, an important way is to exploit its approximated matrices, which means that, $X \approx X_\mathrm{k} = U_{m\times \mathrm{k}} \Sigma_{\mathrm{k}\times\mathrm{k}} V^T_{\mathrm{k}\times\mathrm{n}}$, where k is smaller than the rank r of the matrix X. The computation allows to stop at a given k different from the real rank r. Therefore, the singular vectors with largest singular values represent the selected features.

3.3 Fusion and classification using ensemble methods

Feature fusion is an important step to build a good classification model. Several classification methods could be used such as SVM, decision trees, and naive Bayes to perform classification. However, these methods have shown to be less accurate when compared to ensemble methods (Opitz and Maclin 1999). This motivates us to incorporate ensemble methods to build our classification model. The reason to use ensemble methods is to improve the predictive performance of a given model through combining several learning algorithms.

Rotation forest (Rodriguez et al. 2006) is an ensemble method proposed to build a classifier, which uses independently trained decision trees. It is found to be more accurate than bagging, AdaBoost and Random Forest ensembles across a collection of benchmark datasets (Ludmila and Juan 2007). The advantage of rotation forests lays in the use of principal component analysis (PCA) to rotate the original feature axes so that different training sets for learning base classifiers can be formed (Ludmila and Juan 2007).

Formally, let $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x_n}]^\mathbf{T}$ be a data point described by n features, and let X be an $m \times n$ matrix containing the training example. Let $Y = [y_1, \ldots, y_m]^T$ be a vector of class labels for the training data, where y_j takes a value from the class labels $\{w_1, \ldots, w_c\}$. Let $D = \{D_1, \ldots, D_L\}$ be the ensemble of L classifiers and $\boldsymbol{F}$ be a feature set. The idea is that all classifiers can be trained in parallel. Therefore, each classifier D_i is trained on a separate training set T_{D_i} to be constructed as follows (Rodriguez et al. 2006):

1. split the feature vector $\boldsymbol{F}$ into K subsets. The subsets may be disjoint or intersecting. Note that rotation forest aims at building accurate and diverse classifiers. Therefore, to maximize the chance of getting high diversity, it is suggested to take disjoint subsets of features. For instance, this can be obtained by taking $M = n/K$, where K is a factor of n.
2. for each of the subsets, select randomly a nonempty subset of classes and then draw a bootstrap sample of objects.
3. run PCA using only the M features in $\boldsymbol{F}_{i,j}$ and the selected subset of X, where j is the jth subset of features for the training set of classifier D_i. Then, store the obtained coefficients of the principal components $\mathbf{a^1_{i,j}}, \ldots, \mathbf{a^{M_j}_{i,j}}$ in a matrix $C_{i,j}$.
4. rearrange the columns of the matrix $C_{i,j}$ in a new matrix B^a_i so that they correspond to the original features in matrix X.
5. the training set for classifier D_i is XB^a_i.
6. to classify a new sample $\mathbf{x}$, we compute the confidence ψ for each class as follows:

$$\psi_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x}B^a_i), \quad j = 1, \ldots, c \tag{7}$$

   where $d_{i,j}(\mathbf{x}B^a_i)$ is the probability assigned by the classifier D_i indicating that $\mathbf{x}$ comes from class w_j. Therefore, $\mathbf{x}$ will be assigned to the class having the highest confidence value.

In rotation forest, bootstrap samples are taken as the training set for each base classifier, and a transformation of the feature set is performed for each base classifier. Finally, rotation forest combines the results of all base classifiers using majority voting method. The steps of our approach are presented in Algorithm 1. The next section presents the validation of our proposed approach.

```
Algorithm 1: Classification algorithm using Rotation Forest
Input:
- 3D coordinates of skeleton joints for all behavior instances
- L: the number of classifiers in the ensemble method
- K: the number of subsets
- the set of class labels $\{w_1,...,w_c\}$
Output:
- Class labels for new behavior instances
Training phase
foreach Behavior instance do
    foreach Time frame t do
        - Compute the feature vector $\boldsymbol{F}_t$
    end
    - Add $F_t$ to matrix $X$
end
- Compute matrices $U$, $\Sigma$ and $V^T$ form matrix $X$ using SVD Equation ( 6)
- Select k first singular values from matrix $\Sigma$ s.t. $\frac{\sum_{i=1}^{\mathrm{k}} \Sigma_{i,i}}{\sum_{i=1}^{n} \Sigma_{i,i}} \geq 90\%$
- Build the training set $X \approx X_{\mathrm{k}} = U_{m\times\mathrm{k}} \Sigma_{\mathrm{k}\times\mathrm{k}} V^T_{\mathrm{k}\times n}$
for i=1...L do
    - Split $\boldsymbol{F}$ into K subsets: $\boldsymbol{F}_{i,j}$ (j=1...K)
    for j=1...K do
        - Let $X_{i,j}$ be the dataset obtained using the features in $\boldsymbol{F}_{i,j}$
        - Eliminate a random subset of classes from $X_{i,j}$
        - Select a bootstrap sample $X'_{i,j}$ of size 75% of objects from $X_{i,j}$
        - Run PCA on $X'_{i,j}$ and store obtained coefficients in a matrix $C_{i,j}$
    end
    - Rearrange the columns of $C_{i,j}$ in a new matrix $B_i^a$ so that they match the
      order of features in matrix $X$
    - Build classifier $D_i$ using $XB_i^a$ as a training set
end
Classification phase
for a given $\boldsymbol{x}$ do
    - Compute $d_{i,j}(\mathbf{x}B_i^a)$
    - Compute confidence $\psi_j(\mathbf{x})$ using Equation ( 7)
    - Assign $\mathbf{x}$ to class having the largest confidence
end
```

4 Validation

We evaluate the performance of each feature representation described above on five different human action datasets of 3D skeleton data. Each dataset has almost completely distinct set of actions.

4.1 Datasets

In this section we present the datasets we used to validate our proposed approach. Two datasets (TRI) and (Kintense) contain agitated and aggressive behaviors, while the three others UTKinect, Florence, and MSR-Action3D contain different common human behavior actions such as drink, answer phone, tie lace, bow, and read watch. The goal of using these datasets is twofold:

– demonstrate the suitability of our approach for the recognition of aggressive and agitated behaviors.
– demonstrate that our proposed approach is generic and can be applied to different behavior actions.

In addition, the UTKinect, Florence, and MSR-Action3D datasets contain some actions that are common for people with dementia when they get agitated such as sit down and stand up and clap hands repetitively (Manoochehri and Huey 2012).

4.1.1 TRI dataset

The first dataset is obtained by conducting an experiment in Toronto Rehabilitation Institute-UHN (TRI-UHN). Ten (10) participants (6 males and 4 females, 3 among them were left-handed) were involved in this experiment to conduct six (6) actions (hitting, pushing, throwing, tearing, kicking and wandering) in front of a Kinect sensor v2. These actions have been identified as the most common challenging aggressive and agitated behaviors[1] observed from persons with dementia. These behaviors were selected from Cohen-Mansfield Agitation Inventory (CMAI) Scale (Cohen-Mansfield 1991). These behaviors are described as follows:

1. *Hitting* to perform this behavior, participants were asked to raise one of their hands up and pretend to hit something in front of them.
2. *Pushing* to perform this behavior, participants were asked to use their both hands at the same time and pretend to push something in front of them.
3. *Throwing* to perform this behavior, participants were given an object and asked to throw it out as far as possible using one hand. The object is a piece of light foam cut from a camping mattress.
4. *Tearing* to perform this behavior, participants were given a piece of paper and asked to tear it using both hands.
5. *Kicking* to perform this behavior, participants were asked to raise one of their feet up and pretend to kick something in front of them.
6. *Wandering* to perform this behavior, participants were asked to look for something that they couldn't find. They were asked to make a step forward and look for something on the ground from side to side and then look up for something from side to side, and then make a step backward and redo the same movements.

Participants were asked to perform the full set of actions using the right side of the body. For instance, hitting and kicking with the right hand and right foot respectively. Note that two of these actions, pushing and wandering, are not specific to one side of the body. In order to ensure the study is generic and takes into account both left-handed and right-handed people, participants were then requested to repeat the four laterally specific actions, hitting, kicking, throwing and tearing, using the left side of the body. Participants performed all the actions in front of a Kinect sensor five times while facing each of three directions (front, left and right). For example, during the hitting action, participants did first the action facing the Kinect sensor five times, then repeated the action another five times with their left side facing the Kinect and then another five times with their right side facing the Kinect. This is to ensure that we take into account different situations that might occur when a person is being monitored. A total of ((10 (participants) × 4 (behaviors) × 3 (sides) × 5 (repetitions) × 2 (left hand and right hand)) + (10 (participants) × 2 (wandering and pushing) × 3 (sides) × 5 (repetitions) × 1 (one side of body)) = 1200 + 300 = 1500) behavior instances have been collected in our experiment. Figure 4 shows an example of a skeleton and a depth image for each action performed by one participant.

Each action was performed using three different directions with respect to the Kinect sensor: front side facing the Kinect, right side facing the Kinect and left side facing the Kinect as shown in Fig. 5. To extract the skeleton data, we used the Kinect Stream Saver application developed by Dolatabadi et al. (2013) in our laboratory. Therefore, each skeleton data consists of 3D coordinates of 25 joints with time stamp indicating the time when the joint coordinates were recorded at each frame. All the skeleton data were recorded at 30 frame per second rate.

4.1.2 Kintense action dataset

In the Kintense action dataset (Nirjon et al. 2013), 19 healthy participants performed 4 different aggressive actions collected using a Kinect sensor. These actions were hitting, kicking, pushing and throwing. Each participant performed the four actions in different distances and different angles with respect to the Kinect sensor. Skeleton joint locations for 20 joints were provided in this dataset. Each action was performed 4–8 times by each participant. About 13000 action instances were collected in this dataset.

4.1.3 UTKinect action dataset

In the UTKinect action dataset (Xia et al. 2012), 10 participants performed 10 different action classes collected using a Kinect sensor. These actions were walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands. Skeleton joint locations were provided in this dataset. Altogether, the data set contained 6220 frames of 200 action samples. The length of sample actions ranged from 5 to 120 frames.

4.1.4 Florence action dataset

The Florence action dataset (Seidenari et al. 2013) that was collected at the University of Florence was captured using a Kinect sensor. It included 9 activities: wave, drink from a bottle, answer phone, clap, tie lace, sit down, stand up, read watch, and bow. During acquisition, 10 participants were asked to perform the above actions for 2–3 times. There was a total of 215 activity samples.

4.1.5 MSR-action3D dataset

MSR-Action3D dataset (Li et al. 2010) was a dataset of depth sequences captured by a depth camera. It contained 20 actions: high arm wave, horizontal arm wave, hammer,

[1] Here we use the terms Behavior and Action interchangeably.

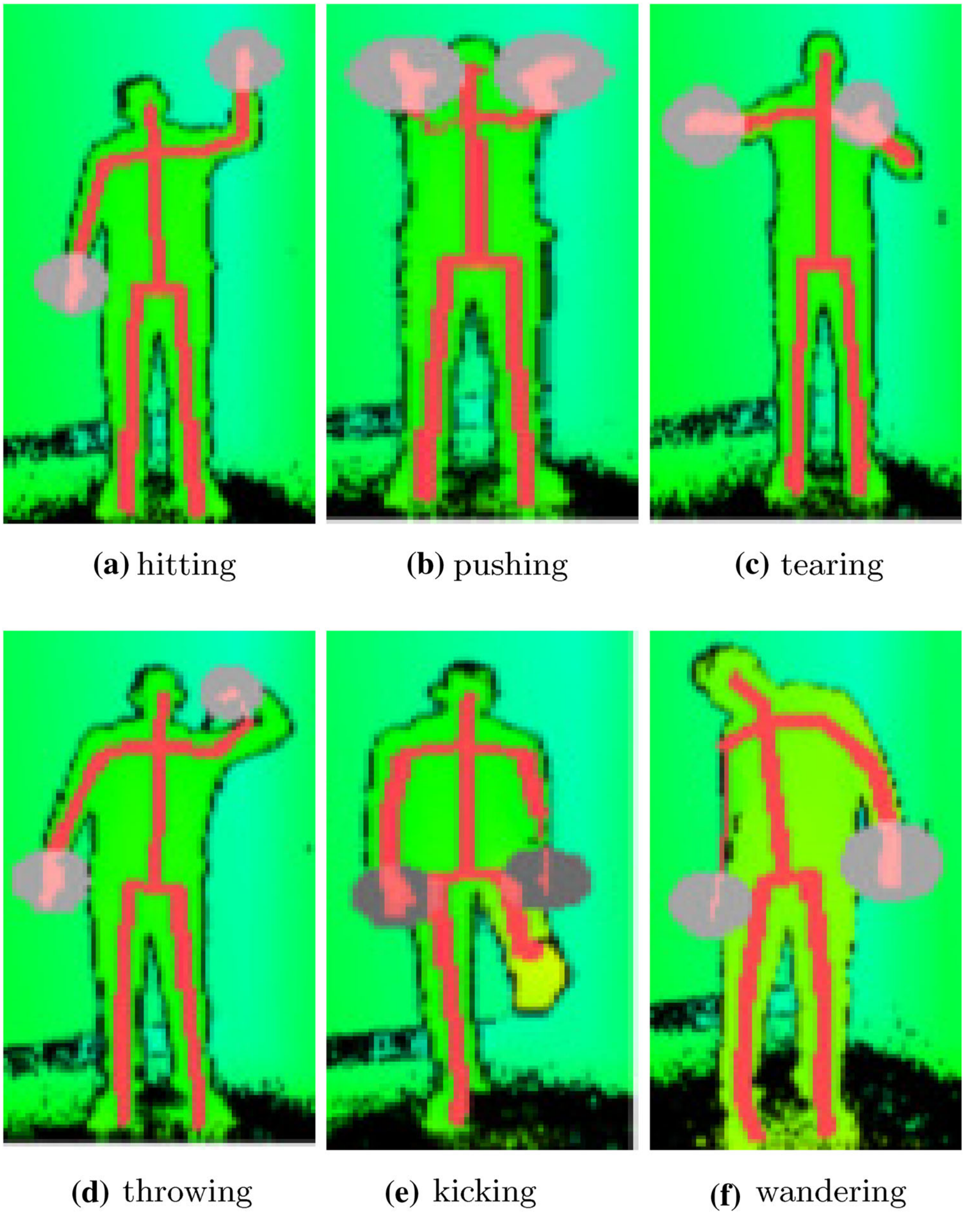


Fig. 4 Example of a skeleton and a depth image for each action performed by one participant

hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw. Ten (10) participants were involved in the study and were asked to perform each action for three times. The frame rate was 15 frames per second. This dataset is challenging because many of the actions in the dataset are highly similar to each other. We used the the same experimental setup as described in (Li et al. 2010; Yang et al. 2012) where the 20 action classes were divided into three main action sets, each containing 8 action classes with some overlap between action sets. All the classifiers were trained to distinguish between actions in the same action set only. The reported accuracy is the average over the three action sets.

4.2 Experimental results

We first evaluate the performance of our proposed approach using all the datasets. Then, we compare our results to the state-of-the-art methods to demonstrate the superiority and effectiveness of our proposed approach. In our experiments, we used the F-Measure, Accuracy and Mean Absolute Error (MAE) to present the results. We show first the recognition results before applying the SVD feature selection method. We then apply the SVD feature

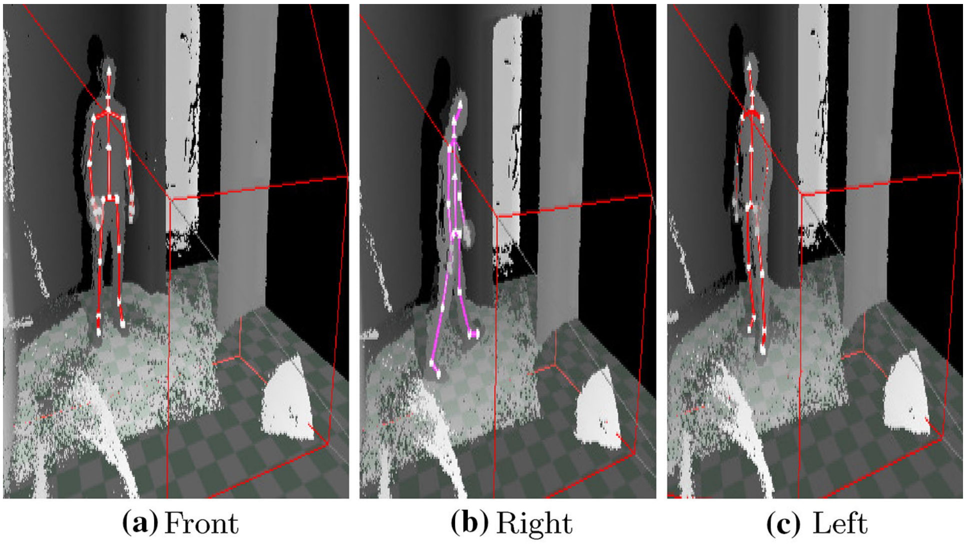


Fig. 5 Three different angles with respect to the Kinect sensor used during our experiments

selection method to show how a small set of features can achieve good recognition results comparing to the whole set of features.

4.2.1 Recognition results without feature selection

In order to show how the combination of features leads to a significant increase of the recognition accuracy, we have included the recognition accuracy obtained for each set of features separately and the combination of all these features as shown in Tables 1 and 2. For the TRI dataset, the results were computed for each direction (Front, Left and Right) with respect to the Kinect sensor, and averaged for the right-handed and left-handed. We used a 10-fold cross validation method to evaluate our approach.

As shown in Table 1, the combination of all the features leads to a higher recognition accuracy and a lower MAE when compared to the recognition results using features taken separately. For example, in the TRI dataset with right-handed, the recognition accuracy has improved with 1 % (99.75) when we combined distance to hip center feature (98.88) with the remaining three features. In addition, the MAE has decreased from (0.02) to (0.01) when all features are combined together. Similarly, in the TRI dataset with left-handed, the recognition accuracy has increased with 2 % from (97.76) using the absolute angle feature to (99.57) using all the features combined. In addition, the MAE has decreased drastically from (0.05) with the absolute angle feature to (0.02) with all the features combined. The high recognition results obtained could be explained by the fact that the features used in our approach are discriminative so that they allow to distinguish with high accuracy between the different behaviors. The same observations were found in the Kintense dataset.

In the other three datasets (UTKinect, Florence and MSR-Action3D dataset) with various human behavior actions, our approach also achieves high recognition accuracy as shown in Table 2. For example, in the UTKinect dataset, the recognition accuracy has improved with 4.29 % (98.37) when we combine the absolute angles feature (94.08) with the remaining three features (i.e. relative angles, distance to hip center and distance between current and the initial frame). Moreover, the MAE has decreased from (0.03) using absolute angles feature to (0.02) when all features are combined together. Similarly, the recognition accuracy has increased with 4.1 % (97.26) in the Florence dataset when compared to the recognition accuracy obtained using the distance between body parts and the hip center feature (93.15). Moreover, the recognition error has decreased from (0.05) to (0.03). In the MSR-Action3D dataset, the recognition accuracy has increased drastically with 8.75 % when we combine the absolute angles feature (82.65) with the remaining three features. Similarly, the MAE has decreased from (0.03) using the absolute angles feature to (0.02) when all features are combined together. This in turn demonstrates the effectiveness of the feature combination in terms of accuracy and recognition error. Therefore, our approach recognizes the actions with high accuracy and low recognition error in all the datasets.

Although the results show that when features were taken separately, such as absolute angles and distances with respect to the hip centre, are promising, these features may not be discriminative for actions involving same body parts such as the Hitting and Pushing actions, and the Kicking and Wandering actions. For example, as shown in Fig. 6a, b, the joint angles of the Elbow, Wrist and Handtip are involved in both the Hitting and Pushing actions, which

Table 1 Agitated and aggressive behavior recognition results obtained from TRI and Kintense datasets

Dataset	Features	F-Measure	MAE	Accuracy (%)
TRI Right-Handed	$\theta_{relative-angle}$	0.93	0.06	93.70
	$\theta_{absolute-angle}$	0.97	0.04	97.30
	$D_{HipCenter}$	0.98	0.02	98.88
	$D_{Initial}$	0.94	0.05	94.77
	ALL	0.99	0.01	99.75
TRI Left-Handed	$\theta_{relative-angle}$	0.94	0.09	94.78
	$\theta_{absolute-angle}$	0.97	0.05	97.76
	$D_{HipCenter}$	0.96	0.06	96.20
	$D_{Initial}$	0.83	0.15	83.58
	ALL	0.99	0.02	99.57
Kintense dataset	$\theta_{relative-angle}$	0.95	0.08	95.45
	$\theta_{absolute-angle}$	0.97	0.07	96.95
	$D_{HipCenter}$	0.98	0.04	98.63
	$D_{Initial}$	0.92	0.10	92.79
	ALL	0.99	0.04	99.12

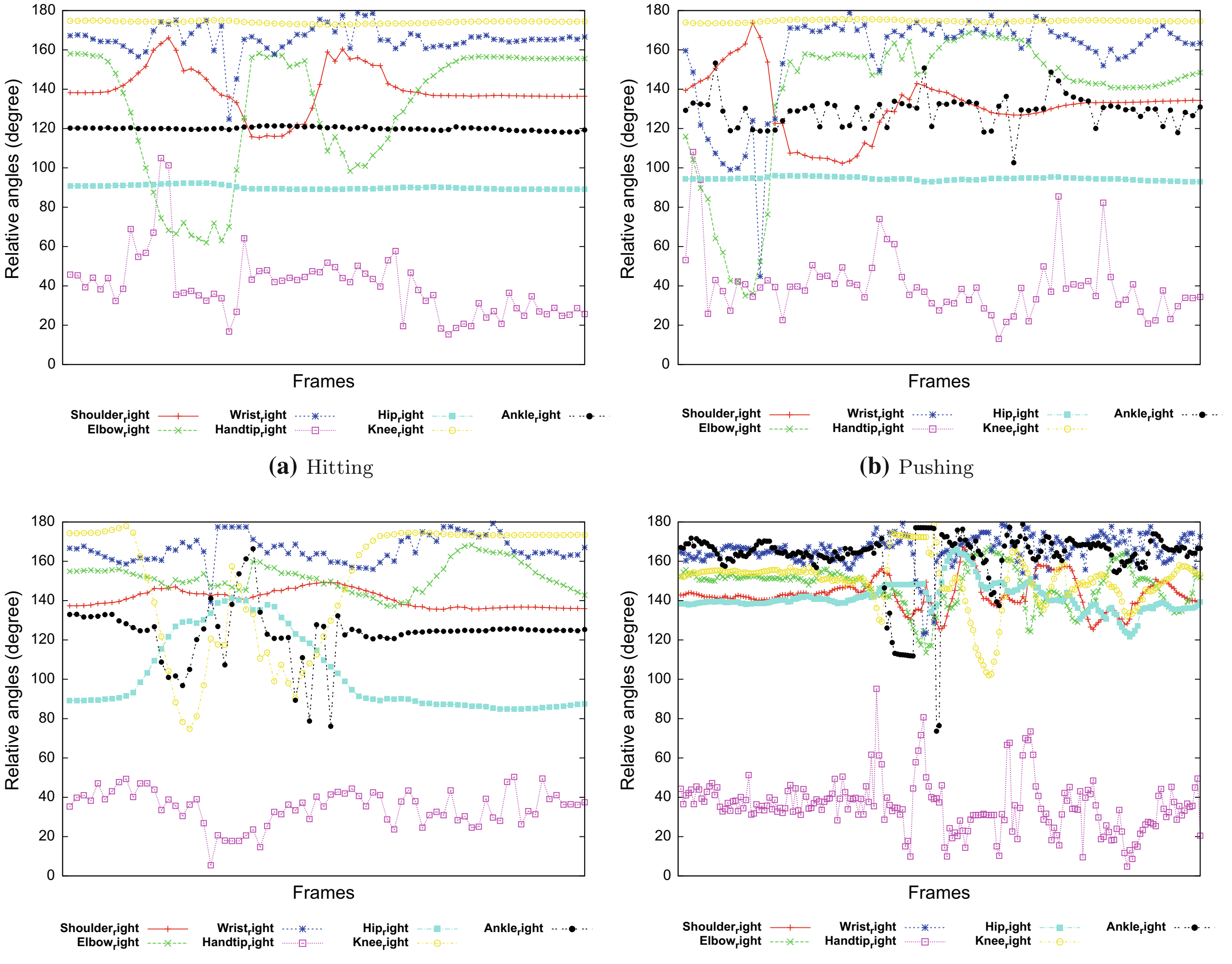
Table 2 Common behavior recognition results obtained from UTKinect, Florence and MSR-Action3D datasets

Dataset	Features	F-Measure	MAE	Accuracy (%)
UTKinect dataset	$\theta_{relative-angle}$	0.87	0.06	87.91
	$\theta_{absolute-angle}$	0.94	0.03	94.08
	$D_{HipCenter}$	0.92	0.04	92.79
	$D_{Initial}$	0.90	0.04	90.09
	ALL	0.98	0.02	98.37
Florence dataset	$\theta_{relative-angle}$	0.84	0.07	84.48
	$\theta_{absolute-angle}$	0.92	0.05	92.20
	$D_{HipCenter}$	0.93	0.05	93.15
	$D_{Initial}$	0.70	0.09	70.34
	ALL	0.97	0.03	97.26
MSR-Action3D dataset	$\theta_{relative-angle}$	0.78	0.04	78.59
	$\theta_{absolute-angle}$	0.82	0.03	82.65
	$D_{HipCenter}$	0.81	0.04	82.00
	$D_{Initial}$	0.65	0.05	65.98
	ALL	0.91	0.02	91.41

increases the similarity between these two actions. Similarly, in the Wandering action shown in Fig. 6c, more joint angles are involved during the performance of this action. This in turn increases the similarity between the Wandering and Kicking actions, which makes difficult to differentiate between them. This demonstrates how our approach performs better when all features are combined together. Therefore, the combination of these features yields a much better performance in terms of recognition accuracy and recognition error rate. This is a very important observation in gesture recognition applications. Indeed, in gesture recognition applications, the recognition accuracy is of high importance in order to personalize and adapt services according to the user gesture.

Despite with high recognition accuracy of our approach when all features are combined together, some actions are still misclassified with other actions as shown in Tables 3, 4, 5 and 6. For space limitation, we present confusion matrices only for TRI Right-Handed, Kintense, UTKinect and Florence datasets.

For instance, in TRI dataset shown in Table 3, 9 instances of the Hitting action were misclassified as the Throwing action and 23 instances of the Kicking action were misclassified as the Wandering action. Similarly, in Kintense dataset shown in Table 4, 290 instances of the Hitting action were misclassified as the Throwing action, and 208 instances of the Throwing action were misclassified as the Hitting action. In UTKinect dataset shown in Table 5, 12 instances of



(a) Hitting

(b) Pushing

(c) Kicking

(d) Wandering

Fig. 6 Example of relative angles computed for some actions performed by a participant in TRI right-handed dataset

Table 3 Confusion matrix obtained from the TRI Right-Handed dataset

	Hitting	Kicking	Pushing	Tearing	Throwing	Wandering
Hitting	**4211**	1	4	0	9	3
Kicking	0	**4518**	1	1	0	23
Pushing	3	3	**4114**	15	3	1
Tearing	0	1	7	**5648**	1	0
Throwing	10	1	1	6	**4845**	6
Wandering	0	1	0	0	1	**18857**

the Pick up action were misclassified as Walk action and 13 instances as Carry action. In Florence dataset shown in Table 6, 21 instances of the Tie lace action were misclassified as Bow action, and 9 instances of the Sit down action were misclassified as Bow action. The reason of the misclassified actions is due to the involvement of the same body parts to perform the action. For example, both the actions of Hitting and Throwing involve the arm to perform the action. This is also the case for the Kicking and Wandering actions that both involve the movements of the leg, and Tie lace and Bow actions that involve the movement of upper part of the body. However, this is a challenging and common issue to any classification algorithm where similarities are observed among the data.

Table 4 Confusion matrix obtained from the Kintense dataset

	Hitting	Kicking	Pushing	Throwing
Hitting	**27190**	29	51	290
Kicking	30	**33498**	17	52
Pushing	73	47	**16718**	60
Throwing	208	60	34	**30302**

4.2.2 Recognition results using feature selection

In order to show the effectiveness of the feature selection method we propose in our approach, we apply different values of k to all the datasets to select the relevant number of features for classification. The k used here is smaller than the rank *r* of the matrix *X* of the training data. We used k = 5, 10, 15, 20, 25, 30, 35, 40 and 45 in our experiments. Figure 7 shows the recognition results in terms of F-Measure obtained from all the datasets using different values of k.

The results show that with 20 features (k = 20), we are able to reach a high recognition accuracy. For example, in the TRI Right-Handed dataset, the recognition accuracy increased from 0.54 using k = 5 to 0.77 using k = 10, and reach to an accuracy of 0.96 when k = 20. Similarly, in the Kintense dataset, the recognition accuracy has increased from 0.77 using k = 5 to 0.93 using k = 10, with an improvement of 15.8 %. In addition, an improvement of 3.9 % has been achieved using k = 20 when compared to k = 10. Although the recognition accuracy continues to increase as k increases, the increase is much smaller when k is greater than 20. This indicates that the 20 features selected using SVD are able to describe the variability of the data and they are sufficient to describe and represent with high accuracy the different behaviors.

As shown in Fig. 7, when k = 20, the recognition results tend to be similar to those when k = 25, 30, 35, 40 and 45. Therefore, choosing k = 20 is considered to be a good empirical choice for the number of features in all the datasets.

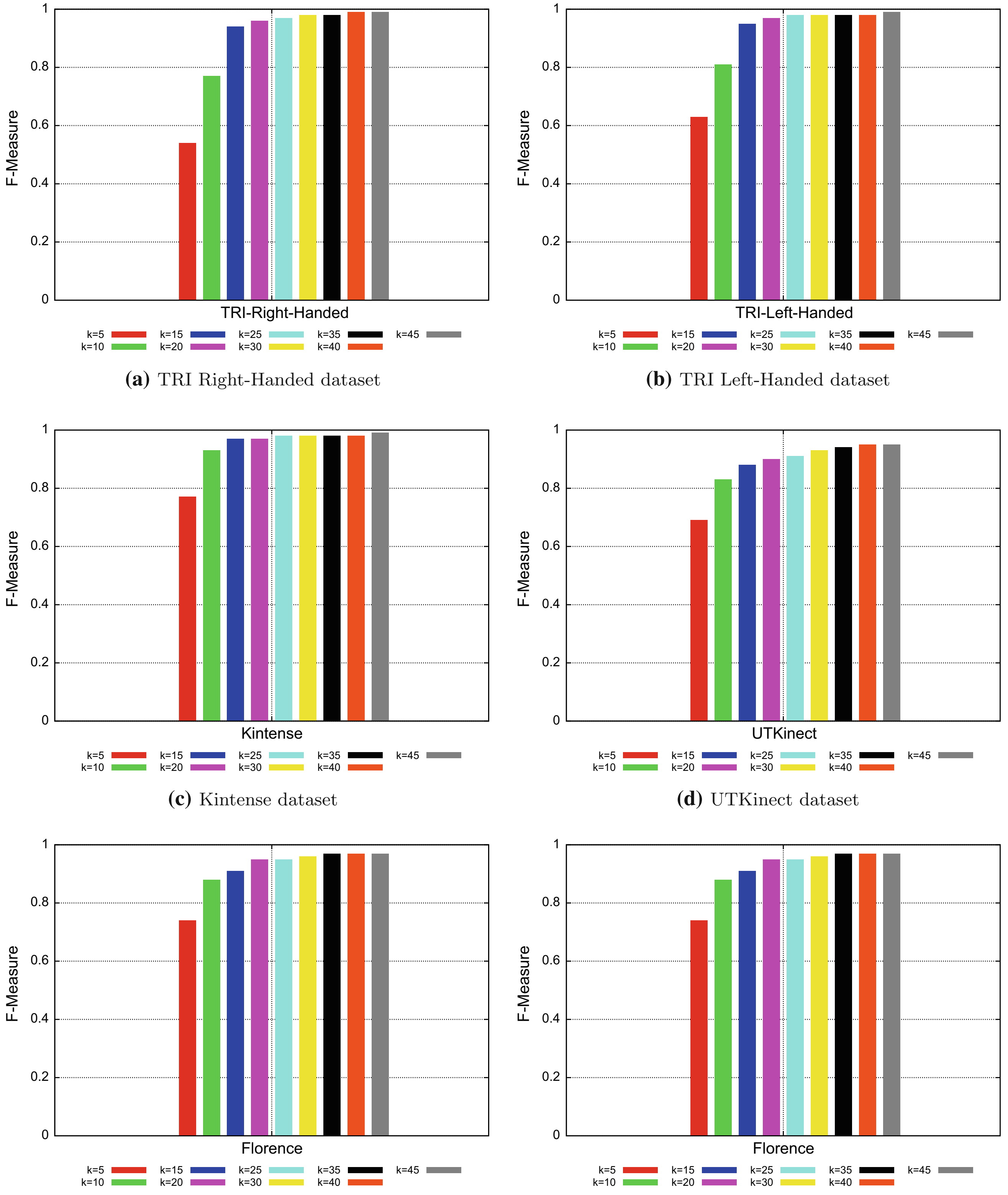
The potential of reducing the dimensionality is not only to reduce the size of feature set, but also to rely on the gain of the time when processing high dimensional data. For example, in a machine with 6 GB of memory and 2.5 GHz processor, taking all the features in our approach with TRI Right-Handed dataset resulted in a training set with more than 70 features and the time taken to build the model and

Table 5 Confusion matrix obtained from the UTKinect dataset

	Walk	Sit down	Stand up	Pick up	Carry	Throw	60Push	60Pull	60Wave hands	Clap hands
Walk	**808**	1	0	4	4	0	0	0	0	0
Sit down	3	**652**	0	8	2	0	0	0	0	0
Stand up	0	0	**493**	0	0	0	0	0	0	0
Pick up	12	8	2	**667**	13	0	1	0	0	1
Carry	3	0	0	1	**886**	0	0	0	0	1
Throw	0	0	0	0	1	**236**	2	2	1	1
Push	0	1	0	0	0	0	**184**	11	0	1
Pull	0	0	0	0	0	0	0	**273**	0	0
Wave hands	0	0	0	0	0	0	1	0	**876**	4
Clap hands	0	0	0	0	1	0	0	0	3	**566**

Table 6 Confusion matrix obtained from the Florence3D dataset

	Wave	Drink	Answer phone	Clap	Tie lace	Sit down	Stand up	Read watch	Bow
Wave	**432**	3	1	3	1	0	0	2	2
Drink	3	**417**	1	0	0	0	0	2	4
Answer phone	0	4	**384**	1	0	1	0	1	0
Clap	3	1	2	**361**	0	3	0	0	0
Tie lace	2	1	1	2	**529**	0	1	0	21
Sit down	0	0	0	4	3	**403**	1	1	9
Stand up	0	0	0	0	1	0	**421**	0	0
Read watch	3	1	2	11	1	2	0	**361**	1
Bow	0	0	0	1	1	2	0	0	**598**



(a) TRI Right-Handed dataset

(b) TRI Left-Handed dataset

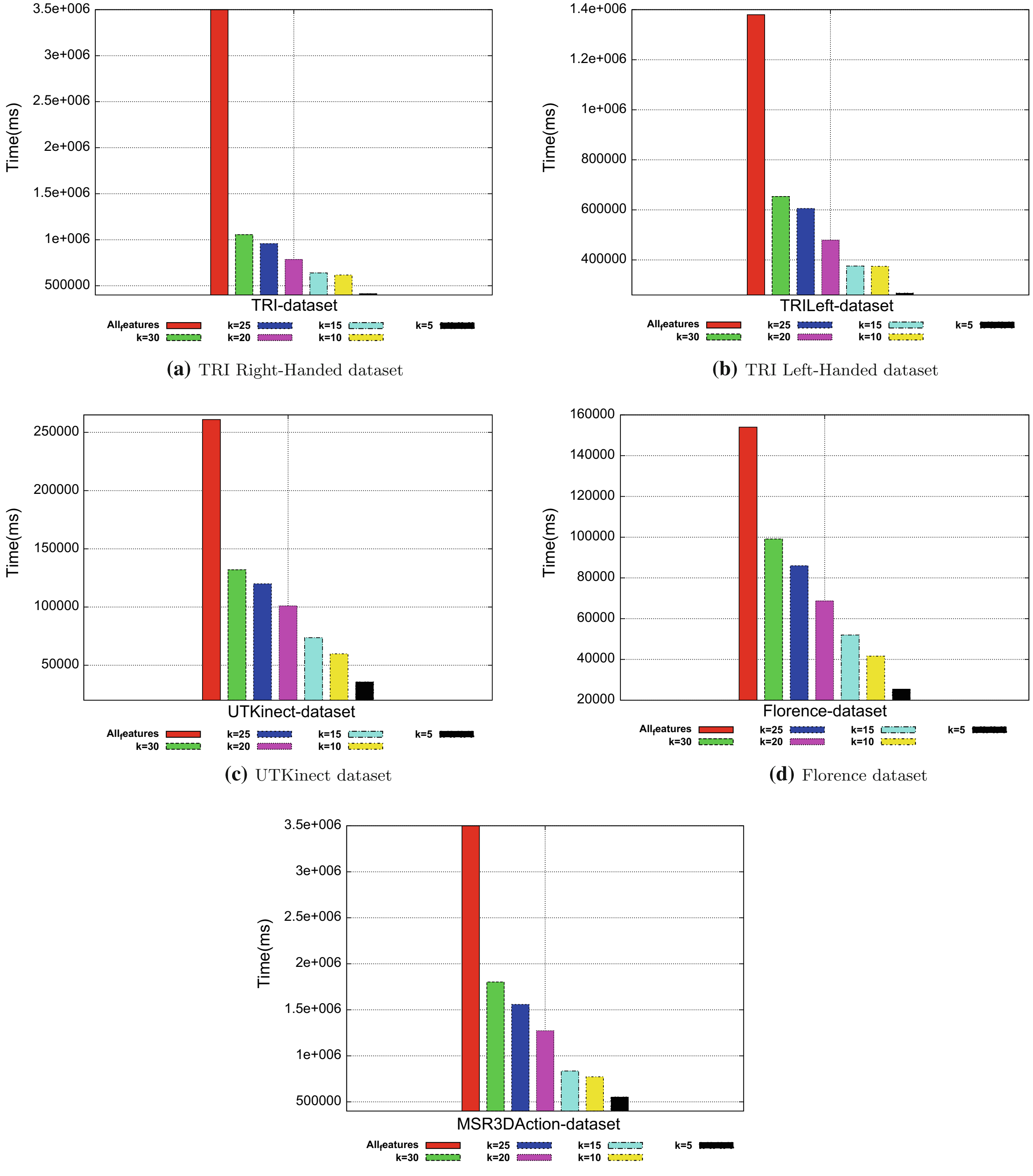
(c) Kintense dataset

(d) UTKinect dataset

(e) Florence dataset

(f) MSR-Action3D dataset

Fig. 7 Recognition results using different values of k



(a) TRI Right-Handed dataset

(b) TRI Left-Handed dataset

(c) UTKinect dataset

(d) Florence dataset

(e) MSR-Action3D dataset

Fig. 8 Classification time using different values of k

classify the data was 4,337,080 ms. However, the time taken to build the model and classify the data when k = 20 was only 785,462 ms, which was approximately 1/5 of the processing time when all features were taken into account.

Similarly, in the MSR-Action3D dataset, the time taken to build the model and classify the data with all features was 3,823,729 ms, and the time taken to build the model and classify the data was only 1,273,561 ms when k = 20. This

Table 7 Comparison of the recognition accuracy results obtained from the conventional classifiers and our approach in all the datasets

Classifiers	Datasets					
	TRI right-handed	TRI left-handed	Kintense	UTKinect	Florence	MSR-Action3D
DT (Quinlan 1999)	89.41	92.12	88.64	78.33	79.38	69.56
MLP (Haykin 1998)	76.89	79.72	65.24	78.75	71.76	44.62
SVM (Burges 1998)	52.71	50.32	73.31	37.93	55.35	53.55
BN (Friedman et al. 1997)	71.81	74.5	64.26	65.03	61.9	47.56
RF (Breiman 2001)	96.29	96.9	95.91	87.14	93.25	80.52
Decorate (Melville and Mooney 2004)	**96.6**	96.57	96.57	87.05	91.88	81.72
MetaCost (Domingos 1999)	44.31	29.83	29.99	15.5	14.99	8.85
AdaBoost (Freund and Schapire 1997)	48.44	35.07	34.1	30.64	19.32	11.5
Bagging (Breiman 1996)	94.31	94.3	92.47	82.17	87.3	77.6
Our approach	96.5	**97.5**	**97.2**	**90.2**	**95**	**82.3**

further indicates that k = 20 is a good choice for the number of features. Therefore, choosing k = 20 makes our approach more practical to deploy for real time applications. Figure 8 shows the execution time using different values of k.

4.2.3 *Ensemble method versus conventional classifiers*

One of the potentials of our approach is the use of the ensemble method based classification that aggregates many other classifiers (i.e. decision trees). In order to validate the superiority and performance of the ensemble method based classification over the conventional classification algorithms, we compared our approach with several conventional classification algorithms including single classifiers such as decision trees[2] (DT), multilayer perceptron (MLP), support vector machines (SVM)[3], and bayesian networks (BN)[4], and ensemble methods such as random forests (RF)[5], Decorate ensemble method[6] (Melville and Mooney 2004), MetaCost ensemble method[7] (Domingos 1999), AdaBoost ensemble method[8] (Freund and Schapire 1997), and Bagging ensemble method[9] (Breiman 1996). We used default settings of all classifiers provided by the Weka framework[10] Table 7 compares the recognition accuracy results obtained from the ensemble method and the conventional classifiers in all the datasets.

Although the conventional classifiers such as RF, Decorate and Bagging achieve good results, overall our approach performs better than all the conventional classifiers in all the datasets as shown in Table 7. The only dataset where the Decorate classifier achieves relatively better results (96.6 %) compared to our approach (96.5 %) is the TRI-Right-Handed. It is shown that overall SVM and BN classifiers achieve the lowest results in all the datasets. Similarly, MetaCost and AdaBoost ensemble methods achieve the lowest results in all the datasets. Table 7 also shows that RF has good recognition accuracy results in TRI, Kintense, UTKinect and Florence datasets as our approach. This can be explained by the fact that RF is considered as an ensemble method classifier as RF combines several decision tree based classifiers. However, in MSR-Action3D dataset with the large number of overlapping actions and the small number of instances for each action, all the classifiers do not show as high accuracy as our approach. Indeed, the RF classifier achieves an accuracy of 80.52 %, the Decorate classifier achieves an accuracy of 81.72 %, and the Bagging classifier only achieves an accuracy of 77.6 %, while our approach achieves the best results with an accuracy of 82.3 %. This further demonstrates the suitability and superiority of the rotation forest ensemble learning method over the conventional single and ensemble method classifiers.

4.2.4 *Comparison with state of the art approaches*

In order to compare our approach with the state of the art methods, we compared our approach with the approach of (Nirjon et al. 2013) for behavior recognition using expert classifiers, the approach of (Guo 2011) for behavior recognition using log covariance matrices, and the approach of (Zhu et al. 2013) for behavior recognition

[2] Confidence factor C = 0.25.

[3] SVM with radial basis function kernel.

[4] BN with K2 search algorithm.

[5] Number of trees n = 10.

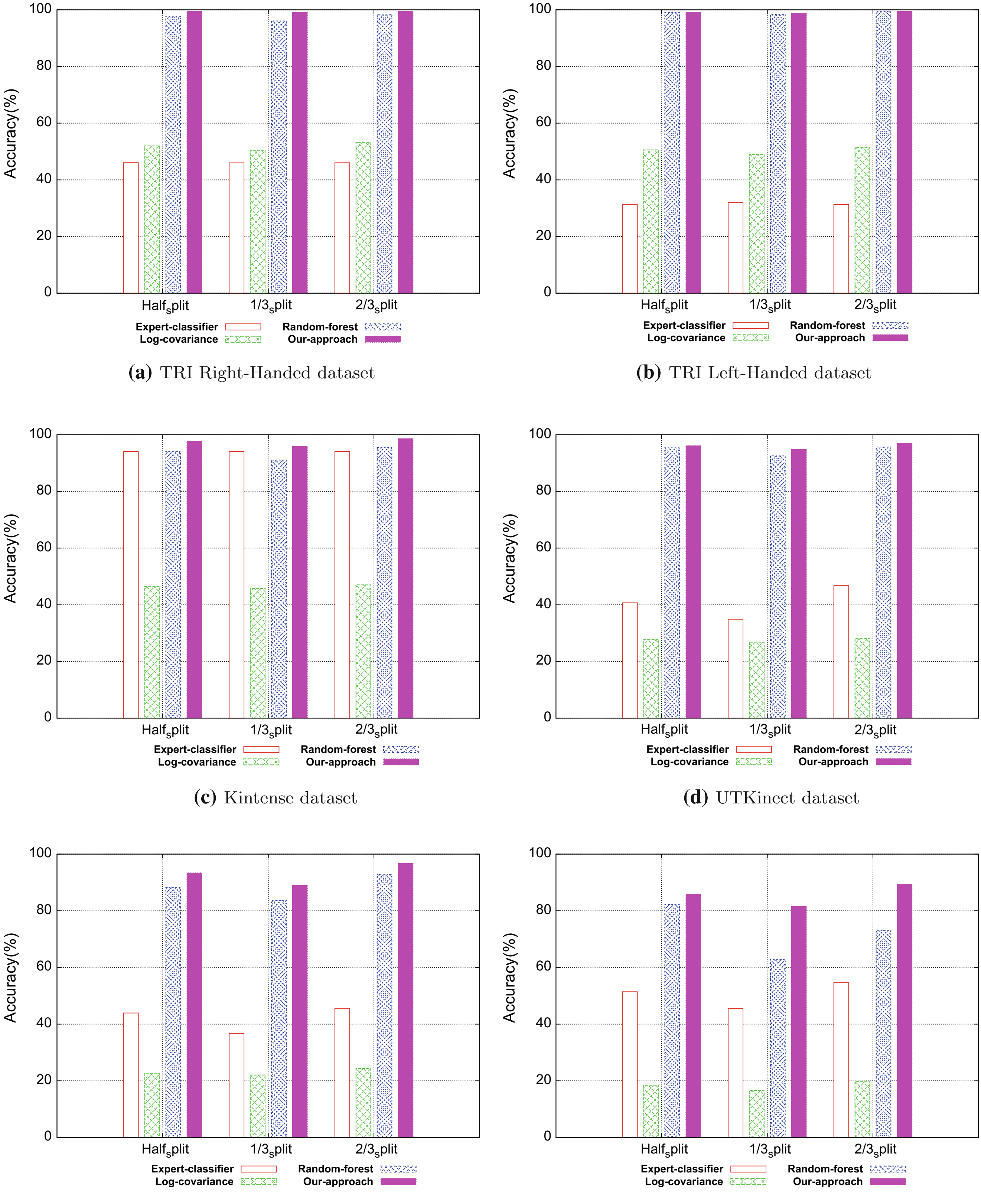
[6] Decision tree as base classifier.

[7] ZeroR as base classifier.

[8] Decision stump tree as base classifier.

[9] RepTree as base classifier.

[10] http://www.cs.waikato.ac.nz/ml/weka/.



(a) TRI Right-Handed dataset

(b) TRI Left-Handed dataset

(c) Kintense dataset

(d) UTKinect dataset

(e) Florence dataset

(f) MSR-Action3D dataset

Fig. 9 Comparison results with state of the art methods

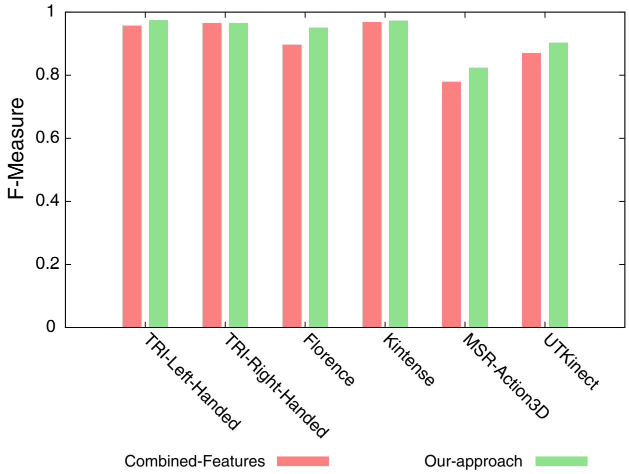


Fig. 10 Comparison between our approach and combined features from existing methods

using random forest based classification. We used different experimental settings such as 50 % subject split, 1/3 of data for training, 2/3 of data for training. The rational of using all these settings is: (1) to test the inter-subject generalization of our approach while using as much data as possible for training, (2) to test the sensitivity of our ensemble method classifier to reducing the number of training samples, and (3) to test the performance improvement when samples are used in training and testing. The comparison results are presented in Fig. 9 for each dataset.

As shown in Fig. 9, our approach achieves almost the best recognition results in all the datasets. The methods of (Nirjon et al. 2013) and (Guo 2011) achieves the lowest results in all the datasets, except for the Kintense dataset where the approach of (Nirjon et al. 2013) achieves good results. As shown in Fig. 9, the method of (Zhu et al. 2013) also achieves good results. This can be interpreted by the use of random forest classifier which is considered as an ensemble classifier that combines several decision trees classifiers. Overall, our approach performs better than the other approaches in all the datasets.

To highlight the importance of features used in our approach and their discriminative power, we compared the results obtained by our approach and those obtained by combing our features and features used in existing methods. Figure 10 shows a comparison of the results obtained.

As shown in Fig. 10, our approach performs better compared to the approach where features are combined from previous methods and our approach in all datasets. This demonstrates the importance of the features selected by the SVD method and their discrimination power compared to features from previous methods.

5 Conclusion

In this paper we have studied the problem of agitated and aggressive behavior recognition. We have proposed an effective approach based on feature fusion extraction from skeleton joint data. Our approach extracts first features such as absolute and relative angles, joint distance with respect to the hip center, and joint distances with respect to the initial frame. Then, a feature selection method is proposed based on the singular value decomposition in order to reduce dimensionality and to select the best features that are relevant to represent the different behaviors and to distinguish between them. For classification, we proposed an ensemble method classification based on rotation forest.

We have illustrated the effectiveness and suitability of our approach through extensive experiments on multiple real agitated and aggressive behavior datasets and common human behavior datasets. The experimental results show the suitability of our approach in representing behaviors and distinguishing between them. In addition, we have also illustrated how our approach outperformed several of the state of the art methods.

The work we have proposed in this paper constitutes a first step towards the development and deployment of a practical system for the recognition of agitated and aggressive behaviors for people with dementia. This in turn, opens new research directions in the ambient assisted living regarding the prediction of the occurrence of agitated and aggressive behaviors in people with dementia, and the issue of big data, specifically with images, videos and audio data, that require efficient and scalable algorithms for processing and management.

References

Aggarwal J, Cai Q (1999) Human motion analysis: a review. Comput Vis Image Underst 73(3):428–440

Aggarwal J, Ryoo M (2011) Human activity analysis: a review. ACM Comput Surv 43(3):1–16

Andreu J, Angelov P (2013) An evolving machine learning method for human activity recognition systems. J Ambient Intell Humaniz Comput 4(2):195–206

Ashok Krishnamoorthy DA (2011) Managing challenging behaviour in older adults with dementia. Prog Neurol Psychiatry 15(3):20–26

Bankole A, Anderson M, Smith-Jackson T, Knight A, Oh K, Brantley J, Barth A, Lach J (2012) Validation of noninvasive body sensor network technology in the detection of agitation in dementia. Am J Alzheimer's Disease Other Dement 27(5):346–354

Beeri MS, Werner P, Davidson M, Noy S (2002) The cost of behavioral and psychological symptoms of dementia (bpsd) in community dwelling alzheimer's disease patients. Int J Geriatr Psychiatry 17(5):403–408

Benayed S, Eltaher M, Lee J (2014) Developing kinect-like motion detection system using canny edge detector. Am J Comput Res Repos 2(2):28–32

Biswas J, Jayachandran M, Thang PV, Fook V FS, Choo TS, Qiang Q, Takahashi S, Jianzhong EH, Feng CJ, Kiat P YL (2006) Agitation monitoring of persons with dementia based on acoustic sensors, pressure sensors and ultrasound sensors: a feasibility study. In: International conference on aging, disability and independence, pp 3–15

Bouchard K, Bouchard B, Bouzouane A (2014) Spatial recognition of activities for cognitive assistance: realistic scenarios using clinical data from Alzheimer's patients. J Ambient Intell Humaniz Comput 5(5):759–774

Bouziane A, Chahir Y, Molina M, Jouen F (2013) Unified framework for human behaviour recognition: an approach using 3d zernike moments. Neurocomputing 100:107–116

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2(2):121–167

Chen L, Wei H, Ferryman J (2013) A survey of human motion analysis using depth imagery. Pattern Recogn Lett 34(15):1995–2006

Chikhaoui B, Wang S, Pigot H (2012) Adr-splda: activity discovery and recognition by combining sequential patterns and latent dirichlet allocation. Pervasive Mobile Comput 8(6):845–862

Chikhaoui B, Wang S, Xiong T, Pigot H (2014) Pattern-based causal relationships discovery from event sequences for modeling behavioral user profile in ubiquitous environments. Inf Sci 285:204–222

Cohen-Mansfield J (1991) Instruction manual for the cohen-mansfield agitation inventory (cmai). Research Institute of the Hebrew Home of Greater Washington

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

Desai AK, Grossberg GT (2001) Recognition and management of behavioral disturbances in dementia. Primary Care Companion J Clin Psychiatry 3(3):93

Dolatabadi E, Taati B, Parra-Dominguez GS, Mihailidis A (2013) A markerless motion tracking approach to understand changes in gait and balance: a case study. In: Proceedings of the RESNA annual conference, pp 391–400

Domingos P (1999) Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, pp 155–164

Duong TV, Bui HH, Phung DQ, Venkatesh S (2005) Activity recognition and abnormality detection with the switching hidden semi-markov model. In: Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on, vol 1, pp 838–845 (**IEEE**)

Fallucchi F, Massimo ZF (2009) Svd feature selection for probabilistic taxonomy learning. In: Proceedings of the workshop on geometrical models of natural language semantics, pp 66–73

Fook VFS, Thang PV, Mon T, Htwe QQ, Phyo A AP, Jayachandran BJ, Yap P (2007) Automated recognition of complex agitation behavior of demented patient using video camera. In: 9th international conference one-health networking, application and services, pp 68–73

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29(2–3):131–163

Gantenbein D (2012). Kinect launches a surgical revolution. http://research.microsoft.com

Gray KF (2004) Managing agitation and difficult behavior in dementia. Clin Geriatr Med 20(1):69–82

Guo K (2011) Action recognition using log-covariance matrices of silhouette and optical-flow features. PhD thesis, Boston University

Haykin S (1998) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall PTR, Upper Saddle River

Hussein ME, Torki M, Gowayyed MA, El-Saban M (2013) Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In: Proceedings of the twenty-third international joint conference on artificial intelligence, IJCAI '13, AAAI Press, pp 2466–2472

Kläser A, Marszalek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: Proceedings of the British machine vision conference 2008, Leeds, September 2008, pp 1–10

Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: Computer vision and pattern recognition workshops (CVPRW), 2010 IEEE computer society conference on, pp 9–14

Liu B (2006) Web data mining: exploring hyperlinks, contents, and usage data (data-centric systems and applications). Springer, New York

Lu C, Jia J and Tang CK (2014) Range-sample depth feature for action recognition. In: Computer vision and pattern recognition (CVPR), 2014 IEEE conference on, pp 772–779

Ludmila K, Juan R (2007) An experimental study on rotation forest ensembles. In: Proceedings of the 7th international conference on multiple classifier systems, pp 459–468

Luo J, Wang W and Qi H (2013) Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: Computer vision (ICCV), 2013 IEEE international conference on, pp 1809–1816

Maleki-Dizaji S, Siddiqi J, Soltan-Zadeh Y, Rahman F (2014) Adaptive information retrieval system via modelling user behaviour. J Ambient Intell Humaniz Comput 5(1):105–110

Mallidou A, Oliveira N, Borycki E (2013) Behavioural and psychological symptoms of dementia: are there any effective alternative-to-antipsychotics strategies? OA Fam Med 1(1):1–6

Manoochehri M, Huey ED (2012) Diagnosis and management of behavioral issues in frontotemporal dementia. Curr Neurol Neurosci Rep 12(5):528–536

Melville P, Mooney RJ (2004) Creating diversity in ensembles using artificial data. Inf Fusion 6:99–111

Mihailidis A, Boger JN, Craig T, Hoey J (2008) The coach prompting system to assist older adults with dementia through handwashing: an efficacy study. BMC Geriatr 8(1):28

Moore P, Xhafa F, Barolli L, Thomas A (2013) Monitoring and detection of agitation in dementia: towards real-time and big-data solutions. In: P2P, parallel, grid, cloud and internet computing (3PGCIC), eighth international conference on, pp 128–135

Mori T, Fujii A, Shimosaka M, Noguchi H, Sato T (2007) Typical behavior patterns extraction and anomaly detection algorithm based on accumulated home sensor data. In: Future generation communication and networking (FGCN 2007), vol 2, pp 12–18 (**IEEE**)

Nazerfard E, Cook DJ (2015) Crafft: an activity prediction model based on Bayesian networks. J Ambient Intell Humaniz Comput 6(2):193–205

Nirjon S, Greenwood C, Torres C, Zhou S, Stankovic JA, Yoon HJ, Ra HK, Basaran C, Park T, Son SH (2013) Kintense: a robust, accurate, real-time and evolving system for detecting aggressive actions from streaming 3d skeleton data. In: Proceedings of the 11th ACM conference on embedded networked sensor systems, pp 1–9

Ohn-Bar E, Trivedi M (2013) Joint angles similarities and hog2 for action recognition. In: Computer vision and pattern recognition workshops (CVPRW), 2013 IEEE conference on, pp 465–470

Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198

Oreifej O, Liu Z (2013) Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences. In: Computer vision and pattern recognition (CVPR), 2013 IEEE conference on, pp 716–723

Osunkoya T, Chern J-C (2013) Gesture-based human-computer-interaction using kinect for windows mouse control and power point presentation. Chicago State University, Chicago (Department of Mathematics and Computer Science 60628)

Plötz T, Hammerla NY, Rozga A, Reavis A, Call N, Abowd GD (2012) Automatic assessment of problem behavior in individuals with developmental disabilities. In: Proceedings of the 2012 ACM conference on ubiquitous computing, pp 391–400

Qiang Q, Fook FS, Phyo WAA, Thang PV, Jayachandran M, Jit B, Philip Y (2007) Multimodal information fusion for automated recognition of complex agitation behaviors of dementia patients. In: Information fusion, 2007 10th international conference on, pp 1–8 (**IEEE**)

Quinlan J (1999) Simplifying decision trees. Int J Hum Comput Stud 51(2):497–510

Rajasekaran S, Luteran C, Qu H and Riley-Doucet C (2011) A portable autonomous multisensory intervention device (pamid) for early detection of anxiety and agitation in patients with cognitive impairments. In: Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE, pp 4733–4736

Rodriguez J, Kuncheva L, Alonso C (2006) Rotation forest: a new classifier ensemble method. Pattern Anal Mach Intell IEEE Trans 28(10):1619–1630

Roy N, Misra A, Cook D (2016) Ambient and smartphone sensor assisted adl recognition in multi-inhabitant smart environments. J Ambient Intell Humaniz Comput 7(1):1–19

Sakr G, Elhajj I, Huijer H-S (2010) Support vector machines to define and detect agitation transition. Affect Comput IEEE Trans 1(2):98–108

Seidenari L, Varano V, Berretti S, Del Bimbo A and Pala P (2013) Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: Computer vision and pattern recognition workshops (CVPRW), 2013 IEEE conference on, pp 479–485

Sheng B, Yang W, Sun C (2015) Action recognition using direction-dependent feature pairs and non-negative low rank sparse model. Neurocomputing 158:73–80

Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: Proceedings of the 2011 IEEE conference on computer vision and pattern recognition, pp 1297–1304

Tampi RR, Williamson D, Muralee S, Mittal V, McEnerney N, Thomas J, Cash M (2011) Behavioral and psychological symptoms of dementia: parti epidemiology, neurobiology, heritability, and evaluation. Clin Geriatr 1–6

van Teijlingen W, van den Broek EL, Könemann R, Schavemaker JG (2012) Towards sensing behavior using the kinect. In: 8th international conference on methods and techniques in behavioural research, pp 372–375 (**Noldus Information Technology**)

Wang J, Liu Z, Chorowski J, Chen Z, Wu Y (2012) Robust 3d action recognition with random occupancy patterns. In: Proceedings of the 12th European conference on computer vision—volume part II, pp 872–885

Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Computer vision and pattern recognition (CVPR), 2012 IEEE conference on, pp 1290–1297

Wang Y, Tran D, Liao Z, Forsyth D (2012) Discriminative hierarchical part-based models for human parsing and action recognition. J Mach Learn Res 13(1):3075–3102

Xia L, Chen CC, Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: CVPR workshops, pp 20–27 (**IEEE**)

Yang X, Zhang C and Tian Y (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM international conference on multimedia, pp 1057–1060

Ya-Xuan H, Chih-Yen C, Hsu SJ, Chia-Tai C (2010) Abnormality detection for improving elder's daily life independent. In: Aging friendly technology for health and independence. Springer pp 186–194

Ye M, Zhang Q, Wang L, Zhu J, Yang R, Gall J (2013) A survey on human motion analysis from depth data. In: Time-of-flight and depth imaging. Sensors, algorithms, and applications: Dagstuhl 2012 seminar on time-of-flight imaging and GCPR 2013 workshop on imaging new modalities, pp 149–187

Zhan Y, Kuroda T (2014) Wearable sensor-based human activity recognition from environmental background sounds. J Ambient Intell Humaniz Comput 5(1):77–89

Zhu Y, Chen W, Guo G (2013) Fusing spatiotemporal features and joints for 3d action recognition. In: Computer vision and pattern recognition workshops (CVPRW), 2013 IEEE conference on, pp 486–491