IET Computer Vision

Research Article

Two-person activity recognition using skeleton data

ISSN 1751-9632
Received on 22nd February 2017
Revised 28th July 2017
Accepted on 27th August 2017
doi: 10.1049/iet-cvi.2017.0118
www.ietdl.org

Alessandro Manzi[1] ✉, *Laura Fiorini*[1], *Raffaele Limosani*[1], *Paolo Dario*[1], *Filippo Cavallo*[1]
[1]*The BioRobotics Institute, Scuola Superiore Sant'Anna, Viale Rinaldo Piaggio, 34, 56026 Pontedera (PI), Italy*
✉ *E-mail: alessandro.manzi@santannapisa.it*

Abstract: Human activity recognition is an important and active field of research having a wide range of applications in numerous fields including ambient-assisted living (AL). Although most of the researches are focused on the single user, the ability to recognise two-person interactions is perhaps more important for its social implications. This study presents a two-person activity recognition system that uses skeleton data extracted from a depth camera. The human actions are encoded using a set of a few basic postures obtained with an unsupervised clustering approach. Multiclass support vector machines are used to build models on the training set, whereas the *X*-means algorithm is employed to dynamically find the optimal number of clusters for each sample during the classification phase. The system is evaluated on the Institute of Systems and Robotics (ISR) - University of Lincoln (UoL) and Stony Brook University (SBU) datasets, reaching overall accuracies of 0.87 and 0.88, respectively. Although the results show that the performances of the system are comparable with the state of the art, recognition improvements are obtained with the activities related to health-care environments, showing promise for applications in the AL realm.

1 Introduction

In recent years, ambient-assisted-living (AAL) solutions have been developed in an attempt to meet the needs of the older population (and other relevant stakeholders) by enabling their independent living and promoting their quality of life and well-being [1]. AAL solutions integrate robots and other smart devices which are able to assist, guide, and support senior citizens in their daily lives [2]. In particular, these sophisticated agents need advanced human–machine interaction capabilities to cooperate effectively with users to perform specific tasks [3] such as to support the management of daily activities [4], improve social relationships [5], commitments and medicine reminding [6], and reveal potential dangerous situations [7]. In this context, it is evident how the capability to recognise human behaviours and the interaction between two (or more) persons plays an important role. This feature can be used by smart agents to understand a particular situation reacting properly. In the literature, most of the researches in the field of activity recognition are primarily focused on a single user performing daily activities (walking, sitting, and sleeping) [8], using objects (eating with a spoon and cooking) [9], and interacting with a robot [10].

On the other hand, the interaction between two or more persons represents a fundamental aspect of human life. However, compared to single-user activity recognition this issue is less popular among researchers and requires further investigations. The identification of this type of interactions involves the recognition of body postures, gestures, and key poses, and can introduce also complex social and psychological aspects, which depends on people's feelings and thoughts, and is influenced by context, culture, and personal attitude [11]. Consequently, the study of social activity is an emerging field of research in various communities including human–computer interaction, machine learning, speech processing, and computer vision [12].

Certainly, a system which is able to detect and recognise two-interacting people automatically can be applied in AAL contexts. Common actions requiring assistance in the health-care environment are walking, standing up, and drawing attention. Help is also needed for aggressive behaviours such as fighting and normal interactions such as handshake and conversation.

Over the past few years, technologies used to address these human activities have varied. Certain solutions have employed wearable sensors to obtain data based on body postures [13]. However, though these devices have some advantages, they can be cumbersome and invasive. In specific situations such as with older persons who have dementia, Alzheimer's, or other cognitive disorders, wearable solutions are not a pragmatic solution for activity recognition systems [14]. Other approaches employ the use of video cameras to extract features based on human silhouettes [15] or spatiotemporal scale-invariant properties [16]. In general, these solutions suffer from problems related to computational efficiency and robustness to illumination changes [17, 18]. An alternative to two-dimensional (2D) images is presented by depth cameras (also known as RGB-D cameras), which provide 3D data at a reasonable frame rate. These particular devices make available both colour and depth information simultaneously. Moreover, specific software trackers that can extract human skeleton models from depth maps have been implemented [19]. These skeleton features can be used to develop technology and innovative services [20] in AL applications [21].

This paper presents a human activity recognition system for two-person interaction based on skeleton data extracted from a depth camera. The use of skeleton data allows to have a system that is robust to illumination changes and, at the same time, can provide much more privacy compared with standard video cameras. These features are extremely valuable in the AAL context. The developed system is an adaptation of the algorithm presented in [22]. The basic idea behind the implementation is to represent an activity with a few and basic postures that can be used to generate multiple and minimal activity features. The basic postures are obtained employed an unsupervised clustering technique. The current implementation differs from the previous mainly in two aspects. First of all, it can handle two skeletons, modifying the early preprocessing step. It considers an interaction made by an active and a passive person, normalising the last with the first one. Second, instead of using the same number of clusters for all the activity samples, the system builds a set of classifier models trained with activity features generated on a range of clusters. During the classification, it calculates on-the-fly the optimal number of clusters for the unseen input sequence and retrieves the pre-trained model which corresponds to the value found.

The system is tested on two public datasets: the Institute of Systems and Robotics (ISR) - University of Lincoln (UoL) 3D social activity dataset [23] and the SBU Kinect interaction dataset [24], both containing skeleton data of two-interacting persons.

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)

Both datasets are composed of eight activities and three common actions (i.e. shaking hands, pushing, and hugging). While the first one contains longer and complex sequences, the second one has more samples of few seconds.

The remainder of this paper is organised as follows. Section 2 focuses on the related works on this field. Section 3 describes the developed system, detailing the generation of the skeleton, and activity features. The experimental results, using three different classification strategies, are presented in Section 4, whereas Section 5 concludes this paper.

2 Related works

Activity recognition is an important and active area in computer vision. Even though early studies in this field began in the 1980s, the field still presents challenges. In the literature, various technologies are used to address this issue, ranging from wearable [25] to smartphones sensors (e.g. accelerometer, gyroscope, and magnetometer) [26], and images from standard video cameras [17, 18].

Recently, researchers have evolved into the use of consumer depth cameras [27], also known as RGB-D devices, which, with inexpensive costs and smaller sizes, are able to provide both colour and depth information simultaneously. Concerning this technology, most research is focused on the recognition of activities performed by a single person. Numerous datasets have been built, and, among these, the most popular are the Cornell Activity Dataset (CAD)-60 [28], CAD-120 [29], RGB-D-HuDaAct [30], and MSRDailyActivity3D [31]. Other works underline the importance of the user's point of view, analysing the information gathered by a first-person camera to recognise the interaction level of human activities from a first-person viewpoint [32]. Comprehensive surveys on human motion analysis with depth cameras can be found in [27, 33].

Differently from the aforementioned works, this paper focuses on the analysis of two-person interaction activities for AAL applications. Over the last year, different research groups have focused their efforts on different aspects of the interaction between two or more persons. For instance, some authors investigate the social aspects involved in the interactions. Rehg *et al.* [34] investigate the interactions between children aged 1–2 years and an adult. They fuse information acquired from different sensors (RGB-D cameras, microphones, and physiological sensors) to capture also the social aspects of the interaction performed over the 160 sessions. Kong *et al.* [35] present a method based on primitive interactive phrases for recognising complex human interactions. The idea is to describe a relationship through a set of primitives which are common to different relationships. Another similar work is presented by Raptis and Sigal [36], where they recognise the most discriminative keyframes while also learning the local temporal context between them. Huang and Kitani [37] aim to build a system based on reinforced learning able to predict and to simulate human behaviour in both space and time from partial observations. When two people interact, they observe only the actions of the initiator on the right-hand side and attempt to forecast the reaction on the left-hand side. Blunsden and Fisher [38] use the movement as a social feature to classify the interaction in a group of people using a hidden Markov model. Another important aspect related to the interaction between two persons is the body orientation which could express a connection with the person's perceived emotions. In this context, Lopez *et al.* [39], in a recent study, analyse the body posture assumed by 26 undergraduate students and recognise six basic emotions. Additionally, some researchers focused their efforts on what can be defined as 'group activity recognition', in which activities are performed by groups. For instance, Cheng *et al.* [40] consider a group of people (more than two persons) and their spatiotemporal relationships. Indeed they model the interaction considering three different layers which are semantically interpretable and described by a set of heterogeneous features. Their analysis starts at the individual level, then each pair of users is considered, and at the end the analysis moves to the group level.

Another advantage of the depth cameras is the possibility to obtain human skeleton data in real time. These software trackers are able to provide the 3D joint coordinates of a human model. A comprehensive survey on skeleton-based classification with a single user can be found in [41]. In the literature, there are few public datasets containing two (or more) interacting persons. In [42], Ibrahim *et al.* propose a deep model based on long–short-term memory models to classify actions on the collective activity dataset [43]. The same dataset was used by Tran *et al.* [44], where a graph-based clustering algorithm is used to discover interacting groups in crowded scenes; a bag-of-words approach is employed to represent group activity, and a support vector machine (SVM) classifier is applied for activity recognition. Even if the collective activity dataset is a popular dataset for 'group activity recognition', it is clear that the classes (e.g. crossing, waiting, queuing, walking, and talking) are activities performed by a single person in the context of a group of people. Another interesting dataset is the Nanyang Technological University (NTU) RGB + D [45], which contains actions made by one and two people. To overcome the limitation of other datasets related to the small number of subjects and very narrow range of performers' ages, and the highly restricted camera views, the authors captured the actions performed by 40 subjects from 80 different viewpoints.

Among the skeleton-based classification method of two-interacting persons, Yun *et al.* [24] present the SBU Kinect dataset [46] composed of eight activities (*approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands*), containing images and skeleton data of seven participants and 21 pairs of two-actor sets. The authors use relational body-pose features which describe geometric relations between specific joints in a single pose or a short sequence of poses such as joint distance, joint motion, velocity, and plane. Regarding the classification, they employ linear SVMs and multiple instance learning (MIL) with a bag of body poses.

Another interesting dataset is the ISR-UoL, introduced in [47], which contains eight activities and ten sessions involving six participants. The authors implement an activity recognition system using a probabilistic ensemble of classifiers called dynamic Bayesian mixture model (DBMM) adapted from [48]. The authors merge spatiotemporal features from individual bodies and social features from the relationship between two individuals to classify the actions. Additionally, the authors aim to learn priority between subjects applying proximity theories to feed the classifiers.

The aim of this paper is to present and evaluate a system for two-person interactions using skeleton-based features for the AAL context. The current system is adapted from a previous version used for single-user activity classification [22]. Hence, we present a system, which implements a simple technique to extract relevant activity features suitable both for single-user and two-interacting persons. Conversely, from other works that groups a sequence of skeletons using a high number of clusters [49, 50], the developed system models an activity using a few and basic informative postures, ranging from 2 to 5 clusters. During the classification, it calculates on-the-fly the optimal number of clusters for the unseen input sequence, loading the pre-trained model correspondent to the obtained value. The algorithm is tested on the ISR-UoL and SBU Kinect datasets evaluating both a fusion at decision and at feature-level scheme, providing the running time for the best configuration.

3 Two-person activity recognition system

The developed system implements a human activity recognition method using 3D skeleton joint data extracted from a depth camera. The choice of excluding the image information is driven by the fact that the final system has to guarantee as much privacy as possible for the final users. Furthermore, the use of skeleton data gathered from depth maps is more robust to illumination variance than standard images. The basic concept of the developed system is to describe an activity using several sequences of a few basic informative postures. The idea is to model an activity sequence using simple and general features, extracting basic postures that are common to the execution of a specific activity. First of all, the skeleton data are normalised to be as independent and general as

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)

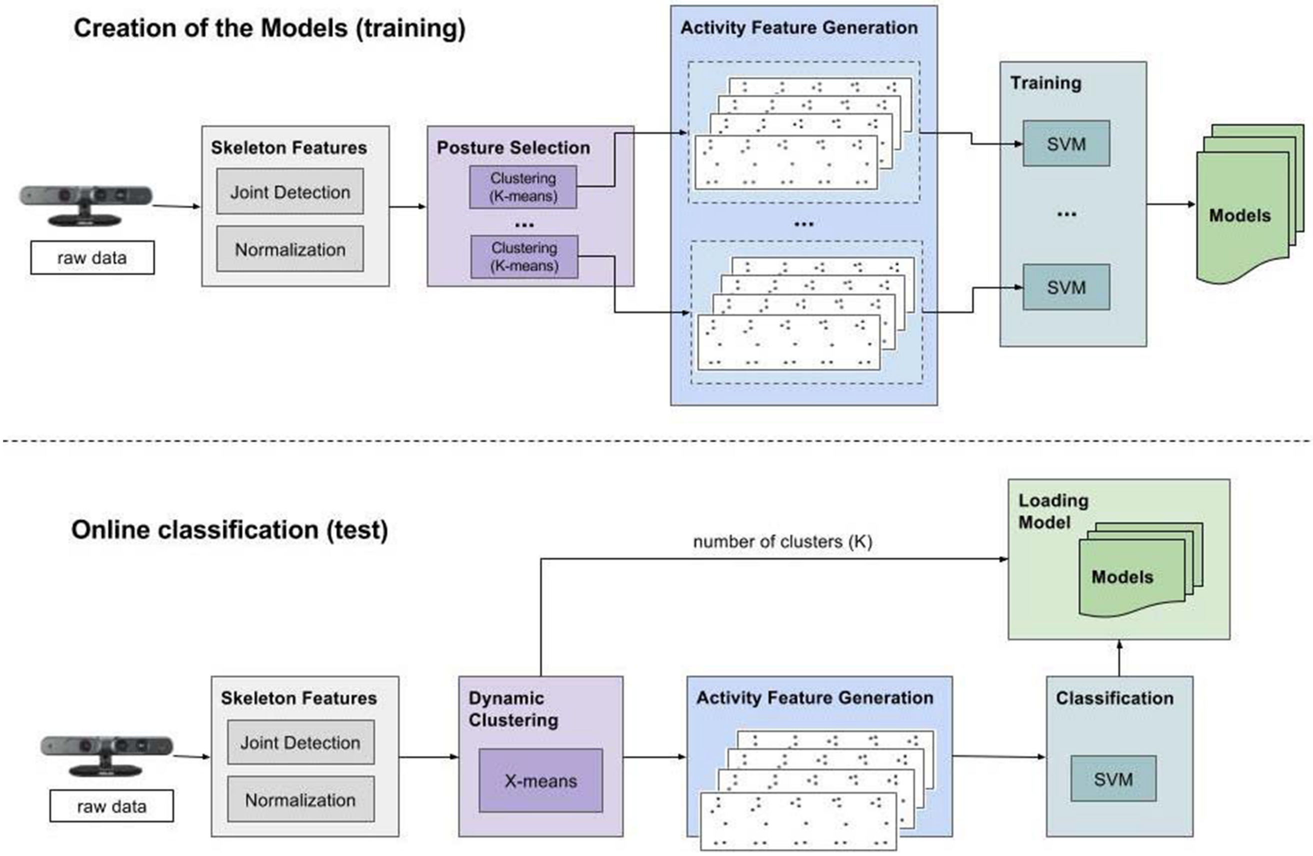


Fig. 1 *Software architecture of the activity recognition system. It is composed of two phases: creation of the models (top) and online classification (bottom). The first phase trains a set of SVM classifier that will be used during the second phase. These phases share some software blocks, i.e. extraction of the skeleton features and generation of the feature activity*

possible with respect to the current person. The key poses are calculated employing an unsupervised clustering technique, to obtain a few and basic informative postures. Then, suitable activity features are calculated from the original sequence and a supervised classifier is adopted for the training and classification. This system is an improvement of the work already presented in [22]. The current implementation differs from the previous one, mainly in two aspects. First of all, the preprocessing step is adapted to handle an activity frame composed of two skeletons. Second, instead of using the same number of clusters for all the activities, now the optimal number of basic skeleton poses can vary according to the complexity of the activity that the system tries to describe. Therefore, the classification phase clusters unseen data dynamically, without forcing a priori the number of groups. For this reason, the system adopts two different procedures regarding the train and test phases (see Fig. 1). The first implementation concerns the creation of the models, and the number of clusters is varied for each activity. The latter procedure uses one of the generated models to infer the class to which the input sequence belongs. These two phases share most of the same software structure, with the exception that the aim of the first phase is to generate and save suitable activity models, whereas the second phase uses these models to perform the classification.

The aim of this paper is to evaluate a system for two-person interactions using skeleton-based features for the AAL context. The developed system implements a simple technique to extract relevant activity features using few and informative postures, ranging from 2 to 5 clusters. Our aim is to evaluate the generality of the current system, which is already tested for single-user activity, also in the case of two-interacting persons. The algorithm is evaluated on the ISR-UoL and SBU Kinect datasets. Since the first dataset contains long and complex sequences, we also evaluate the performances of the system considering a subset of the original sequence, without changing the implementation. In Section 4, we investigate the use of the developed system with features extracted from each person independently. We then fuse the features at the decision and feature levels. The remainder of this section focuses on the implementation of the aforementioned phases.

3.1 Generation of the classification models

The generation of the classification models is composed of four main steps. At the beginning, the skeleton software tracker [19] detects the joints of the humans from the depth camera device. Then, a clustering algorithm is used to retrieve the group of similar postures, iterating on several numbers of clusters. Next, the activity features are generated from the previous sequence, and finally, several classifiers are trained for each set of obtained clusters. The remainder of this section details further steps in the process.

3.1.1 Skeleton feature extraction: The input data of the system is a sequence of two human skeletons extracted using a software tracker [19]. Each sequence contains an active and passive user. Each human skeleton is modelled with 15 joints that are represented as 3D Cartesian coordinates with respect to the sensor. Since these data depend on the distance between the users and the sensor, they cannot be used directly and need to be normalised. Therefore, the reference frame is moved from the camera to the torso joint of the active person, while the passive person is referenced to the active one. In addition, the joints are scaled with respect to the distance between the neck and the torso joint. This last step makes the data more independent to the specific subject dimensions such as height and limb length [50, 51].

Formally, if we consider a skeleton with N joints, the skeleton feature vectors of the first ($\boldsymbol{f}_A$) and the second ($\boldsymbol{f}_B$) person are defined as

$$\boldsymbol{f}_A = (\boldsymbol{j}_{A_1}, \boldsymbol{j}_{A_2}, \ldots, \boldsymbol{j}_{A_N}) \quad \boldsymbol{f}_B = (\boldsymbol{j}_{B_1}, \boldsymbol{j}_{B_2}, \ldots, \boldsymbol{j}_{B_N}) \qquad (1)$$

where each $\boldsymbol{j}$ is the vector containing the 3D normalised coordinates of the joint $\boldsymbol{J}$ detected by the sensor. Therefore, $\boldsymbol{j}_{A_i}$ and $\boldsymbol{j}_{B_i}$ are defined as

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)

$$j_{A_i} = \frac{J_{A_i} - J_{A_1}}{\| J_{A_2} - J_{A_1} \|} \quad i = 1, 2, \ldots, N \tag{2}$$

and

$$j_{B_i} = \frac{J_{B_i} - J_{A_1}}{\| J_{B_2} - J_{B_1} \|} \quad i = 1, 2, \ldots, N \tag{3}$$

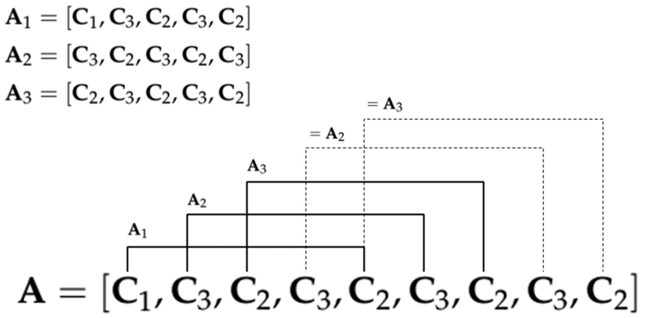
where J_{A_1} and J_{B_1} are the coordinates of the first and second user torsos, respectively, and J_{A_2} and J_{B_2} are the coordinates of the neck of the first and second persons, respectively.

The number of attributes for each feature vector in (1) is equal to $3N$. During our experiments, we found that using a number of joints of $N = 11$ (excluding the hip and shoulder joints) produce the best performance for our system. Consequently, a posture feature is made by 33 attributes per skeleton. The next section describes how the most informative postures are selected from the entire activity sequence.

3.1.2 Posture selection: The purpose of this phase is to group, for each activity, the skeletons that share similar features to obtain a few basic postures. Therefore, a clustering algorithm is applied to find the centroids describing these groups. Nevertheless, some activities can be more complex than others; thus, the optimal number of clusters may vary according to this complexity. For this reason, the input sequence is clustered multiple times, ranging from 2 to 5, using the K-means algorithm [52] and varying the number of clusters (K) to generate multiple samples representing the same activity. At the end of this phase, the set of posture features f_{X_i} representing an activity sequence is replaced by the centroid to which the posture feature belongs. Hence, the centroids can be seen as the key poses of the activity.

3.1.3 Activity feature generation: The activity feature generation step is the core of the system. The same implementation is used for both the model generation and the classification phase. The aim of this step is to properly represent an activity by means of suitable features. The output of the posture selection module (see Section 3.1.2) contains a temporally ordered sequence of the centroids representing the most important postures of the original input. In this phase, all the equal centroids that are temporally consecutive are discarded. This means that the temporal sequence is simplified to include only the transitions between clusters. The obtained representation is more compact, lowering the overall complexity. In addition, it is speed invariant, which is an important property, considering that people perform activities at different speeds. At this point, the obtained sequence is still dependent to the current input sample in terms of the posture transitions. The idea is to exploit all the information contained in this sequence to represent a specific action in a way that is as general as possible using only a few basic postures. Therefore, we generate n-tuple from the current sequence by means of a sliding window, creating a set of new instances to represent the specific activity. Consequently, the new instances are composed of a set of features with a size of $3LN$, where L is the length of the sliding window and N is the number of selected skeleton joints. For instance, if an activity has three clusters, a possible compressed sequence can be

$$A = [C_1, C_3, C_2, C_3, C_2, C_3, C_2, C_3, C_2] \tag{4}$$



A1 = [C1, C3, C2, C3, C2]
A2 = [C3, C2, C3, C2, C3]
A3 = [C2, C3, C2, C3, C2]
= A3
= A2
A3
A2
A1
A = [C1, C3, C2, C3, C2, C3, C2, C3, C2]

Fig. 2 *Example of activity feature instances using a sliding window of L = 5 elements. Duplicates increase the instance weight*

A sliding window with a size of $L = 5$ elements produces three activity feature instances as depicted in Fig. 2. Each instance has a weight that is increased if there is a duplicate.

Fig. 3 shows a feature example of the handshake activity using a window length of 5 elements and skeletons with 11 joints (shoulders and hips omitted). The cardinality of the instances is related to the number of different transitions between the different key poses. This means that actions which are repetitive during the time will have fewer feature instances than the ones with more variability between key poses. The output of this module is a new dataset containing the activity features whose instances have a weight that is increased if the same sub-sequence is already present. At the end of this step, each input activity is represented with several new activity features generated from the different sets of basic clusters.

3.1.4 Training: In this last step, several multiclass SVMs [53] are trained using the activity features obtained in the previous step (see Section 3.1.3). Specifically, SVMs are supervised learning models used for binary classification to calculate the optimal hyperplane that separates two classes in the feature space. SVMs can perform a non-linear classification efficiently using what is called the kernel trick, which implicitly maps their inputs into high-dimensional feature spaces. In our case, we employ the SVM with the radial basis function kernel that yields better performances compared with other kernel types. The multiclass version is implemented using a one-versus-one strategy. The output of this phase is the creation of different models related to the number, K, of clusters.

3.2 Online classification

The activity recognition system is tested by providing the input sequence piece-by-piece in a serial fashion. In other words, the input is fed to the algorithm without having the entire input available from the beginning. This is more realistic and is similar to an actual application. We refer to this approach as online classification. The general architecture of the classification phase is depicted in Fig. 1 (bottom). The skeleton feature extraction and the activity feature generation modules are shared with the model generation phase. The online classification primarily differs in the clustering step, which is dynamic, and, of course, does not generate the models (i.e. it does not perform any training).

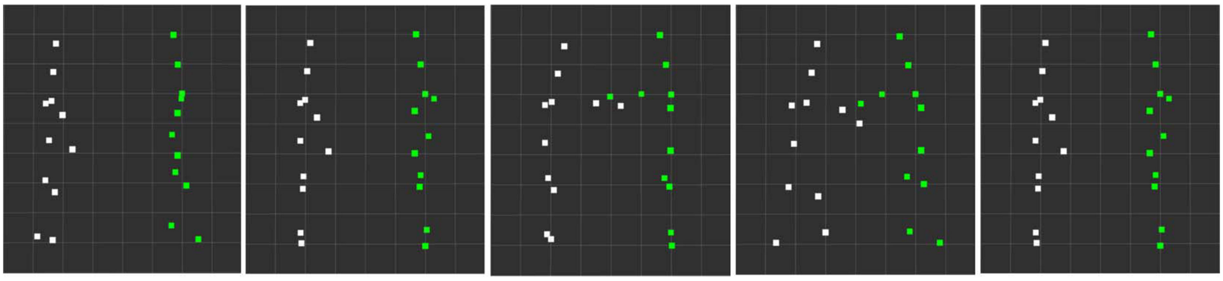
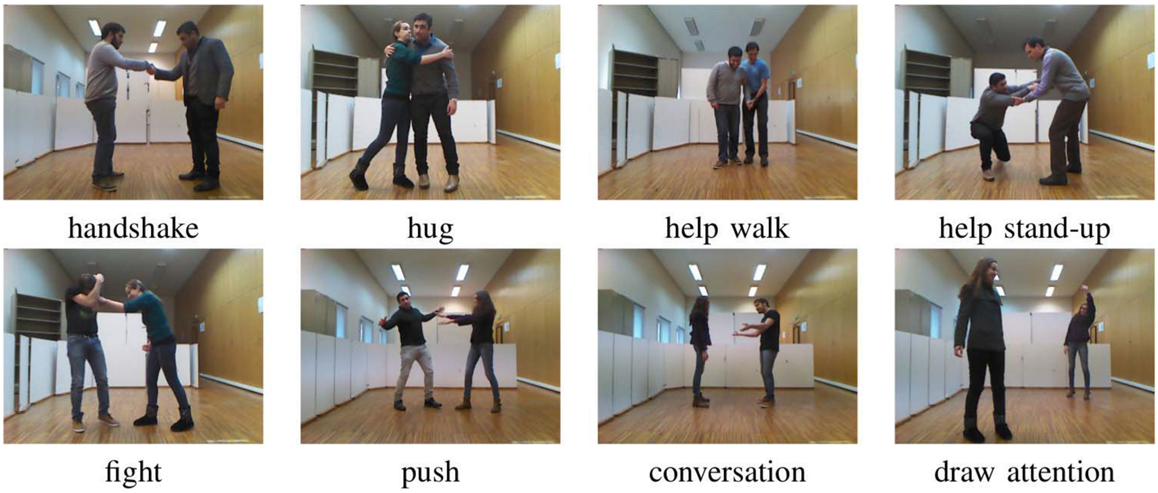


Fig. 3 *Instance example of an activity feature using a window length equal to 5 and a skeleton of 11 joints*

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)



handshake

hug

help walk

help stand-up

fight

push

conversation

draw attention

Fig. 4 *Samples of the ISR-UoL 3D social activity dataset*

3.2.1 Dynamic clustering: The core of this phase is to dynamically cluster the input sequence to use a different number of centroids to describe the activity, considering the heterogeneity of the action types. The aim of this step is to find the optimal number of clusters for the current data sequence that will be used to retrieve the relative pre-trained model. In this phase, the system cannot use the *K*-means, because it does not know a priori the number of clusters for the actual sequence. Hence, the optimal number of clusters is obtained using the *X*-means algorithm [54], which is an optimised version of the *K*-means that does not need to know a priori the number of classes. It attempts to split the centres into regions and to select the number of clusters using a probabilistic scheme called Bayes information criterion. The *X*-means is much faster rather than repeatedly running *K*-means using different numbers of clusters. Thus, *X*-means has proved to be less prone to local minima than its counterpart [55]. For each activity, the *X*-means is applied using the Euclidean distance function as a metric. In detail, given an activity composed of *M* posture features ($\boldsymbol{f}_1$, $\boldsymbol{f}_2$, …, $\boldsymbol{f}_M$), the *X*-means gives *k* clusters ($\boldsymbol{C}_1$, $\boldsymbol{C}_2$, …, $\boldsymbol{C}_k$), so as to minimise the intra-cluster sum of squares

$$\arg\min_{\boldsymbol{C}} \sum_{j=1}^{k} \sum_{f_i \in C_j} \| \boldsymbol{f}_i - \mu_j \|^2 \qquad (5)$$

where μ_j is the mean value of the cluster $\boldsymbol{C}_j$.

3.2.2 Classification: The classification step takes as input the activity features generated from the clustered data (see Section 3.1.3) and the trained model relative to the number of clusters *k* obtained from the dynamic clustering step (see Section 3.2.1). Each generated activity instance is classified with the same classifier as the training phase (Section 3.1.4), and each observed class is summed. Hence, the final result is the class that has the greater value; therefore, it incorporates all the classification results for each feature instance.

4 Experimental results

The system is implemented in Java using the Weka library [56], which is an open-source software containing a collection of machine learning algorithms for data mining tasks. The system is tested on the ISR-UoL 3D social activity dataset [23], and on the SBU Kinect interaction dataset, both containing skeleton data of two-interacting persons. The activity recognition system is evaluated following the procedure described in the paper [47] for the first dataset and in [24] for the second dataset. During the generation of the models, the input skeleton data are mirrored on the sagittal plane to increase the generality of the samples. The tests are conducted using different subsets of skeleton joints, and varying the length of the sliding window during the activity feature generation process, ranging from 5 to 12 elements. Furthermore, we compare the results obtained with three different classification strategies according to the feature used: independent, fusion at decision level, and fusion at the feature level.

4.1 Datasets

The adopted datasets contain eight actions of two-interacting persons. The ISR-UoL dataset contains longer sequences, from 40 to 60 s, whereas the SBU has short segmented actions of ∼4 s. Some of these actions are present in both of them: *handshake*, *hug*, and *push*. The SBU dataset has *punching* and *kicking* actions, whereas the ISR has a general *fight* interaction. The actions contained in the ISR datasets are more complex and realistic. Furthermore, it also includes *call attention*, *help walk*, and *help stand-up* particularly suitable for AAL applications.

4.1.1 ISR-UoL 3D social activity dataset: The ISR-UoL 3D social activity dataset [23] incorporates interaction between two subjects. This dataset consists of RGB and depth images, and tracked skeleton data acquired by an RGB-D sensor. It includes eight social activities: *handshake, greeting hug, help walk, help stand-up, fight, push, conversation, and call attention* (see Fig. 4). Each activity is recorded in a period of ∼40–60 s of repetitions within the same session, at a frame rate of 30 fps. The only exceptions are *help walk* (at a short distance) and *help stand-up*, which is recorded four times as the same session, regardless of the time spent on it. The activities are selected to address the AL scenario (e.g. happening in a health-care environment: *help walk*, *help stand-up*, and *call attention*), with potentially harmful situations such as aggression (e.g. *fight*, *push*), and casual activities of social interactions (e.g. *handshake*, greeting *hug*, and *conversation*). The activities are performed by six persons, four males, and two females, with an average age of 29.7 ± 4.2, from different nationalities. A total of ten different combinations of individuals (or sessions) is presented, with variation of the roles (active or passive person) between the subjects. Each subject has participated in at least three combinations, acting each role at least once to increase the generalisation of the study regarding individual behaviour. The dataset is evaluated using the leave-one-out cross-validation method, given ten sessions and eight activities [47]. The dataset contains approximately more than 120,000 data frames.

4.1.2 SBU Kinect interaction dataset: The SBU Kinect interaction dataset consists of RGB, depth images, and tracked skeleton data acquired by an RGB-D sensor. It includes eight activities: *approaching, departing, pushing, kicking, punching, exchanging objects, hugging,* and *shaking hands* (see Fig. 5). The dataset is composed of 21 sets, where each set contains data of different persons (7 participants) performing the actions with a frame rate of 15 frames per second. The dataset is composed of manually segmented videos for each interaction (∼4 s), but each

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)

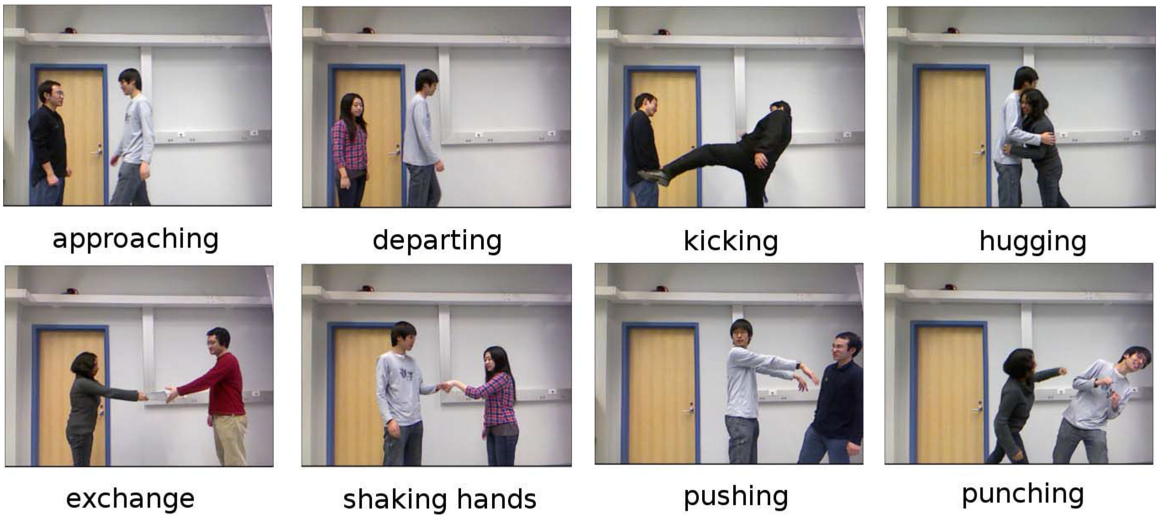


Fig. 5 *Samples of the SBU Kinect interaction dataset*

video roughly starts from a standing pose before acting and ends with a standing pose after acting. The evaluation is done by five-fold cross-validation, i.e. four folds are used for training and 1 for testing. The partitioning of the datasets into folds is performed, so that each two-actor set is guaranteed to appear only in training or only in testing [24]. The dataset contains ∼10,000 data frames.

4.2 Classification results

According to the skeleton features used in the first module of the system, we adopt three classification strategies. The first strategy is person independent, the second employs a fusion of the features at decision level, and the last uses the two-person skeletons from the beginning of the entire process. In details:

Independent: The skeletons of the active and passive persons are processed and classified separately. This means that the system takes as inputs $\boldsymbol{f}_A$ and $\boldsymbol{f}_B$ in (1) independently.

Fusion at decision level: The decision-level fusion [57] combines the models of the independent classification. The results are used to train a supervised classifier, i.e. the *k*-nearest neighbours (*k*-NNs) [58] algorithm, yielding the final classification (see Fig. 6).

Fusion at feature level: The feature-level strategy combines the features at the early stage of the process [8]. It uses a skeleton feature that is the combination of the two feature vectors of the input sample ($\boldsymbol{f}_A$ and $\boldsymbol{f}_B$ at the beginning).

All the three classification strategies are evaluated varying two parameters, the skeleton joints and the length of the sliding window applied to the generated activity features. In particular, to find the most informative joints for the actions, the system is tested using all joints ($N = 15$), removing shoulder and hip ($N = 11$), and removing also elbow and knee joints ($N = 7$). For each configuration, different lengths for the sliding window are adopted, ranging from 5 to 12 elements. Since both datasets have unbalanced classes, the performances are evaluated in terms of accuracy, calculated taking into account the number of class instances. The average accuracy of the evaluation sessions are reported in Table 1 for the ISR dataset and in Table 2 for the SBU dataset. Looking at these results, it is possible to note how the worst classification performances are obtained in the independent use case. This consideration is somewhat expected since the data of the active and passive skeletons alone are not enough to classify the activities. It is interesting to note, how the performances of the independent classification of the ISR dataset is far better than the SBU dataset. This can be explained by the fact that the samples of the first dataset are longer than the second, producing more activity features. The fusion at decision-level scheme, which combines the results of the independent classifiers, slightly improves the accuracy of the ISR dataset. Conversely, the performances are worst using the SBU dataset. In this case, in fact, the results of the independent case are too low and the fusion at decision level does not produce any advantages. On the contrary, the fusion at feature-level scheme proves to be the best approach for both datasets. The accuracy greatly improves in all the test configurations, achieving comparable results on all the different joint and sliding window settings. The best performances are obtained using 11 skeleton joints and a sliding window with six elements. In this configuration, the running time for the ISR samples are ∼100 ms for the shortest sequences and ∼300 ms for the longest on an Intel i7 with 2.4 GHz quad-core processors, whereas the SBU samples take <100 ms to be classified.

4.3 Discussion

Tables 3 and 4 report the confusion matrices of the best configuration ($N = 11$, $L = 6$) for the ISR and SBU datasets, respectively. The values are normalised between 0 and 1, while the rightest column reports the total number of test samples according to the cross-validation method used for the dataset. Considering the first dataset, the developed system reaches 0.80, 0.80, and 0.87 values of average precision, recall, and accuracy. The classification system implemented in [47], which implements the activity recognition using a probabilistic ensemble of classifiers called DBMM and proximity priors, reaches a slightly better value of precision and recall (0.85). However, the average accuracy of our system, calculated considering the total number of activity instances, has good performances (0.87). Comparing the results of the two methods, we can note that our system obtains better performances on activities related to AL context, i.e. *help stand-up* (0.95 versus 0.82), *help walk* (1 versus 0.98), *hug* (1 versus 0.78), and *draw attention* (1 versus 0.97). In addition, the *push* activity is often misclassified as *fight*, because of their similarity. The worst case is represented by the *conversation* activity, which is the most static of the dataset.

Regarding the SBU dataset, the classification system implemented in [24], which trains a MILBoost classifier using specific joint features, reaches an average accuracy of 0.87 on the same dataset, comparable with our system. A deeper comparison is not possible because precision and recall values and confusion matrices are not provided. In general, all the activities have a high true positive rate, while *punching* is misclassified with *pushing*, and *shaking hands* with *exchanging*, or *punching*.

In general, the developed activity recognition system has good classification performances comparable with the state of the art, even if the developed system is much simpler. Static actions are the most difficult to classify because they generate less informative activity features. On the other hand, the system reaches optimal results with interactions of AL context such as *help stand-up*, *help walk*, and *draw attention*.

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)

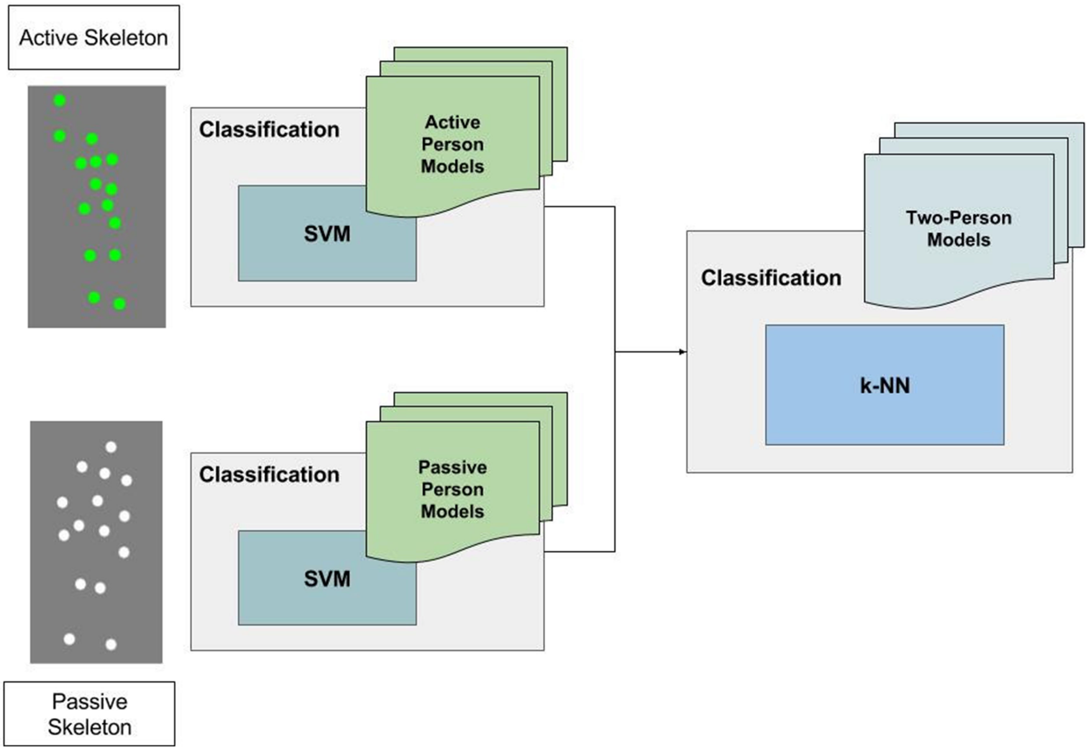


Fig. 6 *Fusion at decision level combines the models of the independent classification by training a k-NN classifier*

Table 1 Average accuracy for each configuration of the different classification strategies for the ISR dataset

Window\joints	Active user			Passive user			Decision level			Feature level		
	7	11	15	7	11	15	7	11	15	7	11	15
5	0.68	0.69	0.72	0.70	0.71	0.72	0.74	0.73	0.73	0.84	0.85	0.85
6	0.67	0.70	0.71	0.70	0.72	0.73	0.74	0.74	0.74	0.85	**0.87**	0.85
7	0.66	0.67	0.71	0.67	0.74	0.74	0.72	0.75	0.74	0.85	0.86	0.86
8	0.66	0.66	0.71	0.67	0.72	0.72	0.74	0.74	0.72	0.83	0.86	0.84
9	0.68	0.64	0.69	0.67	0.71	0.70	0.76	0.74	0.71	0.86	0.86	0.86
10	0.65	0.63	0.70	0.68	0.70	0.70	0.75	0.74	0.70	0.85	0.86	0.86
11	0.68	0.65	0.70	0.68	0.71	0.69	0.78	0.73	0.70	0.85	0.86	0.86
12	0.68	0.63	0.68	0.65	0.72	0.69	0.78	0.75	0.68	0.85	0.86	0.85

Bold values represent the best accuracy obtained with the configuration

Table 2 Average accuracy for each configuration of the different classification strategies for the SBU dataset

Window\Joints	Active user			Passive user			Decision level			Feature level		
	7	11	15	7	11	15	7	11	15	7	11	15
5	0.58	0.56	0.60	0.37	0.36	0.37	0.50	0.54	0.31	0.87	0.86	0.87
6	0.57	0.57	0.58	0.40	0.38	0.35	0.50	0.52	0.42	0.88	**0.88**	0.88
7	0.57	0.58	0.61	0.43	0.39	0.36	0.44	0.47	0.48	0.87	0.86	0.84
8	0.55	0.56	0.60	0.43	0.39	0.40	0.47	0.45	0.41	0.88	0.87	0.86
9	0.56	0.57	0.59	0.41	0.37	0.37	0.50	0.50	0.36	0.88	0.87	0.86
10	0.56	0.56	0.60	0.40	0.38	0.36	0.45	0.50	0.33	0.87	0.87	0.84
11	0.55	0.56	0.60	0.42	0.38	0.37	0.44	0.50	0.47	0.88	0.87	0.85
12	0.56	0.54	0.57	0.43	0.37	0.37	0.48	0.52	0.41	0.88	0.86	0.85

Bold values represent the best accuracy obtained with the configuration

5 Conclusion

This paper describes an activity recognition system for two-interacting persons using skeleton data extracted from a depth camera. The activity sequence is encoded using a few basic informative postures, which are computed using a clustering approach. Differently from state-of-the-art methods, the system uses few number of clusters to calculate the features, ranging from 2 to 5 classes. During the training phase, the system creates several classification models according to a different number of clusters, whereas in the classification process the relative pre-trained model is retrieved by dynamically calculating the optimal number of clusters for current unseen data. In this way, the activities can be modelled precisely.

The system is tested on two public datasets. The ISR dataset contains long and complex activity sequences, whereas the SBU has shorter samples. Best results are obtained employing a fusion at feature-level approach, producing an overall accuracy of 0.87 for the first dataset and 0.88 for the second one. The running time of the classification process takes from 100 to 300 ms according to the length of the input sequence. Results are comparable with the state of the art, even if the developed system employs a simpler

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)

Table 3 Confusion matrix of the fusion at feature level for the ISR dataset, using $N = 11$ skeleton joints, and a window length of $L = 6$ elements. The rightest column reports the total number of activity samples for the leave-one-subject-out evaluation

conversation	*0.30*	0.00	**0.10**	*0.20*	0.00	0.00	**0.10**	*0.30*	10
draw attention	0.00	***1.00***	0.00	0.00	0.00	0.00	0.00	0.00	10
fight	0.00	**0.06**	***0.75***	**0.06**	0.00	0.00	**0.06**	**0.07**	16
handshake	**0.10**	0.00	**0.10**	***0.70***	0.00	0.00	**0.10**	0.00	10
help stand-up	0.00	**0.02**	0.00	0.00	***0.95***	0.00	**0.03**	0.00	40
help walk	0.00	0.00	0.00	0.00	0.00	***1.00***	0.00	0.00	45
hug	0.00	0.00	0.00	0.00	0.00	0.00	***1.00***	0.00	11
push	0.00	0.00	*0.22*	**0.07**	0.00	0.00	0.00	***0.71***	14
	conversation	draw attention	fight	handshake	help standing	help walk	hug	push	

Precision 0.80, recall 0.80, and accuracy 0.87.

Bold italic numbers are the values on the diagonal of the matrix.

Bold numbers are the values not belonging to the diagonal.

Table 4 Confusion matrix of the fusion at feature level for the SBU dataset, using $N = 11$ skeleton joints, and a window length of $L = 6$ elements. The rightest column reports the total number of activity samples for the five-fold evaluation. Precision 0.86, recall 0.87, and accuracy 0.88

approaching	***0.91***	0.00	0.00	0.00	0.00	**0.02**	*0.05*	**0.02**	42
departing	0.00	***0.98***	0.00	0.00	0.00	0.00	**0.02**	0.00	42
exchanging	0.00	0.00	***0.82***	0.00	**0.05**	0.00	0.00	*0.13*	38
hugging	0.00	0.00	0.00	***0.95***	0.00	0.00	**0.05**	0.00	21
kicking	0.00	0.00	**0.02**	0.00	***0.95***	0.00	0.00	**0.03**	38
punching	**0.03**	0.00	**0.05**	**0.03**	0.00	***0.76***	*0.13*	0.00	37
pushing	0.00	0.00	0.00	**0.02**	0.00	**0.10**	***0.88***	0.00	40
shaking hands	0.00	0.00	*0.11*	0.00	0.00	*0.17*	0.00	***0.72***	18
	approaching	departing	exchanging	hugging	kicking	punching	pushing	shaking hands	

Bold italic numbers are the values on the diagonal of the matrix.

Bold numbers are the values not belonging to the diagonal.

approach. In particular, significant improvements involve the activities related to the health-care environments, making feasible the possible application in AL scenarios. It is worth to note that the use of human skeletons acquired from depth cameras is not affected by environmental light variations, ensuring at the same time a higher level of user privacy compared with standard video cameras. Furthermore, the system is more general, and not tailored to the two-person interactions, since it has been previously tested on single action datasets [22]. The current implementation of the system needs to know the role of the persons, i.e. the active and passive users. For this reason, future works will investigate suitable and robust methodologies to solve this issue. In addition, it can be observed that the samples of the SBU dataset are very short, whereas the ISR dataset contains longer actions with more variability. Therefore, future works will focus on the development of a proper action segmentation procedure to be used as preprocessing step that can enhance the classification performances.

6 Acknowledgments

This research was pursued in part within the project ‘CENTAURO – Colavoro, Efficienza, preveNzione nell'industria dei motoveicoli mediante Tecnologie di AUtomazione e Robotica’ supported by a grant from Regione Toscana, Call FAR-FAS 2014, and in part by the ACCRA Project, funded by the European Community's Horizon 2020 Programme (H2020-SCI-PM14-2016) – agreement number 738251.

7 References

[1] Aquilano, M., Cavallo, F., Bonaccorsi, M., *et al.*: ‘Ambient assisted living and ageing: preliminary results of RITA project’. 2012 Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC), 2012, pp. 5823–5826

[2] Garcia, N.M., Rodrigues, J.J.P.C.: ‘*Ambient assisted living*’ (CRC Press, 2015)

[3] Laitinen, A., Niemela, M., Pirhonen, J.: ‘Social robotics, elderly care, and human dignity: a recognition-theoretical approach’, 2016

[4] Portugal, D., Trindade, P., Christodoulou, E., *et al.*: ‘On the development of a service robot for social interaction with the elderly’, 2015

[5] Cesta, A., Cortellessa, G., De Benedictis, R., *et al.*: ‘Supporting active and healthy ageing by exploiting a telepresence robot and personalized delivery of information’. Int. Conf. Intelligent Software Methodologies, Tools, and Techniques, 2015, pp. 586–597

[6] Fiorini, L., Esposito, R., Bonaccorsi, M., *et al.*: ‘Enabling personalised medical support for chronic disease management through a hybrid robot-cloud approach’, *Auton. Robots*, 2017, **41**, (5), pp. 1263–1276

[7] Yürür, Ö., Liu, C.H., Sheng, Z., *et al.*: ‘Context-awareness for mobile sensing: a survey and future directions’, *IEEE Commun. Surv. Tutor.*, 2016, **18**, (1), pp. 68–93

[8] Vrigkas, M., Nikou, C., Kakadiaris, I.A.: ‘A review of human activity recognition methods’, *Front. Robot. AI*, 2015, **2**, p. 28

[9] Dong, Y., Scisco, J., Wilson, M., *et al.*: ‘Detecting periods of eating during free-living by tracking wrist motion’, *IEEE J. Biomed. Health Inf.*, 2014, **18**, (4), pp. 1253–1260

[10] Xiao, Y., Zhang, Z., Beck, A., *et al.*: ‘Human–robot interaction by understanding upper body gestures’, *Presence, Teleoperators Virtual Environ.*, 2014, **23**, (2), pp. 133–154

[11] Picard, R.W.: ‘*Affective computing*’ (MIT Press, Cambridge, MA, USA, 1997)

[12] Vinciarelli, A., Pentland, A.S.: ‘New social signals in a new interaction world: the next frontier for social signal processing’, *IEEE Syst. Man Cybern. Mag.*, 2015, **1**, (2), pp. 10–17

[13] Vázquez, M., Steinfeld, A., Hudson, S.E.: ‘Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation’. 2015 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), 2015, pp. 3010–3017

[14] Adelman, R.D., Tmanova, L.L., Delgado, D., *et al.*: ‘Caregiver burden: a clinical review’, *Jama*, 2014, **311**, (10), pp. 1052–1060

[15] Kellokumpu, V., Pietikäinen, M., Heikkilä, J.: ‘Human activity recognition using sequences of postures’. MVA, 2005, pp. 570–573

[16] Willems, G., Tuytelaars, T., Van Gool, L.: ‘An efficient dense and scale-invariant spatio-temporal interest point detector’. European Conf. Computer Vision, 2008, pp. 650–663

[17] Aggarwal, J.K., Ryoo, M.S.: ‘Human activity analysis: a review’, *ACM Comput. Surv. (CSUR)*, 2011, **43**, (3), p. 16

[18] Weinland, D., Ronfard, R., Boyer, E.: ‘A survey of vision-based methods for action representation, segmentation and recognition’, *Comput. Vis. Image Underst.*, 2011, **115**, (2), pp. 224–241

[19] Shotton, J., Sharp, T., Kipman, A., *et al.*: ‘Real-time human pose recognition in parts from single depth images’, *Commun. ACM*, 2013, **56**, (1), pp. 116–124

[20] Turchetti, G., Micera, S., Cavallo, F., *et al.*: ‘Technology and innovative services’, *IEEE Pulse*, 2011, **2**, (2), pp. 27–35

[21] Cavallo, F., Aquilano, M., Bonaccorsi, M., *et al.*: ‘Multidisciplinary approach for developing a new robotic system for domiciliary assistance to elderly people’. 2011 Annual Int. Conf. IEEE Engineering in Medicine and Biology Society, 2011, pp. 5327–5330

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)

[22] Manzi, A., Cavallo, F., Dario, P.: '*A 3D human posture approach for activity recognition based on depth camera*' (Springer International Publishing, Cham, 2016), pp. 432–447

[23] ISR-UoL 3D Social Activity Dataset. Available at https://lcas.lincoln.ac.uk/wp/isr-uol-3d-social-activity-dataset, accessed June 2017

[24] Yun, K., Honorio, J., Chattopadhyay, D.*, et al.*: 'Two-person interaction detection using body-pose features and multiple instance learning'. 2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 28–35

[25] Lara, O.D., Labrador, M.A.: 'A survey on human activity recognition using wearable sensors', *IEEE Commun. Surv. Tutor.*, 2013, **15**, (3), pp. 1192–1209

[26] Su, X., Tong, H., Ji, P.: 'Activity recognition with smartphone sensors', *Tsinghua Sci. Technol.*, 2014, **19**, (3), pp. 235–249

[27] Aggarwal, J.K., Xia, L.: 'Human activity recognition from 3D data: a review', *Pattern Recognit. Lett.*, 2014, **48**, pp. 70–80

[28] Sung, J., Ponce, C., Selman, B.*, et al.*: 'Unstructured human activity detection from RGBD images'. 2012 IEEE Int. Conf. Robotics and Automation (ICRA), 2012, pp. 842–849

[29] Koppula, H.S., Gupta, R., Saxena, A.: 'Learning human activities and object affordances from RGB-D videos', *Int. J. Robot. Res.*, 2013, **32**, (8), pp. 951–970

[30] Ni, B., Wang, G., Moulin, P.: 'RGBD-HuDaAct: a color-depth video database for human daily activity recognition'. Consumer Depth Cameras for Computer Vision, 2013, pp. 193–208

[31] Wang, J., Liu, Z., Wu, Y.*, et al.*: 'Mining actionlet ensemble for action recognition with depth cameras'. 2012 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1290–1297

[32] Ryoo, M.S., Matthies, L.: 'First-person activity recognition: what are they doing to me?'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2013, pp. 2730–2737

[33] Chen, L., Wei, H., Ferryman, J.: 'A survey of human motion analysis using depth imagery', *Pattern Recognit. Lett.*, 2013, **34**, (15), pp. 1995–2006

[34] Rehg, J., Abowd, G., Rozga, A.*, et al.*: 'Decoding children's social behavior'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2013, pp. 3414–3421

[35] Kong, Y., Jia, Y., Fu, Y.: 'Learning human interaction by interactive phrases'. Computer Vision – ECCV 2012, 2012, pp. 300–313

[36] Raptis, M., Sigal, L.: 'Poselet key-framing: a model for human activity recognition'. Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition, CVPR '13, Washington, DC, USA, 2013, pp. 2650–2657

[37] Huang, D.-A., Kitani, K.M.: 'Action-reaction: forecasting the dynamics of human interaction'. ECCV (7), 2014, pp. 489–504

[38] Blunsden, S., Fisher, R.B.: 'The behave video dataset: ground truthed video for multi-person behavior classification', *Ann. BMVA*, 2010, **4**, (1–12), p. 4

[39] Lopez, L.D., Reschke, P.J., Knothe, J.M.*, et al.*: 'Postural communication of emotion: perception of distinct poses of five discrete emotions', *Front. Psychol.*, 2017, **8**, p. 710. PMC. Web. 13 Oct. 2017

[40] Cheng, Z., Qin, L., Huang, Q.*, et al.*: 'Recognizing human group action by layered model with multiple cues', *Neurocomputing*, 2014, **136**, pp. 124–135

[41] Presti, L.L., Cascia, M.L.: '3D skeleton-based human action classification: a survey', *Pattern Recognit.*, 2016, **53**, pp. 130–147

[42] Ibrahim, M.S., Muralidharan, S., Deng, Z.*, et al.*: 'A hierarchical deep temporal model for group activity recognition', 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2015, pp. 1971–1980. doi: 10.1109/CVPR.2016.217

[43] Choi, W., Shahid, K., Savarese, S.: 'What are they doing?: collective activity classification using spatio-temporal relationship among people'. 2009 IEEE 12th Int. Conf. Computer Vision Workshops (ICCV Workshops), 2009, pp. 1282–1289

[44] Tran, K.N., Gala, A., Kakadiaris, I.A.*, et al.*: 'Activity analysis in crowded environments using social cues for group discovery and human interaction modeling', *Pattern Recognit. Lett.*, 2014, **44**, pp. 49–57

[45] Shahroudy, A., Liu, J., Ng, T.-T.*, et al.*: 'NTU RGB + D: a large scale dataset for 3D human activity analysis'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 1010–1019

[46] SBU Kinect interaction dataset v2.0. Available at http://www3.cs.stonybrook.edu/kyun/research/kinect_interaction/. accessed June 2017

[47] Coppola, C., Faria, D.R., Nunes, U.*, et al.*: 'Social activity recognition based on probabilistic merging of skeleton features with proximity priors from RGB-D data'. 2016 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS, 2016, pp. 5055–5061

[48] Faria, D.R., Premebida, C., Nunes, U.: 'A probabilistic approach for human everyday activities recognition using body motion from RGB-D images'. 23rd IEEE Int. Symp. Robot and Human Interactive Communication, 2014 RO-MAN, 2014, pp. 732–737

[49] Cippitelli, E., Gasparrini, S., Gambi, E.*, et al.*: 'A human activity recognition system using skeleton data from RGBD sensors', *Comput. Intell. Neurosci.*, 2016, **2016**, p. 21

[50] Gaglio, S., Re, G.L., Morana, M.: 'Human activity recognition process using 3D posture data', *IEEE Trans. Hum.–Mach. Syst.*, 2015, **45**, (5), pp. 586–597

[51] Shan, J., Akella, S.: '3D human action segmentation and recognition using pose kinetic energy'. 2014 IEEE Int. Workshop on Advanced Robotics and its Social Impacts, 2014, pp. 69–75

[52] MacQueen, J.: 'Some methods for classification and analysis of multivariate observations'. Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, CA, 1967, pp. 281–297

[53] Chang, C.-C., Lin, C.-J.: 'LIBSVM: a library for support vector machines', *ACM Trans. Intell. Syst. Technol.*, 2011, **2**, pp. 27 : 1–27 : 27. Accessed on June 2017. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

[54] Pelleg, D., Moore, A.W.: '*X*-means: extending *k*-means with efficient estimation of the number of clusters'. 17th Int. Conf. Machine Learning, 2000, pp. 727–734

[55] Witten, I.H., Frank, E., Hall, M.A.: '*Data mining: practical machine learning tools and techniques*' (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011, 3rd edn.)

[56] Hall, M., Frank, E., Holmes, G.*, et al.*: 'The Weka data mining software: an update', *SIGKDD Explor. Newsl.*, 2009, **11**, (1), pp. 10–18

[57] Jiang, B., Martinez, B., Valstar, M.F.*, et al.*: 'Decision level fusion of domain specific regions for facial action recognition'. 2014 22nd Int. Conf. Pattern Recognition (ICPR), 2014, pp. 1776–1781

[58] Aha, D., Kibler, D.: 'Instance-based learning algorithms', *Mach. Learn.*, 1991, **6**, pp. 37–66

This is an open access article published by the IET under the Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/)