A 3D Human Posture Approach for Activity Recognition Based on Depth Camera

Alessandro Manzi(✉), Filippo Cavallo, and Paolo Dario

The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
{a.manzi,f.cavallo,p.dario}@sssup.it

Abstract. Human activity recognition plays an important role in the context of Ambient Assisted Living (AAL), providing useful tools to improve people quality of life. This work presents an activity recognition algorithm based on the extraction of skeleton joints from a depth camera. The system describes an activity using a set of few and basic postures extracted by means of the X-means clustering algorithm. A multi-class Support Vector Machine, trained with the Sequential Minimal Optimization is employed to perform the classification. The system is evaluated on two public datasets for activity recognition which have different skeleton models, the CAD-60 with 15 joints and the TST with 25 joints. The proposed approach achieves precision/recall performances of 99.8 % on CAD-60 and 97.2 %/91.7 % on TST. The results are promising for an applied use in the context of AAL.

Keywords: Activity monitoring systems · Human activity recognition · Depth camera · RGB-D camera · Ambient Assisted Living · Assistive technologies

1 Introduction

Human activity recognition is one of the most important areas of computer vision research today. It can be described as the spatiotemporal evolutions of different body postures and its main goal is to automatically detect human activities analyzing data from various types of devices (e.g. color cameras or range sensors). Regarding the assistive technologies, the possibility of application is really wide, in particular, they can include surveillance and monitoring systems, and a large range of applications involving human-machine interactions [1]. Although the recognition of human actions is very important for many real applications, it is still a challenging problem. In the past, the research has mainly focused on recognizing activities from video sequences by means of color cameras. However, capturing articulated human motion from monocular video sensors results in a considerable loss of information [2]. These solutions are often constrained in terms of computational efficiency and robustness to illumination changes [3]. Another approach is to use 3D data from marker-based motion capture or stereo camera systems, i.e. capturing 2D image sequences from multiple views to reconstruct 3D information [4]. Nowadays, the use of depth cameras has become very

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 432–447, 2016.
DOI: 10.1007/978-3-319-48881-3_30

popular, because the technological progress has made available devices that are cost effective providing also 3D data at suitable resolution rate. Recently, this kind of sensors has led several new works on activity recognition from depth data [5,6]. These inexpensive devices, such as Microsoft Kinect or Asus Xtion, allow capturing both color and depth information. Moreover, specific tracker software can efficiently detect the human skeleton directly from the depth maps [7]. These features can be exploited in order to develop effective solutions for Ambient Assistive Living applications, simplify the problem of human detection and segmentation. In particular, the depth maps are not affected by environment light variations and, at the same time, they guarantee the user privacy more than the color information [8]. Although there are several types of research on this topic, challenges still remain on how to process these kinds of data to effectively detect actions in real world scenarios.

The present work describes a human activity recognition system based on skeleton data extracted from a depth camera. The activity is represented with few and basic postures obtained with the X-means clustering algorithm. The Sequential Minimal Optimization (SMO) is used to train a multiclass SVM in order to classify the different activities. The system is trained on two publicly available datasets, the CAD-60, widely used for activity recognition and the TST, a dataset created for assistive technologies. The performances outperform the state-of-the-art on these datasets making the system suitable for Ambient Assistive Living real applications.

The remainder of the paper is organized as follows. Section 2 summarizes activity recognition methods, while the Sect. 3 details the developed system. The experimental results are presented in the Sect. 4 and Sect. 5 concludes the paper.

2 Related Works

In the past, human activities recognition methods have been focused on the processing of color images from traditional cameras. Some authors focus on the extraction of the human silhouettes from images using HMMs [9,10] or SVMs [11] to classify different postures. The drawbacks of these methods are low robustness in the case of complex environments and light variations. Other approaches are based on the detection of scale-invariant spatiotemporal features [12]. These features are convenient to detect moving objects, but problems arise in the presence of multiple persons and dynamic background. Wearable sensors are another alternative for activity recognition [13,14]. Usually, they provide more accurate information, but they are too intrusive for most people.

Recently, depth cameras are become very popular on the activity recognition topic, because they offer several advantages over traditional video cameras. First of all, they are inexpensive devices able to work also in poor light conditions. In addition, the RGB-D cameras provide human skeleton information that can significantly simplify the task of human detection. Wang et al. [15] introduce the concept of actionlet, which is a particular sequence of features, so-called local

occupancy features. An activity is described as a combination of actionlet. In [16], action recognition is performed extracting human postures. The postures are represented as a bag of 3D points and the actions are modeled as a graph, whose nodes are the extracted postures. Sung et al. [17] represent an activity as a set of subactivities, which is modeled using more than 700 features computing the Histogram of Oriented Gradient both on color images and on depth maps. A hierarchical maximum entropy Markov model is used to associate sub-activities with an high-level activity. Other authors focus on the use of multimodal features, i.e. combining color and depth information [18,19]. Space-Time Occupancy Patterns is proposed in [20] that divides space and time axes in multiple segments in order to embed a depth map sequence in multiple 4D grids. In [21] the EigenJoints feature descriptor is proposed, combining static posture, motion property and dynamics. They use motion energy to select the informative frame and use a Naive-Bayes-Nearest-Neighbor classifier for classification. Koppula et al. [22] use also object affordances and a Markov Random Field to represent their relationship with sub-activities. Zhu et al. [23] employ several spatio-temporal interest point features extracted from depth maps in combination of skeleton joints to classify actions with an SVM. These methods can reach good results, but usually, their performances depend on the complexity of the background scene and on the noise present on the depth data.

Other approaches use only the 3D human skeleton model to extract informative features to classify. Several joints representations have been proposed, Gan and Chen [24] propose the APJ3D representation computing the relative positions and local spherical angles from the skeleton joints. The HOJ3D, presented in [25], associates each joint to a particular area using a Gaussian weight function. The temporal evolution of the postures is modeled with a discrete Hidden Markov Model. Gaglio et al. [26] estimate the postures using a multiclass SVM and create an activity model using discrete HMM. Other works consider also trajectories of joints [27]. Some researchers focus on the selection of the most informative joints to improve the classification results [28,29]. In [30] a clustering method is applied to extract relevant features and a multiclass SVM is used for classification.

Looking at the aforementioned works, it is possible to understand that some authors try to extrapolate relevant features from multimodal data, while others exclusively rely on the human skeleton obtained from tracker software. Using more data not always yields better results, and sometimes simple solutions are preferable. The present work belongs to the latter case and it is based on the concept of informative postures known as “key poses”. This concept has been introduced in [31] and extensively used in the literature [32,33]. Some authors identify key poses calculating the kinetic energy [34,35] to segment an activity in static and dynamic poses. But not all the activities can be represented by alternating static and dynamic motions. Our approach is similar to [30], which uses clustering techniques to extrapolate the key poses, but conversely, our method represents an activity with a set of features based on few and basic informative postures.

3 Activity Recognition System

The aim of the implemented system is to infer the user activity using a combination of machine learning techniques. As already said in Sect. 1, we develop a system based only on the 3D skeleton data in order to deal with less number of features compared to color images. Moreover, using only skeleton data allows having much high privacy for the user. The idea is to describe an activity with a sequence of few and informative basic postures.

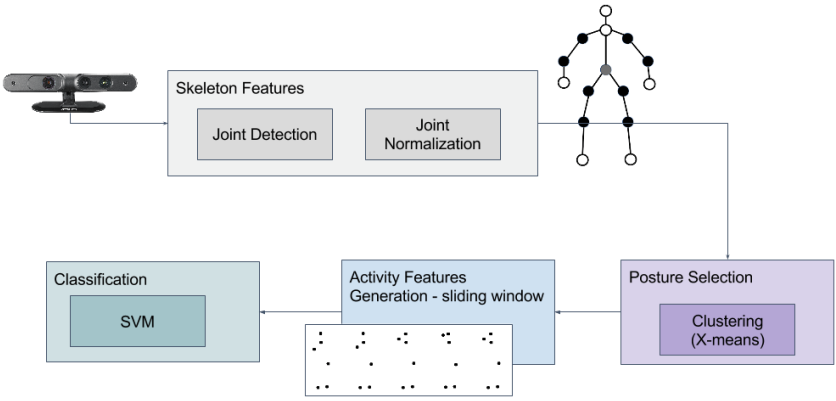

Fig. 1. The system is composed of 4 steps. The skeleton data are obtained from the sensor, and the posture features are processed. Informative postures are extracted from the sequence. Then, the activity features are generated from the basic postures. Finally, a classifier is applied.

The system is composed of four main steps (see Fig. 1). First of all, the relevant skeleton features (i.e. spatial joints) are extracted from the RGB-D device. Then, the basic and informative postures are selected using a clustering method. Afterwards, a new sequence of cluster centroids is built to have a temporal sequence of cluster transitions and an activity window is applied in order to create the activity feature. Finally, a classifier is trained to perform the recognition of the activity.

3.1 Posture Features

The coordinates of the human skeleton are extracted from the depth maps captured by the RGB-D sensor [7]. A human pose is represented by a number of joints that varies depending on the skeleton model of the software tracker (usually it can be 15, 20, or 25). The Fig. 2(a) shows a skeleton made of 15 joints. Each joint is described with three-dimensional Euclidean coordinates with respect to the sensor. The posture features are extracted directly from this skeleton model.

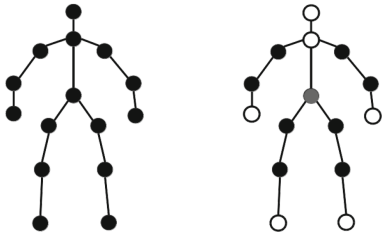

Fig. 2. Representation of the human skeleton: (a) Original 15 joints (b) Subset of joints for the posture feature (white), and the reference joint (torso) is in gray.

These raw data cannot be used directly, since they are dependent on the position of the human with respect to the sensor and also on the subject dimension, such as height and limb length. Therefore, the skeleton data are normalized following a widely used method [26,30,34]. The reference system is changed from the camera to the torso joint, in order to have data that are independent of the position of the sensor. Then, the joints are scaled with respect to the distance between the neck and the torso joint. This allows having data that are more independent with respect to the person specific size.

Formally, considering a skeleton with N joints, the skeleton feature vector $\mathbf{f}$ is defined as:

$$\mathbf{f} = [\ \mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_{N-1}\ ] \tag{1}$$

where each $\mathbf{j}_i$ is the vector containing the 3-D normalized coordinates of the ith joint $\mathbf{J}_i$ detected by the sensor. Therefore, $\mathbf{j}_i$ is defined as

$$\mathbf{j}_i = \frac{\mathbf{J}_i - \mathbf{J}_0}{\|\mathbf{J}_1 - \mathbf{J}_0\|}, \quad i = 1, 2, ..., N-1 \tag{2}$$

where $\mathbf{J}_0$ and $\mathbf{J}_1$ are the coordinates of the torso and the neck joint respectively. These normalized skeleton features can be seen as a set of distance vectors with respect to the torso joint and the number of attributes of the feature vector $\mathbf{f}$ in (1) is $3(N-1)$. Many works use a reduced set of joints since not all of them are really informative for the recognition of an activity. Moreover, using all the joints leads an increase of the complexity of the problem affecting the performances of the recognition phase. In Sect. 4, we will show results using different subsets of joints. However, our best results are obtained using a feature vector with $N = 7$, namely the head, the neck, the hands, and the feet (see Fig. 2(b)) and the torso as a reference. It is worth to note that we have tested our system on two datasets containing a different representation of the skeleton (15 and 25 joints) and we obtained our best performances with the same value of N. This restricted set of joints has shown to be the most discriminative for activity recognition, allowing also to reduce the complexity of the further steps of computation. As a consequence, in this case, a posture feature is made by

18 attributes. The next section describes how the most informative postures are selected from the all activity sequence.

3.2 Selection of Postures

The aim of the system is to represent an activity as multiple sequences of few and basic postures, i.e. key poses. This phase is in charge of select general and informative postures for each activity. Some authors [23,34] identify key poses calculating their kinetic energy and then considering a template of static and dynamic postures. This approach can be successfully used to discriminate some kinds of actions, but it is not necessarily the most suitable for others. In fact, some action samples may not have identifiable poses with zero kinetic energy, as also pointed out in [34]. Conversely, our approach aims to extrapolate few postures that are able to describe a specific activity by means of a clustering technique. One of the most used algorithms is the K-means, introduced in [36] and developed in many variations [37,38]. However, one of the main issues of this method is that the number of desired K clusters needs to be known a priori. Another possible approach is to run the K-means algorithm repeatedly for different values of K. However, this method is time consuming. As an alternative, we adopt the X-means algorithm [39], which is an optimized version of the previous one. It attempts to split the centers into regions and to select the number of clusters using a probabilistic scheme called Bayes Information Criterion. The X-means is much faster that run repeatedly K-means with a different number of clusters and it has proved to be less prone to local minima than its counterpart. In addition, it automatically finds the optimal number of clusters [40].

For each activity, the X-means is applied using the Euclidean distance function as a metric. In detail, given an activity composed by M posture features $[\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M]$, the X-means gives k clusters $[\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k]$, so as to minimize the intra-cluster sum of squares

$$\arg\min_{\mathbf{C}} \sum_{j=1}^{k} \sum_{\mathbf{f}_i \in C_j} \|\mathbf{f}_i - \mu_j\|^2 \tag{3}$$

where μ_j is the mean value of the cluster $\mathbf{C}_j$. At this stage, the set of posture features $\mathbf{f}_i$ representing an activity sequence is replaced with the centroid that the posture feature belongs, hence the centroids can be seen as the key poses of the activity.

3.3 Activity Features Generation

The aim of this step is to generate suitable features able to encode an activity. At this stage, the activity is composed of a set of centroids representing the most important postures for the sequence aligned in temporal order. The cardinality of this set is equal to the number of frames constituting the original action data. The features computed in the Sect. 3.2 need to be reduced in order

to lower the complexity and increase the generality of the representation. An important aspect to take into account is also the speed invariance of the feature since different person performs activities at different speed. In other words, this step considers only the transitions between key poses, i.e. transitions between centroids. For this reason, the temporal sequence obtained in Sect. 3.2 is now simplified in order to include only the transitions between clusters. This means that all the equal centroids that are consecutive in temporal order are discarded. This new compressed sequence allows to have a more compact representation of the activity and it is also speed invariant. At the end of the process, an action sequence is encoded in a temporally ordered sequence of centroids transitions. In order to be as general as possible, our aim is to characterize an activity defining several n-tuples composed of n poses. Therefore, the problem is to segment the obtained sequence. But, what exactly characterize an activity? Let consider a person who is drinking a glass of water. Does the action start when he is actually drinking? Or does it start when he begin to move his hand close to the mouth? And also, a person can drink in one or several gulps. To obtain our n-tuples, we adopt a sliding window on the sequence and we generate a set of new instances to represent the whole activity sequence. Consequently, the new instances are composed of a set of features with a size of $3L(N-1)$, where L is the length of the sliding window and N is the number of the selected skeleton joints.

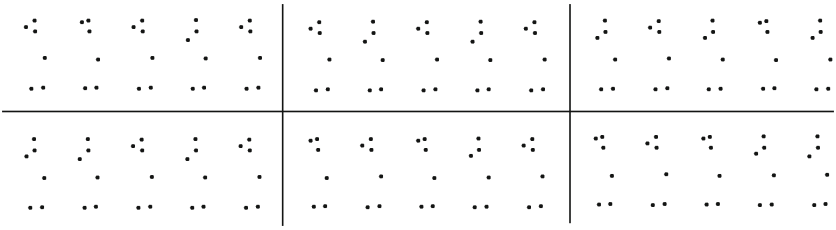

Fig. 3. Subset example of activity feature instances using a window length equals to 5 and a skeleton of 7 joints (torso omitted).

To clarify with a simple example, if an activity has 3 clusters, a possible compressed sequence can be

$$\mathbf{A} = [\mathbf{C}_1, \mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2] \tag{4}$$

if the length of the sliding window is $L = 5$, the cardinality of the activity feature instances is 3:

$$\begin{aligned} \mathbf{A}_1 &= [\mathbf{C}_1, \mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2] \\ \mathbf{A}_2 &= [\mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3] \\ \mathbf{A}_3 &= [\mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_2] \end{aligned} \tag{5}$$

The cardinality of the instances is related to the number of different transitions between different key poses. This means that actions which are repetitive

during the time will have fewer features instances than the ones with more variability between key poses. However, the newly generated instances are directly proportional to the window size in most of the cases. Figure 3 shows some activity feature instances of the drinking action using $L = 5$ and $N = 7$ (torso is omitted). For the training phase of the classifier, a dataset of activity feature instances has been built. The weight of the instances is increased if the same tuple appears in the sequence.

3.4 Activity Recognition

The activity recognition step involves the use of a classifier to associate the features created in the Sect. 3.3. Usually, machine learning techniques are applied to accomplish this task. In this work, we adopt the Sequential Minimal Optimization (SMO) [41] to train a multiclass Support Vector Machine (SVM) [42]. SVMs are supervised learning models used for binary classification calculating the optimal hyperplane that separates two classes in the feature space, while SMO makes use of improved internal structures and linear kernels in order to optimize the training phase. The multiclass SVM version is implemented by combining several binary SVMs using, in our case, a one-versus-one strategy [43].

4 Experimental Results

The system is implemented in Java using the Weka library [44], which is an open source software containing a collection of machine learning algorithms for data mining tasks. We have tested our system on two publicly available datasets, the first one is the well-known Cornell Activity Dataset (CAD-60) [17], widely used in activity recognition, while the second one is the fairly new TST dataset [45], specifically created for Activities of Daily Living (ADL). The skeleton model of these datasets are quite different, the first one represents it with 15 joints, while the second one uses 25 joints.

4.1 CAD-60 Dataset

The dataset focuses on realistic actions from daily life. It has been collected using a depth camera and contains actions performed by four different human subjects, two males, and two females. Three of the subjects use the right hand to perform actions, while one of them uses the left hand. There are 12 types of actions in the dataset, which are: “talking on the phone”, “writing on whiteboard”, “drinking water”, “rinsing mouth with water”, “brushing teeth”, “wearing contact lenses”, “talking on couch”, “relaxing on couch”, “cooking (chopping)”, “cooking (stirring)”, “opening pill container”, and “working on computer”. The dataset contains RGB, depth, and skeleton data, with 15 joints available. Each subject performs the activity twice, so one sample contains two occurrences of the same activity.

Table 1. Overall precision and Recall (%) values using 4 clusters and different window activity sizes on CAD-60.

Window size	"new person"	
	Precision	Recall
5	98.5	98.4
6	98.9	98.6
7	99.4	99.4
8	99.2	99.2
9	99.5	99.5
10	99.6	99.5
11	99.8	99.8
12	99.8	99.8

To be able to compare the results, we employ the same experimental settings of [17]. It consists of two cases: the "have seen" and "new person" setting. In the first case, the classification is done with the data of all the four persons, splitting the data in half. The latter uses a leave-one-out cross-validation approach for testing. This means that the classifier is trained on three of the four people and tested on the fourth. Since one person is left-handed, all the skeleton data are mirrored with respect to the sagittal plane of the skeleton. Conversely from [34], in which the right and left-handed samples are trained and tested separately, our samples contains both original and mirrored data. As for the other works on the CAD dataset, the performance is reported in terms of the average precision and recall among all the activities according to the "new person" test case.

At the beginning, the posture features are extracted as explained in the Sect. 3.1. After extensive tests, we have found that the most informative joints for our system are the head, the neck, hands and feet ($N = 7$). The output of the X-means algorithm (Sect. 3.2) gives 4 clusters as results for the majority of the activities. The activity features are generated following the procedure detailed in the Sect. 3.3. The last parameter that needs to be find is the size of the sliding window. We trained our classifier using different window size, ranging from 5 to 12. Therefore, the length of activity features has a minimum value of 90 and a maximum of 216, according to the window size.

In the "have seen" setting, the overall precision/recall reach always the 100 % for all the values of the clusters and the window size. This is coherent with the other works on the same dataset, in which most of them reaches the same result. More interesting is the outcome of the "new person" test case. The precision and recall values for all the tested window size are reported in Table 1. All of them produce high values in terms of precision and recall. Considering that the number of generated activity instances increases with the size of the window, we decided to select $N = 11$ to minimize the number of instances and maximize the performance. In this case, the total number of activity features is 199, i.e.

Table 2. Precision and Recall values for each activity, using 4 clusters and a window size of 11 elements on CAD-60.

Activity	"new person"	
	Precision	Recall
talking on the phone	1.0	1.0
writing on whiteboard	1.0	1.0
drinking water	1.0	1.0
rinsing mouth with water	1.0	1.0
brushing teeth	1.0	1.0
wearing contact lenses	1.0	1.0
talking on couch	1.0	1.0
relaxing on couch	1.0	1.0
cooking (chopping)	1.0	.977
cooking (stirring)	.969	1.0
opening pill container	1.0	1.0
working on computer	1.0	1.0
Overall	**.998**	**.998**

Table 3. The confusion matrix of the "new person" test case using $K = 4$ clusters and $N = 11$ window activity size on CAD-60.

	talking on the phone	writing on whiteboard	drinking water	rinsing mouth with water	brushing teeth	wearing contact lenses	talking on couch	relaxing on couch	cooking (chopping)	cooking (stirring)	opening pill container	working on computer
talking on the phone	1	0	0	0	0	0	0	0	0	0	0	0
writing on whiteboard	0	1	0	0	0	0	0	0	0	0	0	0
drinking water	0	0	1	0	0	0	0	0	0	0	0	0
rinsing mouth with water	0	0	0	1	0	0	0	0	0	0	0	0
brushing teeth	0	0	0	0	1	0	0	0	0	0	0	0
wearing contact lenses	0	0	0	0	0	1	0	0	0	0	0	0
talking on couch	0	0	0	0	0	0	1	0	0	0	0	0
relaxing on couch	0	0	0	0	0	0	0	1	0	0	0	0
cooking (chopping)	0	0	0	0	0	0	0	0	.98	.02	0	0
cooking (stirring)	0	0	0	0	0	0	0	0	0	1	0	0
opening pill container	0	0	0	0	0	0	0	0	0	0	1	0
working on computer	0	0	0	0	0	0	0	0	0	0	0	1

Table 4. Precision and Recall values for different joint configuration on CAD-60.

Joints	"new person"	
	Precision	Recall
15	0.826	0.851
11	0.816	0.823
7	0.998	0.998

Table 5. State of the art precision and recall values (%) on CAD-60 dataset.

	Precision	Recall
Zhu et al. [23]	93.2	84.6
Faria et al. [46]	91.1	91.9
Shan et al. [34]	93.8	94.5
Parisi et al. [47]	91.9	90.2
Cippitelli et al. [30]	93.9	93.5
Our method	99.8	99.8

size of the window equals to 11, and 7 joints, plus the activity attribute. The Table 2 shows the precision and recall values according to each activity using $K = 4$ clusters and $N = 11$ window activity size for the "new person" test case as usually reported for the CAD-60 dataset, while Table 3 shows the relative confusion matrix. The "cooking (chopping)" activity has been misclassified with the "cooking (stirring)" in only a few instances, indeed these activities are very similar. The algorithms have been run on a machine with a 2.4 GHz quad-core i7 processor and 8 Gb RAM. The activity features extraction took 70 milliseconds per activity. In the "new person" test case, the time to train the classifier was 1.24 s on 3564 instances. The time took for the test phase was 0.13 s on 920 instances. We have also conducted additional tests using different joint configuration: all 15 skeleton joints and 11 joints (excluding only shoulder and hip joints). Results prove that adding more joints does not improve the classification rate (see Table 4). The overall precision and recall value is 0.998 and this result outperforms the other works on the CAD-60 dataset. In the Table 5 is reported the current 5 best results of the CAD-60 dataset.

4.2 TST Dataset

We have tested our system also on the fairly new TST Fall detection database (version 2) [45]. The dataset has been collected using Microsoft Kinect v2 and IMU (Inertial Measurement Unit). It is composed of ADL and fall actions simulated by 11 volunteers. The people involved in the test are aged between 22 and 39, with different height (1.62–1.97 m) and sizes. The actions performed by a single person are separated into two main groups: ADL and Fall. Each activity is repeated three

Table 6. Overall precision and Recall values using 4 clusters and different window activity sizes on TST dataset.

Window size	"new person"			
	ADL		Fall	
	Precision	Recall	Precision	Recall
5	1	1	.960	.955
6	.962	.958	.955	.947
7	.951	.944	.95	.938
8	.940	.929	.946	.929
9	.938	.923	.942	.923
10	.938	.917	.938	.917
11	.938	.917	.938	.917
12	.938	.917	.938	.917

Table 7. Precision and Recall values for each activity, using 4 clusters and a window size of 5 elements on TST Dataset.

	"new person"		
	Activity	Precision	Recall
ADL	sit on chair	1.0	1.0
	walk and grasp	1.0	1.0
	walk back and forth	1.0	1.0
	lie down	1.0	1.0
Overall		**1**	**1**
Fall	frontal fall	1.0	.667
	backward fall	.889	1.0
	side fall	1.0	1.0
	backward fall and sit	1.0	1.0
Overall		**.972**	**.917**

times by each subject involved. The database contains 264 different actions for a total of 46k skeleton samples and 230 k acceleration values. Each person performs the following movements: "sit on chair", "walk and grasp an object", "walk back and forth", "lie down", "frontal fall and lying", "backward fall and lying", "side fall and lying", "fall backward and sit". Also for this dataset, we obtain the best performances using $N = 7$ joints. Applying the X-means algorithm we have an average number of clusters of $K = 4$ as for the CAD-60 dataset. We employ the same experimental settings as before, testing the classifier with the "new person" configuration. The Table 6 reports the overall precision and recall values of the ADL and Fall samples. In this case, the best window size is 5. This is coherent with the fact that each sample of the TST dataset contains a subject that performs the action

only once, while the subjects of the CAD-60 dataset perform the actions twice. We report in Table 7 the precision and recall values for each activity using 4 clusters and a window size of 5 elements. Only few frontal fall samples are misclassified with a backward falling, while the other samples are correctly classified.

5 Conclusion

This paper describes an activity recognition system using skeleton data obtained from an RGB-D sensor. The developed system is based on the idea of representing an activity with few and basic key poses, i.e. informative skeleton postures. The key poses are extracted using the X-means algorithm, and the activity features are built considering the centroid transition during the action. A multiclass SVM, trained with the SOM optimization, is used for the classification step. The system is tested on two publicly available datasets, the CAD-60 and the TST (version 2). The first is a well-known and widely used dataset for activity recognition. The results outperform the state-of-the-art, achieving an overall precision and recall of 0.998. The TST is a new dataset more close to the assistive technologies containing ADL and also Fall samples. The system does not misclassify sample of ADL case, and the overall precision and recall for the falling samples are 0.97 and 0.92 respectively. In particular, the misclassification happens with the frontal fall detected as backward fall. It does not represent a real issue from the perspective of a monitoring application for the ambient assistive living since these actions are essentially the same. It is also worth to note that the most informative skeleton joints are the same for both datasets, despite the fact that the original model is different, 15 against 25 joints.

These encouraging results make it feasible to use the presented system for assistive application and it will concern our future works. Anyway, one of the main drawbacks of the adopted classifier is that it is not able to handle unknown classes. In fact, it will always return one of the classes that better adapts to the observed one. Further study will be conducted in order to develop a real-time system for assistive scenarios exploiting the presented system.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
2. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. (CSUR) **43**(3), 16 (2011)
3. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. **115**(2), 224–241 (2011)
4. Argyriou, V., Petrou, M., Barsky, S.: Photometric stereo with an arbitrary number of illuminants. Comput. Vis. Image Underst. **114**(8), 887–900 (2010)
5. Aggarwal, J.K., Xia, L.: Human activity recognition from 3d data: a review. Pattern Recogn. Lett. **48**, 70–80 (2014)

6. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans. Cybern. **43**(5), 1318–1334 (2013)
7. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
8. Padilla-López, J.R., Chaaraoui, A.A., Gu, F., Flórez-Revuelta, F.: Visual privacy by context: proposal and evaluation of a level-based visualisation scheme. Sensors **15**(6), 12959–12982 (2015)
9. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden markov model. In: 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. Proceedings CVPR 1992, pp. 379–385. IEEE (1992)
10. Kellokumpu, V., Pietikäinen, M., Heikkilä, J.: Human activity recognition using sequences of postures. In: MVA, pp. 570–573 (2005)
11. Scholkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2001)
12. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., et al. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
13. Preece, S.J., Goulermas, J.Y., Kenney, L.P., Howard, D., Meijer, K., Crompton, R.: Activity identification using body-mounted sensorsa review of classification techniques. Physiol. Meas. **30**(4), R1 (2009)
14. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
15. Wang, J., Liu, Z., Wu, Y.: Learning actionlet ensemble for 3d human action recognition. In: Human Action Recognition with Depth Cameras, pp. 11–40. Springer, Heidelberg (2014)
16. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 9–14. IEEE (2010)
17. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 842–849. IEEE (2012)
18. Ni, B., Pei, Y., Moulin, P., Yan, S.: Multilevel depth and image fusion for human activity detection. IEEE Trans. Cybern. **43**(5), 1383–1394 (2013)
19. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: a color-depth video database for human daily activity recognition. In: Fossati, A., et al. (eds.) Consumer Depth Cameras for Computer Vision, pp. 193–208. Springer, London (2013)
20. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.: Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In: Alvarez, L., et al. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 252–259. Springer, Heidelberg (2012)
21. Yang, X., Tian, Y.: Effective 3d action recognition using eigenjoints. J. Vis. Commun. Image Represent. **25**(1), 2–11 (2014)
22. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. Int. J. Robot. Res. **32**(8), 951–970 (2013)
23. Zhu, Y., Chen, W., Guo, G.: Evaluating spatiotemporal interest point features for depth-based action recognition. Image Vis. Comput. **32**(8), 453–464 (2014)
24. Gan, L., Chen, F.: Human action recognition using apj3d and random forests. J. Softw. **8**(9), 2238–2245 (2013)

25. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–27. IEEE (2012)
26. Gaglio, S., Re, G.L., Morana, M.: Human activity recognition process using 3-d posture data. IEEE Trans. Hum. Mach. Syst. **45**(5), 586–597 (2015)
27. Ding, W., Liu, K., Cheng, F., Zhang, J.: Stfc: spatio-temporal feature chain for skeleton-based human action recognition. J. Vis. Commun. Image Represent. **26**, 329–337 (2015)
28. Jiang, M., Kong, J., Bebis, G., Huo, H.: Informative joints based human action recognition using skeleton contexts. Sig. Process. Image Commun. **33**, 29–40 (2015)
29. Chaaraoui, A.A., Padilla-López, J.R., Climent-Pérez, P., Flórez-Revuelta, F.: Evolutionary joint selection to improve human action recognition with rgb-d devices. Expert Syst. Appl. **41**(3), 786–794 (2014)
30. Cippitelli, E., Gasparrini, S., Gambi, E., Spinsante, S.: A human activity recognition system using skeleton data from rgbd sensors. Comput. Intell. Neurosci. **2016**, 14 (2016)
31. Baysal, S., Kurt, M.C., Duygulu, P.: Recognizing human actions using key poses. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 1727–1730. IEEE (2010)
32. Ballan, L., Bertini, M., Bimbo, A.D., Seidenari, L., Serra, G.: Effective codebooks for human action categorization. In: IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 506–513, September 2009
33. Raptis, M., Sigal, L.: Poselet key-framing: a model for human activity recognition. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 2650–2657. IEEE Computer Society, Washington, DC (2013)
34. Shan, J., Akella, S.: 3d human action segmentation and recognition using pose kinetic energy. In: IEEE International Workshop on Advanced Robotics and its Social Impacts, pp. 69–75. IEEE (2014)
35. Zhu, G., Zhang, L., Shen, P., Song, J., Zhi, L., Yi, K.: Human action recognition using key poses and atomic motions. In: 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1209–1214, December 2015
36. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Statistics, Berkeley, Calif., pp. 281–297. University of California Press (1967)
37. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
38. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of carefull seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
39. Pelleg, D., Moore, A.W.: X-means: Extending k-means with efficient estimation of the number of clusters. In: Seventeenth International Conference on Machine Learning, pp. 727–734. Morgan Kaufmann (2000)
40. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco (2011)
41. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)

42. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (2011) 27: 1–27: 27 Software available at. http://www.csie.ntu.edu.tw/cjlin/libsvm
43. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems, vol. 10, MIT Press (1998)
44. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
45. Gasparrini, S., Cippitelli, E., Gambi, E., Spinsante, S., Wåhslén, J., Orhan, I., Lindh, T.: Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. In: Loshkovska, S., Koceski, S. (eds.) ICT Innovations 2015. AISC, vol. 399, pp. 99–108. Springer, Heidelberg (2016)
46. Faria, D.R., Premebida, C., Nunes, U.: A probabilistic approach for human everyday activities recognition using body motion from rgb-d images. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pp. 732–737, August 2014
47. Parisi, G.I., Weber, C., Wermter, S.: Self-organizing neural integration of pose-motion features for human action recognition. Front. Neurorobotics **9**(3) (2015)