Noname manuscript No.
(will be inserted by the editor)

A Multi-Level Refinement Approach Towards the Classification of Quotidian Activities Using Accelerometer Data

Dario Ortega Anderez · Ahmad Lotfi · Caroline Langensiepen · Kofi Appiah

Accepted: 23 October 2018; First Online: 30 October 2018

Abstract Wearable inertial measurement units incorporating accelerometers and gyroscopes are increasingly used for activity analysis and recognition. In this paper an activity classification algorithm is presented which includes a novel multi-step refinement with the aim of improving the classification accuracy of traditional approaches. To do so, after the classification takes place, information is extracted from the confusion matrix to focus the computational efforts on those activities with worse classification performance. It is argued that activities differ diversely from each other, therefore a specific set of features may be informative to classify a specific set of activities, but such informativeness should not necessarily be extended to a different activity set. This approach has shown promising results, achieving important classification accuracy improvements of up to 4% with the use of low-dimensional feature vectors.

Keywords Activity Recognition · Quotidian Activities · Wearable Sensors · Accelerometer · Classification

1 Introduction

Human activity recognition (HAR) is a challenging research area extensively investigated in the field of Ambient Assisted Living (AAL)(Suryadevara and Mukhopadhyay, 2014). The knowledge of the daily behaviour of an elderly person living independently can be valuable information for clinicians (Godfrey et al., 2007). However, self-assessment of daily activities has shown to be subjective and variable, as a subject's own assessment can differ from that

Corresponding Author: Ahmad Lotfi
School of Science and Technology, Nottingham Trent University, Clifton Campus, Clifton Lane, Nottingham, NG11 8NS, United Kingdom
E-mail: ahmad.lotfi@ntu.ac.uk

of an expert in the field (Smith et al., 2005). This fact explains the increasing attention on the development of automatic activities tracking systems for subject monitoring.

Different sensor platforms are utilized with the aim of automatically monitoring a person in a home environment. According to the type of sensor/s employed for data collection, approaches can be roughly divided into three different categories; 1) computer vision (video cameras), 2) ambient sensors such as passive infrared sensors (PIRs), On/Off or open/close sensors, microphones, vibration sensors etc. and 3) wearable devices with inertial sensors such as accelerometers, gyroscopes and magnetometers. Research efforts are currently shifting towards wearable solutions, which avoid occlusion and the privacy concerns related to the use of video cameras in a home environment. In fact, previous surveys regarding the acceptability of the use of wearable devices showed positive results, not only in adults but also within the elderly population (Nelson et al., 2016), (O'Brien et al., 2015).

Several efforts have been made to develop systems for HAR using wearable sensors in view of its many applications in fitness, security and health care. However, predominantly, attention has been given to fitness applications, analysing periodic activities such as walking, running or climbing stairs (Bayat et al., 2014), (Casale et al., 2011), (Capela et al., 2015), (Erdaş et al., 2016). Little research is reported concerning quotidian daily activities, such as eating, drinking or hygiene-related activities, which could also be used as an indicator of a person's well-being. The authors in (Amft et al., 2007), (Dong et al., 2014), (Amft and Tröster, 2008) investigated activity events related to eating as well as dietary periods. Amft et al. (2010) studied fluid intake. However, hygiene-related activities have not yet been studied in depth.

A HAR system normally embodies data collection, signal pre-processing, feature extraction, feature selection/reduction and activity classification. Given that human activity takes places at frequencies of up to 20 Hz, data collection is normally performed at frequencies of 50 Hz to 100 Hz avoiding an aliasing effect on the collected signals (Wang et al., 2011a). In the pre-processing step, different filters are applied to eliminate potential noise outside the human activity frequency range, as well as to separate the low frequency component (gravity) from the high frequency component (body motion) (Bayat et al., 2014), (Casale et al., 2011), (Erdaş et al., 2016), (Mannini and Sabatini, 2010). After filtering, the signal is normally segmented into sliding windows (Bayat et al., 2014), (Casale et al., 2011), (Ravi et al., 2005), (Wundersitz et al., 2015), (Erdaş et al., 2016), (Capela et al., 2015) or shorter fundamental movements known as primitives (Zhang and Sawchuk, 2012), (Garcia-Ceja and Brena, 2013) from which feature vectors are extracted.

Features are normally calculated in the time and frequency domains. The extraction of features in the wavelet domain have been investigated (Erdaş et al., 2016), however, the classification results did not improve when added to the features calculated in the other two domains. Within the time domain, statistical features have demonstrated good classification results. Examples include mean, standard deviation, inter-quartile range and correlation between

accelerometer axes (Casale et al., 2011). Algorithms such as Dynamic Time Warping (DWT) (Billiet et al., 2016), (Bruno et al., 2013) and Histogram of Oriented Gradients (Miyamoto and Ogawa, 2014), typically used in computer vision, are also employed as feature descriptors. Once the feature vector is defined, a step of feature selection/reduction is normally applied to remove those redundant and irrelevant features not contributing significantly to improve the signal description. Selection models such as Chi-Square (Erdaş et al., 2016), Analysis of Variance (ANOVA) (Wundersitz et al., 2015), Lasso Regression (Wundersitz et al., 2015), or filter-based approaches such as Relief-F or Correlation-based Feature Selection (CFS) (Capela et al., 2015), have been used in previous work for dimensionality reduction purposes. Finally, the classification step takes place. Although unsupervised classification approaches have been investigated in the past (Kwon et al., 2014), human activity recognition is typically tackled as a supervised learning problem. A wide range of supervised classifiers such as Support Vector Machines (SVM), Neural Networks, Random Forest (RF) or K-Nearest Neighbours (KNN) have been widely used in previous studies (Jafari et al., 2007).

The aim of the research reported here is to develop a novel model to detect different sets of quotidian activities of daily living (ADL) using data collected with a wrist-mounted tri-axial accelerometer worn on the dominant hand of the user. The rest of this paper is organized as follows. Section 2 reviews previous work on the use of wearable devices for activity recognition. Section 3 describes the methodology followed to develop the multi-level refinement classification approach. Section 4 discusses the results. Section 5 justifies the use of the multi-level refinement step in future activity classification studies.

2 Related Work

Many attempts have been made to develop HAR systems using wearable sensors. Studies have varied with regards to the type of sensors, the number of them, their location, the activities to be tracked, the pre-processing techniques, the feature extraction and selection methods, the classification approaches as well as the research purpose itself. Bayat et al. (2014) used a smart-phone with a single tri-axial accelerometer to test the activity recognition results on a set of six activities, carrying the phone in the pocket and carrying the phone in the hand. They achieved a maximum accuracy of 91.15% on the 'in-hand' experiment using a combination of different classifiers. Casale et al. (2011), studied a group of five activities using a single tri-axial accelerometer worn on the chest, achieving a maximum accuracy of 94% using the Random Forest Classifier. Ravi et al. (2005) used a tri-axial accelerometer worn near the pelvic region to study eight different activities, obtaining an accuracy of over 99% combining different classifiers by applying plurality voting. However, data was collected only from two subjects, thus compromising the variability of the experiment.

Wundersitz et al. (2015) studied a circuit composed of eight fitness activities using a system embodying a tri-axial accelerometer and a tri-axial gyroscope embedded in a vest and located at the upper trunk of the experiment participants, achieving a maximum accuracy of 92% using a Logistic Model Tree (LMT). Zhang and Sawchuk (2012) used data from a tri-axial accelerometer and a tri-axial gyroscope worn at the right hip to develop a Bag-of-Words (BoW) approach applied to activity recognition. They achieved a maximum accuracy of 92.7% using a K-Means clustering algorithm for primitive construction and Support Vector Machine using an RBS kernel for the classification of those primitives. Erdaş et al. (2016), used data from a tri-axial accelerometer worn on the chest to study seven activities. Their approach included an ensemble feature selection, which combined with a Random Forest classifier obtained a maximum accuracy of 88%. Godfrey et al. (2007) used two accelerometers worn on the sternum and on the right thigh to classify sitting, standing and stepping with an accuracy of around 98%. Jafari et al. (2007) classified sit-to-stand, stand-to-sit, lie-to-stand and stand-to-lie movements using a tri-axial accelerometer exhibiting an average of 84% accuracy using a K Nearest Neighbours classifier.

Dong et al. (2014) developed a detector for periods of eating from data collected by a wrist-worn tri-axial accelerometer and gyroscope, combining a custom wrist energy peak detector and a naive Bayes classifier, achieving an accuracy of 81%. Munoz-Organero and Lotfi (2016) developed a stochastic model for recognizing and classifying different types of steps and falls using a tri-axial accelerometer worn on the abdomen for model training and on the chest for validation, achieving a step classification accuracy of 91.14%. Garcia-Ceja and Brena (2013) tackled a long term activity recognition problem as a distribution of simple activities with an accelerometer worn on the belt of the users, achieving a maximum accuracy of 92.5% using a KNN classifier. Wang et al. (2011a) developed a Hidden Markov Model (HMM) to classify six different daily activities using data from a waist-worn tri-axial accelerometer, achieving a recognition accuracy of 94.8%.

Billiet et al. (2016) studied six transition activities in rheumatic and musculoskeletal patients using a bi-axial accelerometer worn on the biceps of the dominant arm. This study included signal-processing features and pattern-based features applying Dynamic Time Warping (WDT), obtaining an average accuracy of 93.5%. The authors in (Mannini and Sabatini, 2010) developed a Gaussian Continuous Hidden Markov Model (cHMM)-based sequential classifier using data from five bi-axial accelerometers to classify seven different activities, achieving a maximum accuracy of 98.4%. Amft and Tröster (2008) investigated dietary events using four inertial measurement units (IMUs) worn on the upper and lower parts of the arms, an ear microphone and a sensor collar composed of a stethoscope microphone and a Electromyogram (EMG), obtaining a maximum recognition rate of 82% for the classification of cutlery usage, drink, spoon usage and eating only using the hand. Using the same sensors, (Amft et al., 2007) studied the intake of six different types of food and water by the use of a Probabilistic Context-Free Grammar (PCFG) parser, achieving

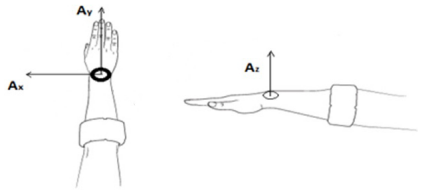

Fig. 1 Diagram showing how the accelerometer is mounted on the wrist and the direction of the orthogonal output signals $[A_x, A_y, A_z]$.

an average classification rate of 80%. Amft et al. (2010) studied the recognition of sips by using the feature similarity search (FSS) as feature descriptor, alongside traditional features, for data collected by a wrist-worn system embodying a tri-axial accelerometer, a tri-axial gyroscope and a compass, as well as a magnetic coupling sensor to measure the relative position between the wrist and the shoulder of subjects. The achieved sip detection rate was 89.2%. Authors in (Bruno et al., 2013), propose a recognition model for motion primitives relying on Gaussian Mixture Modelling (GMM) and Gaussian Mixture Regression (GMR) based on Dynamic Time Warping (DTW) and Mahalanobis distance descriptors, using data from a single wrist-worn accelerometer.

Among the studies above, a common feature can be seen; some activities are less easily discriminated than others. This problem has motivated the development of the multi-level refinement approach presented in this paper. For example, authors in (Wang et al., 2011a) struggled to classify 'sit' and 'fall'. Billiet et al. (2016) observed the highest number of false detections in their study was occurring in two specific activities; 'get up' and 'maxreach'. Amft et al. (2007) found the groups 'spoon' and 'apple' to have considerably lower detection rate than others. Zhang and Sawchuk (2012) found difficulties to classify 'walking upstairs' and 'walking downstairs' among other walking-related activities.

3 Methodology

Using a simple wristband which includes a tri-axial accelerometer sensor has proven to be the most acceptable form of wearable sensor. There are many readily available wristband which incorporate other sensors including gyroscope, gps, heart rate variation (HRV) and temperature. A tri-axial accelerometer is a sensor that provides simultaneous acceleration measurements in three orthogonal directions, namely x, y and z axes to represent acceleration A_x, A_y and A_z respectively. From that data, an informative feature set can be calculated and a posteriori the activity classification is tackled as a supervised classification problem. Using a tri-axial accelerometer sensor on a wrist is illustrated in Figure 1. As an alternative to a tri-axial accelerometer, a gyroscope could also be used. However, given that a gyroscope consumes approximately ten times the power of an accelerometer (Dong et al., 2014), making the use

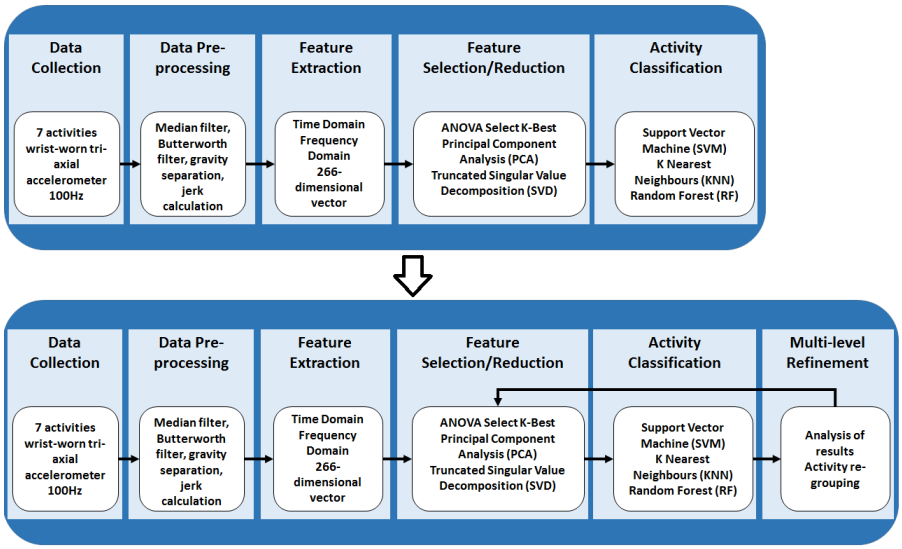

Fig. 2 Diagram representing the steps employed by previous work for activity classification (top) and in this paper (bottom). It can be seen that after the classification step takes place, an additional multi-level refinement step is included, whereby pairs of activities which worsen the performance of the classification model, are grouped together for further inspection.

of the former excessively power consuming for all-day monitoring systems, the decision was taken to use only accelerometer data, which as shown by previous HAR studies, is sufficient to achieve high activity classification rates.

Based on the literature review presented in Section 2, it can be observed that the general question of HAR is being addressed via several approaches. To the best of our knowledge, previous work in the field lacks further analysis of classification results which can lead to better classification performance.

In the remainder of this section, the multi-level refinement activity classification approach will be explained step by step. This investigation differs from earlier studies as follows. Firstly, some of the activities included in the activity set have not been studied together. Secondly, the proposed approach does not finish on the classification step as traditional approaches do. Instead, a novel approach is used whereby the classification performance in the form of confusion matrix is examined to identify those activities which worsen the performance of the proposed model, with the aim of directing the computational efforts towards refining the misclassification rate of the model. A comparison between the methodology used in this paper and that in most previous work is illustrated in Figure 2.

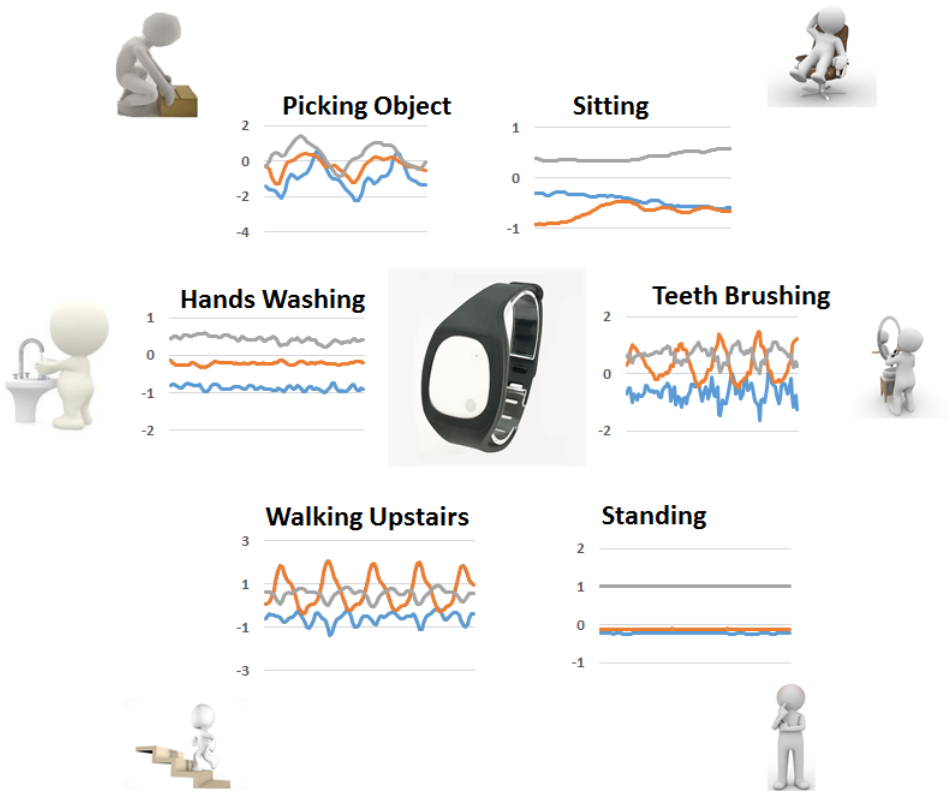

Fig. 3 Diagram of the activities studied and their accelerometer signals (1 second windows).

3.1 Data Collection

Six subjects, two female and four male, between 21 and 36 yeas of age, participated in this research experiment. They were ask to perform a set of seven quotidian daily activities:

1. Hand Washing,
2. Teeth Brushing,
3. Standing,
4. Sitting,
5. Picking up an object from the floor,
6. Walking Upstairs,
7. * Walking Downstairs is also included in the activity set for classification.

The volunteers were asked to wear the sensor on their dominant hand but no instructions were given as to how to perform the activities, adding reality and variability to the data. Since human activity happens at no more than 20 Hz (Wang et al., 2011a), a sampling frequency of 100 Hz was used which

according to the Nyquist theorem will be sufficient to avoid undesirable aliasing effects on the collected signals.

A *Meta Motion R* wristband mounted system was used for data collection (MetaMotionR, 2017). A visual representation of the system alongside one second windows of the signals from the different activities studied in this paper can be observed in Figure 3. This system includes, among other sensors, a tri-axial accelerometer, which was employed in this experiment. After the data was collected, it was sent via low energy Bluetooth to a smart-phone with the use of the Metabase app, which allows for the configuration of the sensors as well as the access to the sensor data. The question on where an inertial sensor should be worn to optimize the information collected keeps unraveled, since the optimal location is dependent on the chosen activity set. However, as compared to other areas of the human body, the wrist enables a high degree of freedom. Additionally, the wrist is a natural place for instrumentation, avoiding undesired obtrusiveness. Given that, the social acceptance is likely to increase due to its resemblance to a common watch.

3.2 Pre-processing

The accelerometer data is composed of three different time series $a_{x_t}, a_{y_t}, a_{z_t}$, which correspond to the medio-lateral, vertical and antero-posterior acceleration inputs respectively. A fourth time series, namely $|a_t|$, is calculated as the argument of the tri-dimensional vector.

Different smoothing and filtering techniques are applied to each time series. First of all, a median filter with a window length $w_l = 7$ has been used for smoothing purposes. Wang et al. (2011b) demonstrated that this filter has a competitive signal to noise ratio (SNR) as compared to other types of filters used for accelerometer data. On top of the median filter, a low-pass 20 Hz Butterworth filter is applied to filter out those frequencies not related to human activity. According to previous studies, human activity takes place at frequencies lower than 20 Hz (Wang et al., 2011a).

Two different components can be extracted from the accelerometer raw data; The gravity component, which is associated with the low frequency component of the signal, and the linear acceleration caused by the motion itself, which is associated to the high frequency component of the original signal. As done by previous work (Casale et al., 2011), a cut-off frequency of 1 Hz was used to extract the gravity component and the motion component from the signal.

In addition to the time series above, the rate of change of acceleration (jerk) of the the signal before being split into motion and gravity components was calculated, therefore obtaining four additional time series. The resultant time series are as follows:

- Original signal: $[a_{x_t}, a_{y_t}, a_{z_t}, |a_t|]$
- Motion: $[a_{x_{m_t}}, a_{y_{m_t}}, a_{z_{m_t}}, |a_{m_t}|]$
- Gravity: $[a_{x_{g_t}}, a_{y_{g_t}}, a_{z_{g_t}}, |a_{g_t}|]$

- Jerk: $[a_{x_{j_t}}, a_{y_{j_t}}, a_{z_{j_t}}, |a_{j_t}|]$, where $|a_{j_t}| = \frac{\partial |a_t|}{\partial t}$

The segmentation of the signals is performed using sliding windows with a window length of 1 second and a 40% overlapping percentage. Features will be calculated from these windows a posteriori.

3.3 Feature Extraction

Guided by previous work (Ravi et al., 2005), (Erdaş et al., 2016), (Dargie, 2009), different features in the time and frequency domains were evaluated. Within the time domain, different statistical features were explored. These include measures of central tendency like the mean and root mean square (RMS), measures of statistical dispersion such as range, standard deviation and inter-quartile range, measures of distribution shape such as kurtosis or skewness, measures of dependence between different axes, such as Pearson's correlation and measures of magnitude of varying quantity such as signal magnitude area. On top of the statistical features, measures such as peak frequency, zero-crossings frequency and signal entropy were calculated. After converting the signal to the frequency domain through the Fast Fourier Transform (FFT), features such as largest magnitude of the signal spectrum, index of the spectrum component with the highest magnitude and the energy of the signal across all the spectrum components were explored. Except few cases where it was not appropriate, features were calculated over all the time series exposed in Section 3.2. The dimensionality of the resultant feature vector is $n = 266$.

This feature set has been carefully selected so as to provide informative and discriminative information with regards to a wide array of signal characteristics, such as range, dispersion, central tendency, periodicity, frequency distribution, magnitude and changes in direction. Description of the selected features are presented below:

- Mean:

$$\overline{a} = \frac{1}{W_l} \sum_{t=0}^{W_l} a_t \tag{1}$$

  where a_t is the acceleration at time t and W_l is the window length expressed as number of samples.

– Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{t=0}^{W_l} (a_t - \overline{a})^2}{W_l - 1}} \tag{2}$$

  where a_t is the acceleration at time t, W_l is the window length expressed as number of samples, and $\overline{a}$ the mean acceleration of the corresponding window.

- Signal Magnitude Area:

$$sma = \frac{1}{W_l} \int_0^{T=W_l} (|a_{x_t} - \overline{a_x}| + |a_{y_t} - \overline{a_y}| + |a_{z_t} - \overline{a_z}|)dt \tag{3}$$

where $a_{x_t}, a_{y_t}, a_{z_t}$ are the acceleration at time t on the x, y and z axes respectively, W_l is the window length expressed as number of samples, and $\overline{a_x}$, $\overline{a_y}$, $\overline{a_z}$ the mean acceleration on the corresponding axis in the corresponding window.

- Signal Entropy:

$$H(a) = \sum_{t=0}^{Wl} |a_t - \overline{a}| \log_{10} |a_t - \overline{a}| \tag{4}$$

where a_t is the acceleration at time t, W_l is the window length expressed as number of samples and $\overline{a}$ the mean acceleration in the corresponding window.

- Correlation:

$$r_{xy} = \frac{Cov(a_x, a_y)}{\sigma(a_x)\sigma(a_y)} \tag{5}$$

where $Cov(a_x, a_y)$ is the covariance of the acceleration on the axes x and y, and $\sigma(a_x) and \sigma(a_y)$ are the standard deviation for the acceleration on the axes x and y respectively.

- Skewness:

$$\gamma_1 = \frac{\frac{1}{W_l} \sum_{t=0}^{W_l} (a_t - \overline{a})^3}{(\sigma(a))^3} \tag{6}$$

where a_t is the acceleration at time t, W_l is the window length expressed as number of samples and, $\overline{a}$ and $\sigma(a)$ are the mean acceleration and the standard deviation in the corresponding window respectively.

- Kurtosis:

$$\beta_2 = \frac{\frac{1}{W_l} \sum_{t=0}^{W_l} (a_t - \overline{a})^4}{(\sigma(a))^4} \tag{7}$$

where a_t is the acceleration at time t, W_l is the window length expressed as number of samples and, $\overline{a}$ and $\sigma(a)$ are the mean acceleration and the standard deviation in the corresponding window respectively.

- Root Mean Square

$$RMS = \sqrt{\frac{1}{W_l} \sum_{t=0}^{W_l} (a_t)^2} \tag{8}$$

where a_t is the acceleration at time t and W_l is the window length expressed as number of samples. The root mean square is calculated on all the time series presented in Section 3.2.

The transformation from the time domain to the frequency domain has been computed using the Fast Fourier Transform:

$$A(k) = \sum_{t=0}^{W_l - 1} a_t e^{(-i2\pi kt/W_l)} \tag{9}$$

where a_t is the acceleration at time t and W_l is the window length expressed as number of samples, A(k) is the sequence of W_l complex-valued numbers given the sequence of data a(t).

- Energy

$$E = \frac{\sum_{k=1}^{W_l} |a_k|^2}{W_l} \tag{10}$$

where a1,a2,... a_{W_l} are the FFT components of the corresponding window of length W_l.

3.4 Feature Selection/Reduction

As stated by Mannini and Sabatini (2010), when the dimension of a feature space is considerably high, learning the parameters for a classifier becomes a difficult and consuming task. In addition, feature selection/reduction can maintain or even increase the discriminative capability of the whole feature set. The current study has explored the use of three different methods for dimensionality reduction. First an analysis of variance (ANOVA) is conducted, where features were ranked according to their F measure, which is calculated as the ratio of the variance between classes and the variance within the class. After features are ranked, the subset that maximizes the classification result is selected. The other two approaches explored were Principal Component Analysis (PCA) and Truncated Singular Value Decomposition (SVD). These two approaches perform an orthogonal transformation of the data into a new coordinate system where the new coordinates are those which maximize the variance of the data, being the difference between the two approaches that PCA centres the data before computing the singular value decomposition.

3.5 Classification

As mentioned above, seven activities were investigated in this study. Each activity was recorded for each person separately and then labelled accordingly to fit them into the selected classifiers after feature extraction and selection. The performance of three different classifiers were evaluated: K-Nearest Neighbours (KNN), Random Forest (RF) and Support Vector Machine (SVM) using a Radial Basis Kernel (RBS). Prior to the evaluation, parameter estimation with 10-fold cross validation was computed to optimize the performance of the classification algorithms. In the case of K Nearest Neighbours, its performance was studied across the parameter K, which indicates the number of nearest points to be taken into consideration for the classification. For Random Forest,

its performance was studied across the maximum number of features utilized on each split and the number of decision trees employed for the classification. In SVM, the parameters C and γ which correspond to the penalty parameter of the model and the kernel coefficient respectively were considered.

3.6 Multi-Level Refinement

Our multi-level refinement can be defined as an algorithm that aims at optimizing the classification accuracy of a group of classes by an improvement on the recognition rate of those classes which lower the classification rate of the whole group. Its implementation is justified by the fact that in a classification problem, a classification accuracy lower than 100% is normally caused by the difficulty to classify specific classes, unless the recognition rate is identical for all the classes, though this is not a common occurrence.

After the activity classification takes place, the confusion matrix is further analyzed and activities are compared in pairs. If the classification accuracy between any pair of activities is lower than that on the whole model, those activities are grouped together for refinement. Activities which are found to lower the accuracy of the system due to their misclassification rate with an activity already pertaining to a refinement group are added to that same group, otherwise a new refinement group is constructed with these pair of activities. At this point the feature selection is optimized for each group selecting the most informative feature set for each of them. This process is repeated until groups of two activities are reached.

The multi-level refinement then focuses the computational efforts on the classification of those activities that are more difficult to classify in the first place.

We state the multi-level algorithm as shown in Algorithm 1.

4 Results

In this section, the experimental results reached are presented, explained and discussed. Section 4 is divided as follows; Section 4.1 computes parameter estimation for KNN, RF and SVM and the results of these three classifiers are compared. Section 4.2 examines the feature selection methods proposed for dimensionality reduction. Section 4.3 presents the classification results and the improvement achieved by the multi-level refinement algorithm. Finally Section 4.4 discusses the results obtained.

4.1 Parameter Estimation

Parameter refinement is implemented to estimate the optimal hyper-parameters for the data set. To do so, a 10-fold cross validation approach is used across different arrays of parameters, specific to each of the classifiers, to find the

Algorithm 1 Multi-level Refinement

```
1: top:
2: accuracy ← classification_accuracy
3: c_m ← confusion matrix
4: i ← rows confusion matrix
5: j ← columns confusion matrix
6: for n_rows do
7:     for n_columns do
8:         if (row ≠ column) then
9:             if ((1 − c_m[i, j]/c_m[i, i]) < accuracy) then
10:                activity_pairs.append[(i, j)]
11:            end if
12:        end if
13:    end for
14: end for
15: for activity_pair ∈ activity_pairs do
16:    for activity ∈ activity_pair do
17:        if (activity belongs to a group) then (add its pair to the group)
18:        else(create new group and add both activities)
19:        end if
20:    end for
21: end for
22: if All activities belong to the same group then (remove activity with the highest accuracy)
23: end if
24: for group ∈ groups do
25:    Feature Selection
26:    Run Classifier
27:    if (grouplength > 2) then
28:        goto top.
29:    end if
30: end for
```

combination that leads to the best classification performance. As it can be seen in Table 1, Table 2 and Table 3, Random Forest outperforms KNN and SVM on classification performance. In particular, the best classification performance was obtained computing Random Forest with 40 decision trees and when the number of features to be considered at each split within each decision tree is log_2 of the total number of features, in this case: $log_2(266) \approx 8$.

4.2 Feature Reduction

Having done the parameter estimation, ANOVA K-best, PCA and SVD are computed to find out the optimal subset of features/components for the description of the data set. To do so, the performance of the different feature selection methods is examined across all the possible values ranging from n=1 to n=266 (whole feature set). Although PCA and SVD peak faster than ANOVA K-best, needing fewer dimensions to obtain reasonably good performance, the best classification result (classification accuracy = 98.04%) is obtained using ANOVA K-Best with n=209, n being the number of features after ranking

according to their F ratio. The performance of the different feature selection methods across the number of dimensions can be observed in Figure 4.

4.3 Classification and Refinement

Once the optimal hyper-parameters and feature sub-set are estimated, the data set is divided, so that 80% is used as the training set and the remaining 20% as the test set. The classification reports the following confusion matrix and classification metrics (See Figure 5 and Table 4 respectively).

The classification accuracy achieved by the model is 97.46%, however there exist relevant differences in terms of precision, recall and F-score between dif-

Table 1 Random Forest: Parameter Estimation.

Number of trees	**auto**	log_2	**sqrt**
10	0.9513	0.9469	0.9508
20	0.9583	0.9600	0.9600
30	0.9569	0.9591	0.9553
40	0.9573	**0.9633**	0.9628
50	0.9610	0.9592	0.9611
60	0.9617	0.9626	0.9605
70	0.9624	0.9590	0.9603
80	0.9609	0.9614	0.9609
90	0.9606	0.9622	0.9602
100	0.9621	0.9613	0.9626

Table 2 K Nearest Neighbours: Parameter Estimation.

K Neighbours	**Accuracy**	**K Neighbours**	**Accuracy**
1	0.9511	11	0.9480
2	0.9504	12	0.9473
3	0.9526	13	0.9479
4	0.9520	14	0.9464
5	**0.9528**	15	0.9459
6	0.9527	16	0.9460
7	0.9508	17	0.9447
8	0.9505	18	0.9442
9	0.9493	19	0.9449
10	0.9493	20	0.9442

Table 3 Support Vector Machine: Parameter Estimation.

C	γ	**Accuracy**	**C**	γ	**Accuracy**
0.1	0.001	0.6427	1	0.1	**0.8919**
0.1	0.01	0.8286	1	1	0.8033
0.1	0.1	0.8421	10	0.001	0.8597
0.1	1	0.7271	10	0.01	0.8855
1	0.001	0.8281	10	0.1	0.8917
1	0.01	0.8590	10	1	0.8053

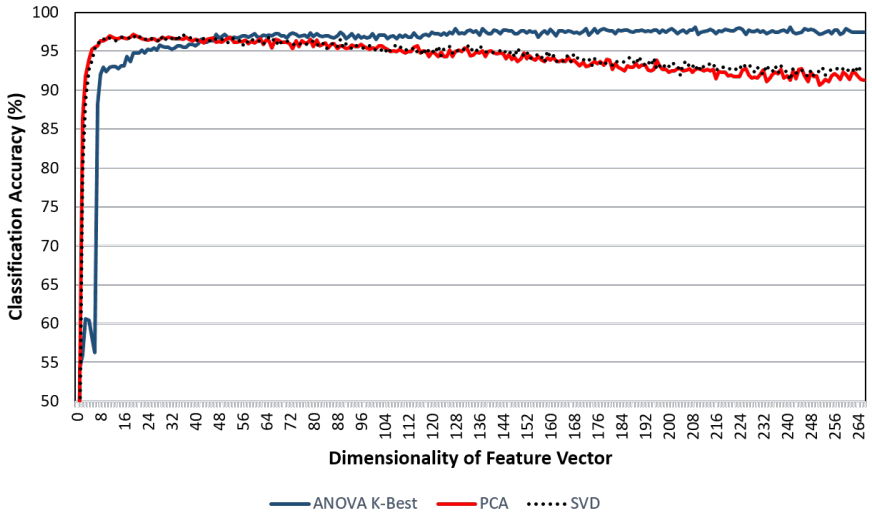

Fig. 4 Accuracy of different feature selection methods employed as a function of the number of dimensions.

	Hands Washing	Teeth Brushing	Standing	Sitting	Picking Object	Walking Downstairs	Walking Upstairs
Hands Washing	456	2	1	0	2	0	1
Teeth Brushing	2	149	0	0	0	0	0
Standing	1	0	241	0	0	0	0
Sitting	0	0	0	328	0	0	0
Picking Object	8	3	2	0	452	0	0
Walking Downstairs	0	6	0	0	0	29	7
Walking Upstairs	0	3	0	0	0	6	36

Fig. 5 Confusion matrix before the first refinement step.

ferent activities. Taking into account the respective binary classification accuracies from the confusion matrix, Algorithm 1 outlines that the activities lowering the performance of the model are Teeth Brushing, Walking Downstairs and Walking Upstairs. Consequently, these activities are selected for further inspection. Considering the interactions between this set of activities before refinement, the resultant confusion matrix and classification metrics are as shown in Figure 6 and Table 5 respectively.

At this point, feature selection and classification are performed again for the new activity set composed by three activities. The classification metrics

	Teeth Brushing	Walking Downstairs	Walking Upstairs
Teeth Brushing	149	0	0
Walking Downstairs	6	29	7
Walking Upstairs	3	6	36

Fig. 6 Confusion matrix before the first refinement step. The activities remaining are 'Teeth Brushing', 'Walking Downstairs' and 'Walking Upstairs'.

observed in Table 6 show that, in comparison to the classification performance before the first refinement step shown in Table 5, the classification accuracy, precision, recall and F-measure improved for all three activities. A visual representation of the improvement on classification accuracy after the first refinement step can be observed in Figure 7. The optimal classification accuracy was achieved using a 54-dimensional feature vector calculated during the refinement feature selection step. From the results obtained, it can be concluded that not only was the classification accuracy improved after refinement, but also that this was achieved using a feature vector with lower dimensionality as compared to the 209-dimensional feature vector used to classify the whole activity set.

Table 4 Classification metrics of the whole model .

	Acc.	**Prec.**	**Rec.**	**F-score**
H. Washing	99.00%	97.64%	98.70%	98.17%
T. Brushing	99.06%	91.41%	98.68%	94.90%
Stand	99.76%	98.77%	99.59%	99.18%
Sit	100.00%	100.00%	100.00%	100.00%
P. Object	99.12%	99.56%	97.20%	98.37%
W. Down	98.89%	82.86%	69.05%	75.32%
W. Up	99.00%	81.82%	80.00%	80.90%
Total	**97.46%**	**97.46%**	**97.46%**	**97.46%**

Table 5 Classification metrics before the first refinement step.

	Accucacy	**Precission**	**Recall**	**F-score**
T. Brushing	95.96%	94.30%	100.00%	97.07%
W. Down	91.85%	82.86%	69.05%	75.32%
W. Up	93.04%	83.72%	80.00%	81.82%
Total	**90.68**%	**90.68**%	**90.68**%	**90.68**%

Table 6 Classification metrics after the first refinement.

	Accucacy	**Precission**	**Recall**	**F-score**
T. Brushing	98.73%	98.09%	100.00%	99.04%
W. Down	95.88%	83.67%	95.35%	89.13%
W. Up	96.28%	100.00%	80.85%	89.41%
Total	**95.49**%	**95.49**%	**95.49**%	**95.49**%

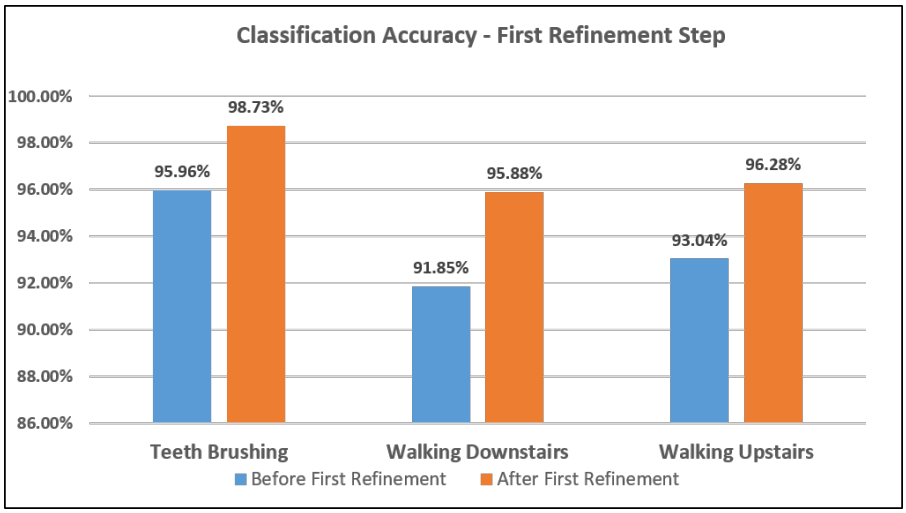

Fig. 7 Comparison of activity classification accuracy before and after the first refinement step.

With the resultant classification metrics in Table 6, obtained during the first refinement step, the same process is now repeated. In this case, Algorithm 1 selects the activities Walking Downstairs and Walking Upstairs for further examination, since according to the calculations performed, they were the classes lowering the classification accuracy of the group. Considering the interaction between these two activities, the resultant confusion matrix and classification metrics before the second refinement step are shown in Figure 8 and Table 7 respectively.

	Walking Downstairs	Walking Upstairs
Walking Downstairs	41	0
Walking Upstairs	8	38

Fig. 8 Confusion matrix before the second refinement step. The activities remaining are 'Walking Downstairs' and 'Walking Upstairs'.

Table 7 Classification metrics before the second refinement.

	Accuracy	**Precision**	**Recall**	**F-score**
W. Down	90.80%	83.67%	100.00%	91.11%
W. Up	90.80%	100.00%	82.61%	90.48%
Total	**90.80%**	**90.80%**	**90.80%**	**90.80%**

During the second refinement step, a 195-dimensional feature vector was found to be the optimal selection for the classification of the remaining two activities. In this case, the dimensionality is also reduced as compared to that for the whole activity set (n=209), though the number of dimensions increased with respects to the previous refinement step. This can be justified by the similarity between walking downstairs and walking upstairs in terms of wrist motion. The classification metrics after the second refinement step are shown in Table 8.

Table 8 Classification metrics after the second refinement step.

	Accuracy	**Precision**	**Recall**	**F-score**
W. Down	92.41%	82.86%	100.00%	90.63%
W. Up	92.41%	100.00%	88.00%	93.62%
Total	**92.94%**	**92.94%**	**92.94%**	**92.94%**

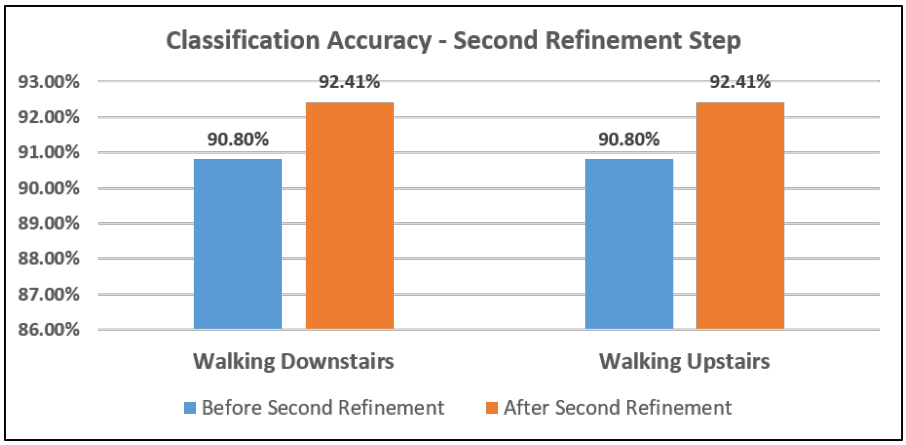

Fig. 9 Comparison of activity classification accuracy before and after the second refinement step.

It can be seen that again an improvement on classification accuracy was achieved. A visual representation of the classification accuracy improvement can be observed in Figure 9.

4.4 Validation and Discussion

To validate the multi-level refinement algorithm, a test was run on the Anguita et al. (2013) data set. The data set consists of data collected from 30 volunteers performing a group of six different activities of daily living (ADLs) while wearing a tri-axial accelerometer and a tri-axial gyroscope on the waist.

After the classification, two groups were formed: 1) Sitting and Laying, 2) Walking Downstairs and Walking Upstairs. An improvement on classification performance was seen in both groups. The first group improved classification accuracy from 96.92% to 100%. The second group went from 97.69% to 99.32%.

These results suggest that the use of the proposed multi-level refinement can improve classification accuracy on those activities that were more difficult to classify between when following a traditional classification approach. In addition, this method will benefit unbalanced experiments where data from specific activities are more difficult to collect as compared to others. After cross-validating the data-set, the data from those specific activities may be not enough to classify them against similar activities. This problem could be reduced by applying the multi-level refinement approach presented.

5 Conclusion

In this paper we propose a novel multi-level refinement technique for optimization of classification results in a HAR problem using accelerometer data. The proposed refinement algorithm autonomously analyses the confusion matrix, finding those activities that worsen the performance of the model and grouping them together for further inspection. Individual groups of activities are then studied separately, this process being repeated until only two activities remain in each group. The proposed approach has demonstrated that the classification results improve on each refinement step. Activities with low classification rates in the first place, obtained better classification accuracy when studied separately, even with the use of lower dimensional feature vectors in some cases.

This suggests that feature informativeness depends on the activity set chosen. Thus computational efforts should be given to particular group of activities (or classes) with lower classification performance, in order to optimize the selection of features and thus their classification rate. This approach could have a significant positive impact when the recognition of a specific activity or class is crucial for the interest of the study as for example a fall detection system.

References

Amft, O., Bannach, D., Pirkl, G., Kreil, M., Lukowicz, P., 2010. Towards wearable sensing-based assessment of fluid intake. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on. IEEE, pp. 298–303.

Amft, O., Kusserow, M., Tröster, G., 2007. Probabilistic parsing of dietary activity events. In: 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007). Springer, pp. 242–247.

Amft, O., Tröster, G., 2008. Recognition of dietary activity events using on-body sensors. Artificial intelligence in medicine 42 (2), 121–136.

Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J. L., 2013. A public domain dataset for human activity recognition using smartphones. In: ESANN.

Bayat, A., Pomplun, M., Tran, D. A., 2014. A study on human activity recognition using accelerometer data from smartphones. Procedia Computer Science 34, 450–457.

Billiet, L., Swinnen, T. W., Westhovens, R., de Vlam, K., Van Huffel, S., 2016. Accelerometry-based activity recognition and assessment in rheumatic and musculoskeletal diseases. Sensors 16 (12), 2151.

Bruno, B., Mastrogiovanni, F., Sgorbissa, A., Vernazza, T., Zaccaria, R., 2013. Analysis of human behavior recognition algorithms based on acceleration data. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, pp. 1602–1607.

Capela, N. A., Lemaire, E. D., Baddour, N., 2015. Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. PloS one 10 (4), e0124414.

Casale, P., Pujol, O., Radeva, P., 2011. Human activity recognition from accelerometer data using a wearable device. Pattern Recognition and Image Analysis, 289–296.

Dargie, W., 2009. Analysis of time and frequency domain features of accelerometer measurements. In: Computer Communications and Networks, 2009. ICCCN 2009. Proceedings of 18th Internatonal Conference on. IEEE, pp. 1–6.

Dong, Y., Scisco, J., Wilson, M., Muth, E., Hoover, A., 2014. Detecting periods of eating during free-living by tracking wrist motion. IEEE journal of biomedical and health informatics 18 (4), 1253–1260.

Erdaş, Ç. B., Atasoy, I., Açıcı, K., Oğul, H., 2016. Integrating features for accelerometer-based activity recognition. Procedia Computer Science 98, 522–527.

Garcia-Ceja, E., Brena, R., 2013. Long-term activity recognition from accelerometer data. Procedia Technology 7, 248–256.

Godfrey, A., Culhane, K., Lyons, G., 2007. Comparison of the performance of the activpal professional physical activity logger to a discrete accelerometer-based activity monitor. Medical engineering & physics 29 (8), 930–934.

Jafari, R., Li, W., Bajcsy, R., Glaser, S., Sastry, S., 2007. Physical activity monitoring for assisted living at home. In: 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007). Springer, pp. 213–219.

Kwon, Y., Kang, K., Bae, C., 2014. Unsupervised learning for human activity recognition using smartphone sensors. Expert Systems with Applications 41 (14), 6067–6074.

Mannini, A., Sabatini, A. M., 2010. Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors 10 (2), 1154–1175.

MetaMotionR, 2017. Mbientlab. Accessed: 2017-07-15.
URL `https://mbientlab.com/`

Miyamoto, S., Ogawa, H., 2014. Human activity recognition system including smartphone position. Procedia Technology 18, 42–46.

Munoz-Organero, M., Lotfi, A., 2016. Human movement recognition based on the stochastic characterisation of acceleration data. Sensors 16 (9), 1464.

Nelson, E. C., Verhagen, T., Noordzij, M. L., 2016. Health empowerment through activity trackers: An empirical smart wristband study. Computers in Human Behavior 62, 364–374.

O'Brien, T., Troutman-Jordan, M., Hathaway, D., Armstrong, S., Moore, M., 2015. Acceptability of wristband activity trackers among community dwelling older adults. Geriatric Nursing 36 (2), S21–S25.

Ravi, N., Dandekar, N., Mysore, P., Littman, M. L., 2005. Activity recognition from accelerometer data. In: Aaai. Vol. 5. pp. 1541–1546.

Smith, B. J., Marshall, A. L., Huang, N., 2005. Screening for physical activity in family practice: evaluation of two brief assessment tools. American journal of preventive medicine 29 (4), 256–264.

Suryadevara, N. K., Mukhopadhyay, S. C., 2014. Determining wellness through an ambient assisted living environment. IEEE Intelligent Systems 29 (3), 30–37.

Wang, J., Chen, R., Sun, X., She, M. F., Wu, Y., 2011a. Recognizing human daily activities from accelerometer signal. Procedia Engineering 15, 1780–1786.

Wang, W.-z., Guo, Y.-w., Huang, B.-y., Zhao, G.-r., Liu, B.-q., Wang, L., 2011b. Analysis of filtering methods for 3d acceleration signals in body sensor network. In: Bioelectronics and Bioinformatics (ISBB), 2011 International Symposium on. IEEE, pp. 263–266.

Wundersitz, D. W., Josman, C., Gupta, R., Netto, K. J., Gastin, P. B., Robertson, S., 2015. Classification of team sport activities using a single wearable tracking device. Journal of biomechanics 48 (15), 3975–3981.

Zhang, M., Sawchuk, A. A., 2012. Motion primitive-based human activity recognition using a bag-of-features approach. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM, pp. 631–640.