IEICE TRANS. INF. & SYST., VOL.E102–D, NO.1 JANUARY 2019

PAPER

Selecting Orientation-Insensitive Features for Activity Recognition from Accelerometers

Yasser MOHAMMAD[†,††*a)], ***Nonmember***, **Kazunori MATSUMOTO**[†††], ***and*** **Keiichiro HOASHI**[†††], ***Members***

SUMMARY Activity recognition from sensors is a classification problem over time-series data. Some research in the area utilize time and frequency domain handcrafted features that differ between datasets. Another categorically different approach is to use deep learning methods for feature learning. This paper explores a middle ground in which an off-the-shelf feature extractor is used to generate a large number of candidate time-domain features followed by a feature selector that was designed to reduce the bias toward specific classification techniques. Moreover, this paper advocates the use of features that are mostly insensitive to sensor orientation and show their applicability to the activity recognition problem. The proposed approach is evaluated using six different publicly available datasets collected under various conditions using different experimental protocols and shows comparable or higher accuracy than state-of-the-art methods on most datasets but usually using an order of magnitude fewer features.

key words: *activity recognition, machine learning, feature selection*

1. Introduction and Motivation

Activity recognition from mobile and wearable devices has several applications for end users, third parties, and social groups. Applications for end users include fitness and health tracking, fall detection, context-aware notifications, self-management apps among many others. Third parties can also utilize outputs from activity recognition for targeted advertising, and corporate management. Combining activity recognition results from several people can have social applications in participatory sensing, connecting people with similar activity profiles, event detection among many others.

Activity recognition from smart-phone sensors can be further divided into systems that deal with the problem as a temporal time-series classification task employing approaches like HMMs [1], and systems that treat activity recognition as a standard classification problem from predefined windows of sensory data [2]–[4]. This paper focuses on feature extraction and selection common to both approaches.

One of the distinguishing features for robust activity recognition on mobile devices is the low computation power available for these devices compared with traditional computers and the rise of concerns regarding privacy issues surrounding uploading raw sensory data from these devices to be processed over the cloud. Because of these considerations, this work limits itself to the most computationally efficient option in every stage of the classification pipeline.

Only accelerometer data is utilized in this work to reduce the energy consumed while reading the data from the sensors and to limit the total number of signals being processed. Both of these factors lead directly to energy saving. Moreover, accelerometers are shown in earlier studies to provide superior activity recognition results compared with gyroscopes [3].

Moreover, the proposed method further reduces expected energy requirements of the algorithm by not requiring continuous sampling of accelerometer data. This means that the algorithms used for classification can only assume the availability of short windows of data that are not related temporally which eliminates algorithms that try to model the temporal structure like HMMs and LSTMs [1], [5], [6]

The main goal of this work is to devise an activity recognition system that achieves consistent state-of-the-art results on a wide range of benchmark datasets using a single sensor (accelerometer) without the need to continuously sample that sensor (sparse sampling).

There are in general two methods to extract features from raw accelerometer data activity recognition: hardwired statistical features and feature learning (usually through deep learning). The first approach has the advantage of speed and reduced energy requirements compared with the second but with the disadvantage of lower recognition accuracy. This paper advocates a middle ground in which a large number of statistical features are compared offline and only the best of them are evaluated in real-time leading to a speed comparable to the first approach but — as will be shown in Sect. 5 — with an accuracy that is on the bar of the best performing algorithms in most datasets considered. Another advantage of the proposed approach that is shared with feature learning is relying mostly on automated feature discovery instead of having to manually define different statistical features for different datasets.

The main contributions of this work are four folds:

- A novel signal for accelerometer readings that is robust to axis permutations.
- A robust feature selection method based on multiple model evaluations that is shown to provide better fi-

Manuscript received March 14, 2018.
Manuscript revised August 21, 2018.
Manuscript publicized October 5, 2018.
[†]The author is with AIST, Tokyo, 135–0064 Japan.
[††]The author is with Assiut University, Egypt.
[†††]The authors are with KDDI Research Inc., Fujimino-shi, 356–0003 Japan.
[*]This work was conducted while the author was with KDDI Research Inc., Fujimino-shi, 356–0003 Japan.
a) E-mail: yasserm@aun.edu.eg
DOI: 10.1587/transinf.2018EDP7092

nal classification results on multiple datasets compared with standard model based feature selection avoiding the usual classifier bias of wrapper methods.
- A systematic analysis of the effect of signal and feature selection on classification performance on a variety of publicly available datasets and using different time-series aware evaluation strategies.

The rest of this paper is organized as follows. Section 2 discusses related work in the field of activity recognition. The methodology used in this paper is described in Sect. 3 and the datasets used are introduced in Sect. 4. Section evaluates the proposed system on all the datasets and on combinations of them. The lessons learned from the analysis of Sect. 5 are discussed in Sect. 6 along with limitations and directions of future research. The paper is then concluded.

2. Related Work

There are two main approaches to feature engineering in activity recognition: Most available systems use handcrafted feature sets that are extracted from the time and frequency domains [7]. The other extreme is to use deep learning methods for feature learning [8], [9])

Barshan and Yuksek [10] applied several classifiers including Support Vector Machines (SVM), and Random Forests (RF) to the problem of recognizing 19 different sports related activities using five Inertial Measurement Units (IMUs) (i.e., 15 sensors) attached to the trunk, hands and knees of eight subjects. Feature extraction employed 1170 statistical features that were then reduced to 30 features using PCA. An accuracy of 99.2% was reported using all sensors employing within-session evaluation. Calatroni et al. [2] studied the collection of activity datasets in highly rich networked sensor environments. The system used 7 IMUs, 12 3D acceleration sensors, 4 3D coordinates from a localization system with a total of 37 body worn sensors. The data was annotated with three types of activities: locomotive activities like walking and setting, low level activities like opening and closing a fridge, and high level activities like relaxation, grooming, eating a sandwich, etc. The maximum accuracy obtained with one of the four subjects using a neural network was 85.3% employing all the sensors.

State of the art feature learning methods employ deep neural architectures. Alsheikh et al. [8] trained a deep model consisting of one input Gaussian Restricted Boltzman Machine followed by four Binary Restricted Boltzman Machines each containing 1000 neurons to achieve 98.23% accuracy on the WISDM [3] dataset. The input to the system was a 303 dimensional spectrogram of the acceleration data. Zeng et al. [9] used convolutional neural networks (CNN) instead of RBMs for building a deep feature selector and classifier. The system used one 36-neuron convolution layer followed by a max-pooling layer and two fully connected hidden layers (1024, and 30 neurons in size) to generate 30 features from 64 input samples. This system achieved 96.88% accuracy on Actitracker [7] and 76.83% accuracy on the Opportunity datasets. Plotz et al. [11], evaluated using the empirical cumulative distribution function (ECDF) combined with a restricted Boltzman machine (RBM) to *learn* suitable features for activity recognition and reported an accuracy of 75.9% compared with 69% using statistical features [4].

A more promising approach for our goal of on-device recognition, is to use feature selection instead of feature learning. Wang et al. [12] recently conducted a comparative study using the UCI dataset in which they used 561 statistical features then compared PCA, a feature filtering approach (FCBF), a greedy wrapping approach (Wrapper), and a novel combination of them and showed that the later is superior to other approaches achieving 85.1% accuracy using 58 features on the UCI HAR [13] dataset using only accelerometer data

Deep learning approaches have also been proposed for both feature learning and classification of activities. Alsheikh et al. [8] trained a deep model consisting of one input Gaussian Restricted Boltzman Machine followed by four Binary Restricted Boltzman Machines each containing 1000 neurons to achieve 97.85% accuracy on the Actitracker dataset. The input to the system was a 303 dimensional spectrogram of the acceleration data. Zeng et al. [9] used convolutional neural networks (CNN) instead of BRMs for building a deep feature selector and classifier. The system used one 36-neuron convolution layer followed by a max-pooling layer and two fully connected hidden layers (1024, and 30 neurons in size) to generate 30 features from 64 input samples. A softmax classifier is then used for activity recognition. This system achieved 96.88% accuracy on Actitracker and 76.83% accuracy on the Opportunity datasets.

3. Methodology

The basic structure of the proposed activity recognition system is a standard machine learning pipeline. Training data from chosen sensor (i.e., accelerometer in this paper) is first preprocessed to increase the signal-to-noise ratio and optionally reduce the effect of gravity on accelerometer data. A set of signals are then extracted from segments of preprocessed data. This signal extraction step is designed to reduce the dependency of the final classification on the details of data collection like sensor orientation. The signals evaluated are then passed through a feature extraction module. The calculated features are analyzed by a feature selection system to keep the minimum number of features that does not significantly reduce the performance. In this work, we evaluate several feature numbers so the feature selector greedily keeps half of the features at every step. Finally only selected features are used to train the classifier. At recognition time, only the selected features need to be calculated on the mobile device. In accordance with our goal of requiring only sparse sampling (i.e. not sampling the sensors continuously), we model the classification problem as a supervised learning from features extracted from a single window in-

stead of a sequence learning problem. That allows the system to sample windows of sensor data with relatively long time delays between these samples but restricts it by being unable to use sequence modeling classifiers based on LSTM and similar methods.

3.1 Preprocessing

Preprocessing can improve the accuracy of later stages by improving the signal to noise ratio (SNR), yet it provides a source of computational complexity that we try to minimize in this paper. For this reason, only a 20Hz low pass filter is applied to the data under the assumption that 90% of the energy in human body motion is found in this frequency range [14].

3.2 Orientation-Insensitive Features

The accelerometer is a sensor that measures acceleration on three orthogonal dimensions that are usually called x,y, and z. These directions refer to a frame of reference intrinsic to the sensor and may be different for different smart-phones. Moreover the orientation of the smart-phone itself is not always the same during data collection in real-world scenarios and do not stay in the same orientation even for a single session of data collection. All of this suggests that relying on these three raw signals (x,y, and z) would reduce the accuracy of any activity recognition system in real-life situations.

To reduce the sensitivity to sensor orientation, some researchers use only the motion signal which is the norm of the raw signal at every time-step. This signal is orientation independent, yet it does not provide any information about the relative values of the three acceleration dimensions. Mizell [15] suggested the use of vertical-horizontal decomposition in which the direction of gravity is estimated as the mean direction of acceleration during a suitably long time-period (10 seconds in our experiments). The motion vector is then projected on the direction of gravity providing the vertical component and the difference between this vertical vector and the original vector is taken as the horizontal dimension.

This vertical-horizontal decomposition reduces the dependence of the classification pipeline on sensor orientation, yet it does not provide enough information about the relative amplitudes of motion on the three directions.

In this paper, we employ a novel axes-permutation independent signal along side motion and vertical-horizontal decomposition described by:

$$c = \frac{y-x}{z-x} + \frac{z-y}{x-y} + \frac{x-z}{y-z} \tag{1}$$

The main property of this novel signal is that it does not depend on permutations of the three axes x, y, and z which means that the same value will be generated for any multiple of 90 degree rotations as can clearly be seen from Eq. (1).

Using motion, horizontal-vertical decomposition, and c as described above reduces the effect of rotation on the signals processed by the next stages in the classification pipeline. Another problem with raw accelerometer signals is that their amplitude depend on the strength by which the motion is executed. For example, a user raising their arm slightly faster will generate higher accelerations. It is not a-priori clear whether or not amplitude information is advantageous for any given activity recognition task. For this reason, we combine the four aforementioned features (motion, horizontal-vertical decomposition, and c) with their z-score normalized versions to provide the set of "invariant signals" that are used for feature extraction and selection.

Another possible approach to reduce the dependence on the raw signals is to use the first two PCA components of the Hankel matrix representing the time-series [16]. In Sect. 5.3, these three approaches to signal generation (raw signals, the proposed invariant signals, and PCA) are compared in terms of the final classification accuracy.

3.3 Feature Extraction

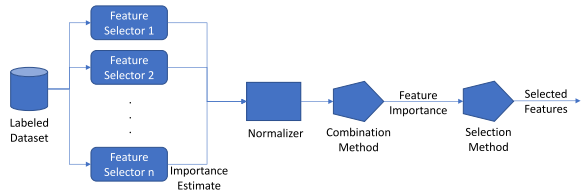
For feature extraction, we employ the TSFRESH library [17] which generates over 800 features from any input time-series (e.g., mean, median, quantiles, entropy based features etc). Feature extraction was applied for all dimensions of the signals generated from the previous step. For some of the databases (WISDM and UCI HAR datasets), handcrafted statistical features utilizing both time-domain and frequency domain information (unlike our time-domain only features) are available. For such cases, we applied feature selection and classification to both TSFRESH features as well as the handcrafted statistical features.

3.4 Feature Selection

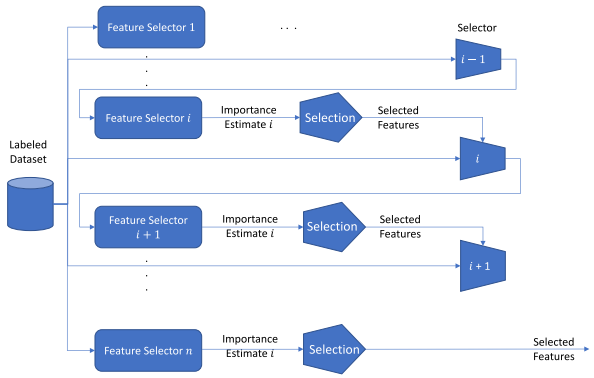
The core step of our approach is aggressive feature selection to reduce the number of features required to reduce the computational overhead of running the classifiers. Several approaches for feature selection have been proposed. Refer to [18] for a recent review.

One common approach for feature selection is randomized logistic regression (RLR) [19] in which randomly selected subsets of training samples are fitted using an L1 sparsity inducing penalty that is scaled for a random set of coefficient. The features that appear repeatedly in such selections (with high coefficients) are assumed to be more *important* and are given higher scores. RLR exemplifies the problems of not taking the time-series nature into account when processing sensory information for activity recognition. Consider applying RLR with some training data collected from the same user (i.e., personal evaluation). During the normal running of the feature selector, the random selection of sampling will result in having a within-session train/test split strategy which will lead to poor generalization in general. As will be shown later, this approach to feature selection does not provide acceptable results for activity recognition on our datasets.

A second common approach is to use a linear model



(a) Combination Method



(b) Switching Method

Fig. 1 The standard combination method and the proposed switching method for employing multiple base feature selectors

(e.g., a linear SVM) with an L1 penalty to fit the data and then select the features that have nonzero coefficients (or coefficients under a given threshold) from in the fitted model. In this work, we used a linear SVM as an example of these feature selection models.

A third approach is to employ algorithms that generate small tress (e.g., Random Forests or Extra Trees). During the process of learning the trees used for classification by these algorithms, an estimate of the quality of each feature is calculated at every step to generate a new node in the tree. These estimates can be used as the basis for feature selection. We employ Extra Trees for this work but experimenting with Random Forests gave similar results.

The final base feature selection approach considered here utilizes Principal Component Analysis (PCA). We find the set of Eigen vectors describing 80% of the variance in the data and use the absolute value of the coefficient for each vector weighted by the Eigen value corresponding to that vector to estimate the *importance* of each feature. To combine the importance measure from each Eigen vector, we use the geometrical mean.

In this paper, and along side the four aforementioned approaches, we proposed combining multiple feature extractors to provide a more robust feature selection strategy. The standard method for combining the scores of multiple feature selectors is shown in Fig. 1 (a) where the *importance measure* or score assigned to features using the base selectors are firstly normalized to a common numeric range and then combined using either geometric or arithmetic mean. The geometrical mean provides a sense of ANDing of feature importances which means that a feature that gets a high score in multiple vectors (after weighting) will have a much higher chance of being selected than a feature that only gets a high score from one or fewer vectors. Comparison of the geometrical mean and the arithmetic mean showed better performance of the geometrical mean in most datasets. Due to lack of space, we will only the results using the geometrical mean.

The alternative proposed in this paper (called "switching" hereafter) uses only one base method at every stage and switches between base feature selection methods in a round-robin fashion. At each stage, a predefined number (or percentage) of features are pruned and the process is repeated with the next base selector. Figure 1 (b) shows the proposed method.

The proposed *switching* method provides a natural method for selecting the number of features to retain in the final set by evaluating the classification accuracy (see Sect. 3.5) on a validation set and stopping when it becomes lower than the required accuracy for the application. Moreover, it runs faster than the standard *combination* method because each successive feature selector is trained using smaller number of features. When using slow selectors such as the L_1SVM method, this can provide substantial increase in selection speed. As will be shown in Sect. 5.4.2, these advantages are achieved without any loss in final classification accuracy on the hold-out set.

3.5 Classification

Several classifiers have been proposed for activity recognition including random forests [3], SVM [20], [21], MLPs [4], etc. We utilized Extremely Randomized Trees (ET) [22], and Random Forests (RF) [23] as examples of tree-based approaches, Multilayer Perceptrons (MLP), and Support Vector Machines (SVM) [24] as examples of statistical and connectionist approaches. The choice of these classifiers is based on previous research that showed that they consistently performed successfully in activity recognition tasks with different datasets [1], [3], [10]. The complexity of any classification problem is inversely correlated with the success of feature selection and extraction. To evaluate the effectiveness of the proposed feature extraction and selection method, we also utilize the simple K-Nearest Neighbor (KNN) method.

For all these methods, cross-validation based model selection was conducted employing an evolutionary algorithm that used cross-validated accuracy as the fitness measure. Best model parameters were then used in a separate cross-validation run to evaluate the classifier selected. Reported results used that classifier's prediction at every data-point when it was in the test set (i.e., cross-validation prediction). All classifier types performed roughly on the same level and due to lack of space comparative analysis between them will not be given in this paper.

4. Datasets

This section describes different public datasets used in this work. Most of the datasets used mobile phone sensors but few of them utilized Inertial Measurement Units (IMUs) and were included for comparison purposes. For each dataset, the collection protocol, classes of activities considered, and base evaluation results will be presented.

4.1 WISDM

The WISDM (WIreless Sensors Data Mining) 1.0 dataset [3] was collected by having 30 subjects carry a mobile phone in their pocket and were asked to perform specific activities while sensory data is collected using 20Hz sampling frequency. Data collection was supervised by one of the WISDM team members [3]. A total of 43 statistical features were calculated from every 10seconds of data in the dataset and the raw sensory data as well as calculated features were made publicly available. Six activities were chosen in this dataset: walking, jogging, climbing upstairs, descending downstairs, standing and sitting.

4.2 Actitracker

The Actitracker dataset [25]. was collected by having 36 subjects carry a mobile phone in their pocket while performing six daily activities (walking, jogging, ascending or descending stairs, standing, sitting and laying down). Accelerometer data were sampled at a 20Hz sampling frequency [7]. The same 43 features used for WISDM dataset were also calculated and shared publicly for Actitracker data [25]. Notice that compared with WISDM dataset, Actitracker combined stair activities and added the laying down activity. When we analyze the combined dataset, we had to combine the stairs activities as well and to combine all static activities (laying down, standing and sitting) due to this inconsistency. The data collection protocol in this case is less restrictive then the one employed in the WISDM dataset.

4.3 UCI HAR Datasets

UCI HAR stands for the UCI repository's dataset on Human Activity recognition Using Smart-phone Datasets [13]. A group of 30 volunteers were recruited for the collection experiment. Each participant carried a Samsung Galaxy II smart-phone while being instructed to do a predefined series of activities comprising 192seconds per session [13]. Each subject performed the same protocol twice. In the first trial, the smart-phone was fixed on the left side of the belt but the user was free to place it as preferred in the second trial. A separation of 5 seconds between each activity within the session were added to facilitates repeatability of the activities. The activity classes considered in this dataset were: walking, climbing upstairs, descending downstairs, standing, sitting and laying down. Other than the accelerometer data, the datasets contained gyroscope readings synchronized with accelerometer data. [14]. A total of 561 features from both accelerometer and gyroscope data were calculated for every 2.56seconds of data with an overlap of 50%. The data was divided into 70% training and 30% testing sets. In this paper we employ only the 386 accelerometer based features and combine the training and test sets then use 10-fold cross validation for all reported results.

4.3.1 HAPT

The Smart-phone-Based Recognition of Human Activities and Postural Transitions Data Set (HAPT) dataset [26] uses the same raw data used by the UCI HAR dataset but adds to ground-truth data six more labels each representing a transition between two of the basic six activities used in UCI HAR. The same preprocessing is used and the same features are extracted. When evaluating the SVM classification algorithm on this dataset, a series of heuristic rules for reducing the fluctuation between basic activities using the postural transition classes were used and the evaluation criterion did not consider errors in classifying the postural activities as long as they are considered either from one of the transitioning classes (e.g., a stand-to-sit misclassified as stand or sit) or to the newly added *unknown* class. In this paper, we count these misclassification as errors as is done with the other datasets. Wu and Zhang reported an accuracy of 89.88% [27] when using interpersonal evaluation without cross validation on the same dataset using the SVM classifier.

4.4 Sensor Datasets

For comparison purposes, we also include two datasets in which data collection employed on-body sensors instead of smart-phones. In both cases, several sensors were attached to the participant body but we will always use a single accelerometer sensor for all evaluations. It will be shown that even with this single sensor, it is possible to achieve similar or higher performance to systems that employ all of the sensors giving more credence to the applicability of smart-phone sensors to the activity recognition problem.

4.4.1 DSA

The Daily Sports Activity database (DSA) [10] was collected at the Bilkent University Sports Hall, in the Electrical and Electronics Engineering Building, and in a flat outdoor area on campus. Eight participants performed 19 different activities repeatedly for 5 minutes. The subjects were asked to perform the activities in their own style and were not restricted on how the activities should be performed. We only employ the accelerometer value from the IMU attached to the left leg in this paper.

4.4.2 Opportunity

The Opportunity Dataset for Human Activity recognition

from Wearable, Object, and Ambient Sensors (hereafter Opportunity dataset) is a dataset devised to benchmark human activity recognition algorithms [2]. The dataset comprises the readings of motion sensors recorded while users executed typical daily activities in a room simulating a studio flat. The following body-worn sensors were used: 7 inertial measurement units (IMUs), 12 3D acceleration sensors, 4 3D coordinates from a localization system. Only the accelerometer sensor attached to the left leg is used in this paper.

4.5 Combined Datasets

Combined datasets are created by concatenating data from two or more of the aforementioned datasets. They are named with the first character of each database used in them. For example, w+a+u is a concatenation of the data in WISDM, Actitracker, and UCI HAR datasets. Because different classes are used in different datasets, some of the class labels had to be combined. Namely, climbing up and down stairs were combined into the same class whenever the Actitracker dataset was used because they are not separated in this dataset. Moreover, Lying down, standing and sitting are also combined whenever the Actitracker dataset is used for the same reason. Three combined datasets were created by combining the WISDM dataset with either UCI HAR dataset (w+u), Actitracker dataset (w+a) or both (w+a+u). The final combined datasets use the predefined 43 handcrafted statistical features used in both WISDM and Actitracker datasets (w+a-pre).

5. Evaluation

To evaluate the proposed approach, we applied it to all the datasets presented in Sect. 4 and compared the results with the best result we could find in the literature for each of these datasets that were achieved using the same evaluation methodology. The proposed approach is designed to be applicable even when data cannot be streamed continuously from the sensors to reduce energy consumption which means that methods like Hidden Markov Models, Long-Short Term Memory (LSTM), etc that has this requirement are not considered in this study. This section starts by presenting the evaluation methodology employed in order to simplify reproducibility of the results (Sect. 5.1). We then present the main findings of the study supporting the claim that the proposed method can achieve comparable performance to state-of-art methods on all datasets while requiring less feature evaluation due to the proposed feature selection scheme in combination with the proposed invariant signals (Sect. 5.2). To evaluate the effect of each of these two factors on the final results, more evaluations of alternatives for each of them are provided in Sect. 5.3 and Sect. 5.4.

5.1 Evaluation Methodology

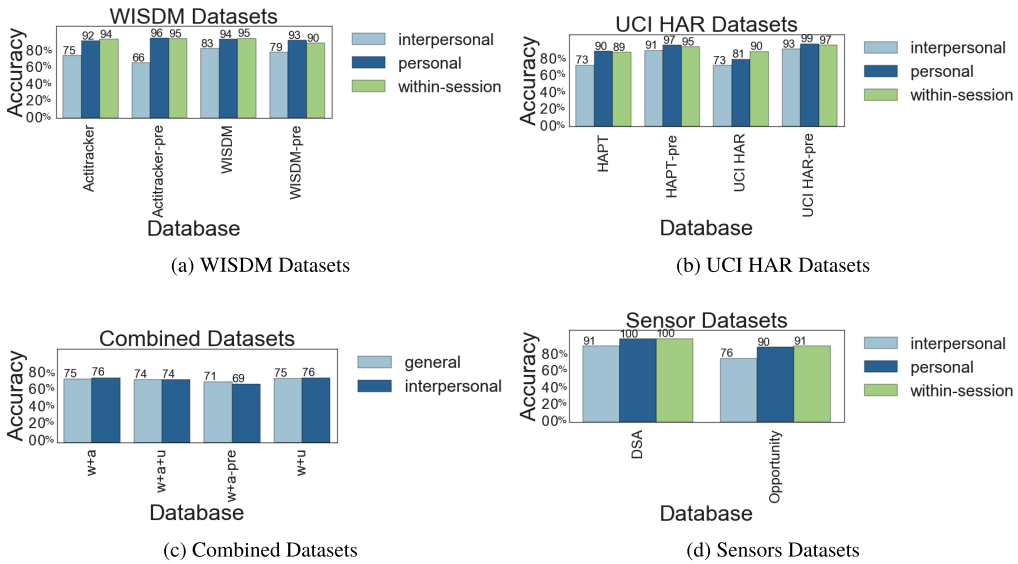
There are three main strategies to evaluate activity recognition systems that we can find in literature. Firstly, data from the same user can be randomly split into training and test sets disregarding the session in which they were collected (within-session evaluation) [10]. A more representative approach is using data from the same person for training and testing but ensuring that the training and testing sets do not overlap in terms of time of data collection (personal evaluation) [2], [28]. Thirdly, training data could be collected from a group of people and then testing data could be collected from different people (interpersonal evaluation). Weiss et al. proposed a fourth approach in which cross-validation is used on the complete dataset disregarding the source (hybrid evaluation) [28]. Finally, for the purpose of this work, we define *general* evaluation as the case when training and testing sets share no participants from the same dataset. This mode of evaluation better reflects expected real-world performance due to the differences in collection protocol that simulate real-world variability.

Regarding the evaluation metric (e.g., accuracy, recall, precision, F-measure), the correlation between any two of these measures in our experiment was higher than 0.95 which means that they all give the same picture of comparative performance between different approaches. The main reason for that is that reported results were obtained using classifier parameters optimized using cross-validation. Only the accuracy metric will be reported in this paper due to lack of space.

Evaluation of activity recognition algorithms is a challenging problem due to the sensitivity of the results to the details of the evaluation criterion. For example, it is common in machine learning tasks to use cross-validation with stratification as a method for evaluating classifiers. Most ML packages (e.g., Weka, PRTools, etc) resort to randomizing the samples before applying cross-validation to it by default which in general is a good practice in machine learning. Nevertheless, due to the fact that the features are extracted from mostly smooth time-series in activity recognition, this randomization may lead to having subsequences that were very near (and are expected to be very similar) in the time series split between the training and testing sets. This converts personal evaluation to a within-session evaluation and may lead to inaccurate characterization of the algorithm accuracy. In this work we strictly reserve the term personal evaluation to the case in which no such randomization is applied at any step of the algorithm. Moreover, no overlap was allowed between segments when using within-session evaluation. Just running normal cross validation with randomization is not considered here as within-session evaluation but simply an evaluation mistake that can easily lead to extremely high accuracy figures.

Throughout this section, when the proposed feature selection algorithm is applied to data extracted using handcrafted statistical features that appeared before in literature, it will be suffixed with the "pre" keyword, otherwise TSFRESH is used for feature extraction.

All evaluations conducted in this paper were conducted on a server with the following specifications: Two 2.20GHz



(a) WISDM Datasets

(b) UCI HAR Datasets

(c) Combined Datasets

(d) Sensors Datasets

Fig. 2 Best accuracy obtained on each of the datasets employed in this work.

Xeon CPUs (E5-2660) each with 8 cores and 256GB memory.

5.2 Main Results

Figure 2 shows the best obtained accuracy using the proposed system on each dataset employed in this work independent of the number of features. An apparent feature of the results is that interpersonal evaluation gives consistently lower accuracy than personalized classifiers. That is in agreement with the analysis done in [28] and shows that the similarity of behavior when using personal evaluation more than compensates for the dramatic reduction in the amount of training data. Moreover, accuracy in combined datasets is lower than that in each individual dataset in most cases. For example combining WISDM and Actitracker datasets (collected with only slightly different protocol using the same group) results in a reduction of mean accuracy of the top ten classifiers from 82% and 78% to 74%. The effect is less for the best performing classifier going down from 83% and 74% to 76%. The following subsections will discuss in details the results in relation to each step of the proposed approach.

Table 1 summarizes the main findings of this paper. The proposed approach was able to achieve a within-session evaluation higher than the state of art results except for the WISDM dataset in which it achieved 94.8% accuracy using 66 features compared with 96.6% accuracy using 540 features [29]. Even in this case the 9 folds reduction in the number of features needed more than compensates for the 2% reduction in accuracy. For personal evaluation, the story is more mixed. The proposed approach achieved higher accuracy than the best performing algorithm in the Opportunity dataset (with a 7.7% accuracy improvement) but not in the Actitracker dataset (with 4.7% reduction in accuracy). Again the proposed method in general needed a smaller number of features in the later case (5 compared with 43). Finally, for the interpersonal evaluation, the proposed method outperformed other approaches (or achieved within 2% of the best performing approach) in all databases except for WISDM in which it could only achieve 83% accuracy compared with 91.7%.

5.3 Signals

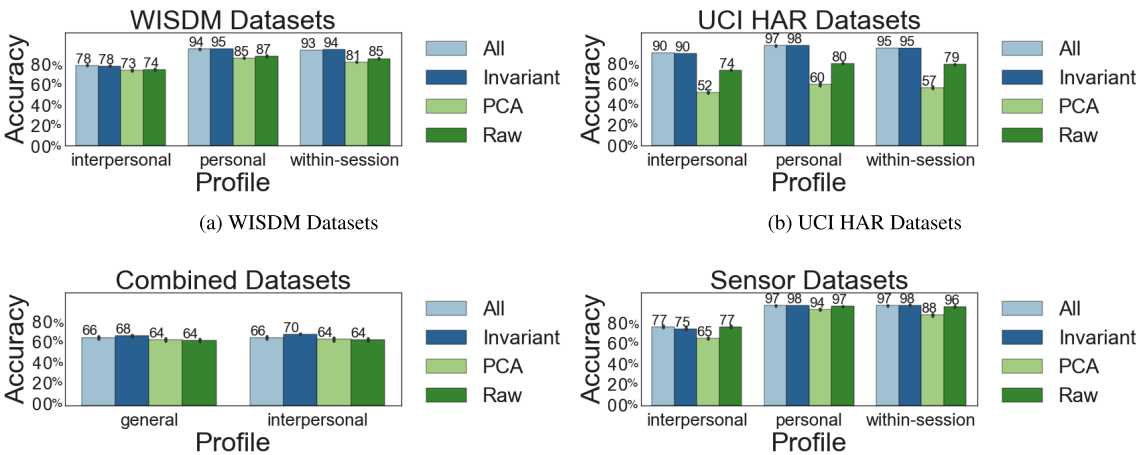
Figure 3 shows the effect of signal choice on the accuracy of the best classifier. Using PCA for signal generation resulted in worse performance on all databases compared with using the raw x,y, and z signals with a difference up to 12% in accuracy for the sensor datasets and 38% for the UCI HAR dataset. Moreover, the proposed orientation invariant signals achieved a performance on bar with utilizing all of the signals together on all databases using all evaluation criteria and performed higher than using the raw signals in most cases with 7% improvement for the personal profile in WISDM datasets, and 18% for UCI HAR dataset. The choice of signals did not affect the accuracy in combined datasets and raw signals outperformed the proposed signals for sensor datasets by 2% for interpersonal evaluation. This final result is not surprising given the fact that on-body sensors are more likely to be fixed in orientation and position during all data collection sessions compared with smart-phone sensors.

Considering the results in Table 1, the proposed invariant signals appear 13 times (out of 18) compared with only once for raw signals. Applying t-tests to check the difference that appear in Fig. 3 with Bonferroni's multiple-comparisons corrections, it was found that the invariant signals lead to higher accuracy compared with raw signals ($t = 17.06$, $p < 0.0001$), and PCA based signals ($t = 19.47$, $p < 0.0001$) but is not different from using all signals ($t = -2.75$, $p = 0.006 > 0.05 \times 12$). Moreover raw sig-

Table 1 Comparison of the proposed approach and state of the art results on different datasets. For each case, the number of features used (#) are shown.

Dataset	Profile	Previous Work			Proposed				
		Accuracy	#	Method	Accuracy	#	Signals	Selector	Classifier
Actitracker	Within-session [30]	92.0	15	ET	**94.3**	**10**	Invariant	Kernel PCA	ET
	Personal [31]	**98.7**	43	MLP	94.0	**5**	Invariant	Switching	SVM
	Interpersonal [31]	**75.9**	43	SVM	74.5	**8**	Invariant	Switching	ET
	Other [32]	*96.9*	*30*	CNN					
WISDM	Within-session [29]	96.6	540	KNN	94.8	**66**	Invariant	Combined	KNN
	Personal				*93.3*	**6**	All	Tree	KNN
	Interpersonal [3]	91.7	43	MLP	83.1	**33**	All	L_1SVM	ET
	Other [8]	*98.23*	*303*	BRM*					
UCI HAR	Within-session [1]	96.8	561	SVM*	**98.9**	**38**	Invariant	RLR	ET
	Personal				**97.0**	**4**	Invariant	Tree	KNN
	Interpersonal				*90.6*	*16*	Invariant	Switching	ET
HAPT	Within-session [27]	89.59	561	SVM	**93.9**	**16**	Invariant	Switching	RF
	Personal				95.5	**10**	All	Combined	ET
	Interpersonal [27]	89.99	561	SVM	**90.0**	**16**	Invariant	Switching	SVM
	Other [26]	*96.7*	*561*	SVM*					
DSA	Within-session [10]	99.2	1170	SVM	**99.5**	**33**	Invariant	Switching	ET
	Personal				97.6	**4**	All	RLR	ET
	Interpersonal [33]	87.6	1170	SVM	**90.9**	**16**	Raw	Combined	ET
Opportunity	Within-session				**90.1**	**4**	Invariant	Tree	ET
	Personal [4]	80.7	3	Clustering*	**88.0**	**16**	Invariant	Combined	KNN
	Interpersonal [9]	**76.8**	64	CNN	76.3	**32**	Invariant	Tree	ET

* Systems that used aggregation or dynamic adaptation.
Bold entries in the accuracy field indicate the highest accuracy for the given dataset and methodology while bold entries in the (#) field indicate the lowest number of features.
Underlined entries are systems that achieved accuracy within 2% of the maximum accuracy for this dataset and methodology.
Italicized items represent systems that did not clearly specify an evaluation methodology of the three considered in this paper.



(a) WISDM Datasets

(b) UCI HAR Datasets

(c) Combined Datasets

(d) Sensors Datasets

Fig. 3 The effect of the signals used on the best accuracy obtained using the proposed system.

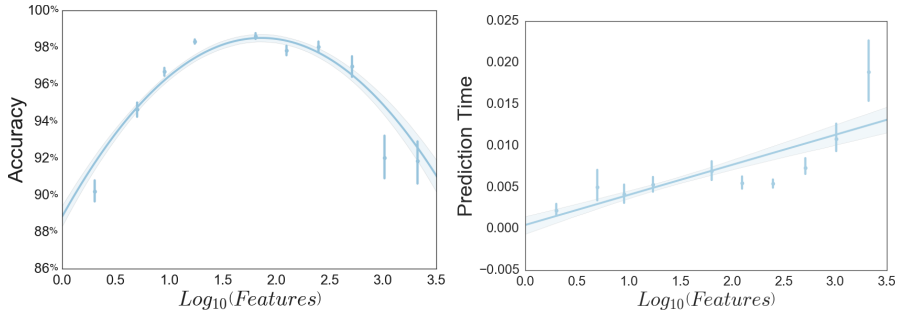
nals were better than PCA signals ($t = 4.5$, $p < 0.05$).

5.4 Feature Selection

One of the main premises of the proposed approach is that greedy feature selection from an off-the-shelf feature selector is effective in reducing computational cost while achieving high accuracy on the final classifier. Figure 4 (a) shows a plot of the mean accuracy of top hundred classifiers employed for every dataset as a function of the logarithm of the number of features used. As the figure clearly shows, the classification accuracy achieves a maximum between 10 and 31 features (which agrees with the findings in Table. 1) when most confusing features are pruned away by the system. Figure 4 (b) shows that the prediction time of the same classifiers changes linearly – as expected – with the number of features over the same range. This suggest that feature selection is effective for this problem. For more details, Fig. 5 shows a breakdown of the results for each dataset studied (results on the combined datasets is not reported due to lack of space). The same insensitivity to the number of features appears in all datasets.

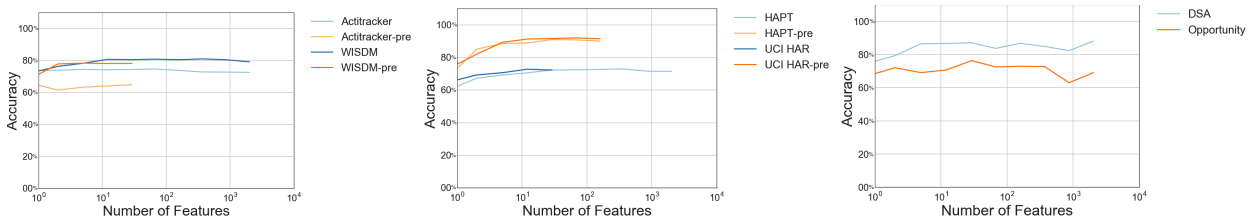
5.4.1 Feature Extractors

Table 2 shows a summary of the best accuracy obtained us-



(a) Accuracy (%) (b) Prediction Time (second)

Fig. 4 The effect of feature selection on classification accuracy and prediction time.



(a) WISDM Datasets (b) UCI HAR Datasets (c) Sensors Datasets

Fig. 5 The effect of the number of features selected on interpersonal accuracy.

Table 2 Effect of feature extraction.

Dataset	Methodology	Statistical	TSFRESH
WISDM	Within-session	90	**95**
	Personal	93	**94**
	Interpersonal	97	**83**
HAPT	Within-session	**96**	89
	Personal	**97**	90
	Interpersonal	**91**	73
UCI HAR	Within-session	**97**	90
	Personal	**99**	81
	Interpersonal	**93**	73
Actitracker	Within-session	**95**	94
	Personal	92	**96**
	Interpersonal	66	**75**
Combined	general	69	**76**
	Interpersonal	71	**75**

ing the general-purpose TSFRESH library as compared with statistical features designed specifically for each dataset. The handcrafted features outperformed the TSFRESH extractor in the UCI HAR and HAPT datasets while having roughly the same performance on WISDM dataset and the TSFRESH suit outperformed the handcrafted features in the interpersonal evaluation on the Actitracker dataset. Three factors contribute to this finding. Firstly, the collection protocol was most rigid on the UCI HAR and HAPT datasets followed by the WISDM dataset then the Actitracker dataset. With more control over the data collection, simpler handcrafted statistical features were effective. Secondly, the UCI HAR and HAPT features included both frequency domain and time-domain features which enhanced their performance. Finally the UCI HAR datasets used far more features (386) compared with the WISDM datasets (43).

Figure 2 shows that in the combined datasets, TSFRESH outperformed simple statistical features with a 75% accuracy compared with 71% in *general* evaluation and 76% compared to 69% for interpersonal evaluation. Statistical analysis using t-test shows that this difference is statistically significant ($t = 11.7612812302$, $p < 0.001$).

5.4.2 Feature Selectors

The performance of different approaches to feature selection was evaluated by calculating the average performance of top hundred classifiers from every type tested. The combined selector — even though more computationally expensive – does not outperform traditional tree-based classifiers. On the other hand, the proposed switching selector outperforms all other selectors consistently with between 1% (ET) and 4% (MLP) improvement in accuracy. The performance of the switching selector is also most stable with a total accuracy range of 18% on the top 100 classifiers compared with 23% for the L_1SVM selector, and 26% for the Tree selector.

Considering the results in Table 1, the proposed switching strategy appear 6 times compared with 4 times for the second best performing selector (Combined and Tree). Applying t-tests to check the differences between selectors with Bonferroni's multiple-comparisons corrections, it was found that the proposed switching selector achieved higher performance compared with L_1SVM ($t = 11.83, p < 0.001$), Tree based selector ($t = 4.46, p < 0.001$), Kernel PCA ($t = 7.19$, $p < 0.001$), and PCA ($t = 10.14$, $p < 0.001$). Its performance was on bar with the more computationally intensive combining selector ($t = 2.64, p = 0.008 > 0.05 * 42$). Figure 6 shows the instability of each selector measured as

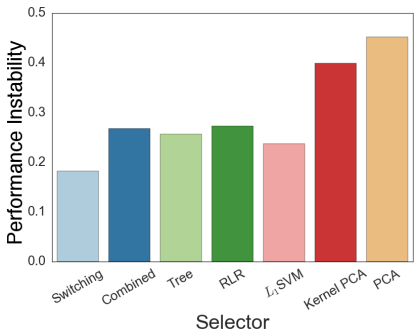


Fig. 6 Performance Instability (Accuracy range for the top 100 classifiers) of different selectors.

the range of accuracies for the top 100 classifiers. It is clear that the proposed switching classifier has the best stability compared with all others.

6. Discussion

The results presented in the previous section can be summarized as follows: Orientation invariant signals can achieve better performance than relying on raw accelerometer data for activity recognition in the more realistic general evaluation strategy. Moreover, feature selection from an off-the-shelf time-domain feature extractor is an effective strategy for reducing the number of required feature evaluations at real time while preserving the accuracy of final classifiers. The proposed feature selection approaches based on combining and switching selectors are effective in preserving the overall accuracy of the system. The switching selector specially can achieve this overall system improvement with the same number of feature selection calls as traditional selectors.

Figure 2 (c) shows that *general* and interpersonal evaluations lead to the same performance even for databases that have different collection procedure (e.g., WISDM and UCI HRI) that is around the same performance of cross-validation on the worst of the combined datasets. Moreover, Fig. 2 shows that in all cases, personalization is useful as it improves accuracy on all datasets (with up to 29% for the Actitracker dataset using handcrafted features). These results agree with the findings of Weiss et al. [28] but extends it to more databases and different feature extraction and selection methods.

One limitation of the analysis conducted in this paper is the reliance solely on accelerometer data. Using other sensors like the magnetometer and gyroscopes can improve accuracy as previous studies have shown [2], [10], [34] at the expense of a larger energy footprint. We only analyzed data from sensors that correspond to phones on the pocket or belt. With the increased availability of smart watches a new venue of research into activity recognition of hand and full body motion is opened. Our initial analysis of the data in DSA and Opportunity datasets show that the sensor on the right hand (the location of a smart-watch) is even more accurate in recognizing both locomotion related and low level hand actions but not high level behaviors. Due to lack of space these results will be omitted from this paper.

One obvious direction of future research for this work is building a smart-phone application to assess the proposed approach in the real world and compare its computational and energy requirements on the target systems to other approaches. Another direction is to include more features in the selection step from the frequency domain. Results from [8] and [32] suggest deep architectures can be effective for feature extraction. As discussed in Sect. 2 though, these methods tend to employ thousands of neurons and/or precalculation of time-frequency features (e.g., spectrogram) which makes them computationally inefficient even at the recognition stage. Network compression can alleviate this problem but at the cost of reduced accuracy. To test this hypothesis, we trained a CNN using the method of Zeng et al. [9] and applied the network compression method proposed by Han et al. [35] which proved effective for ImageNet and other image sets. The final number of weights dropped from 43,188 to 22,105 weights on the top two layers but the accuracy decreased to 91.36% on the Actitracker dataset and 72.9% on the Opportunity dataset (lower than the accuracies reported in this paper). The proposed method shares the advantage of requiring no manual tinkering to design features but lacks the ability of utilizing unlabeled data. Combining these two approaches to feature engineering will be one of our directions of future research.

7. Conclusion

This paper presented a systematic evaluation of different signal extraction and feature selection methods on a set of publicly available smart phone and on-body activity recognition datasets with numbers of activities ranging from 4 to 19. The main findings of the paper are: (1) Invariant signals extracted from accelerometer data allow for more generalizable performance across different datasets. (2) Using an off-the-shelf feature selector combined with aggressive wrapper based feature selection can achieve state-of-the-art performance on most datasets (with the single exception of UCI HAR dataset) with fewer features. The proposed approach provides an appealing alternative to deep learning methods for activity recognition on mobile devices as it provides approximately the same accuracy using much less computational resources. In the future, we will explore using the proposed feature selection approach in combination with feature learning using deep methods.

References

[1] R. San-Segundo, J.M. Montero, R. Barra-Chicote, F. Fernández, and J.M. Pardo, "Feature extraction from smartphone inertial signals for human activity segmentation," Signal Processing, vol.120, pp.359–372, March 2016.

[2] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J.d.R. Millàn, "Collecting complex activity datasets in highly rich networked sensor environments," 2010

Seventh International Conference on Networked Sensing Systems (INSS), pp.233–240, IEEE, June 2010.

[3] J.R. Kwapisz, G.M. Weiss, and S.A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, vol.12, no.2, pp.74–82, Dec. 2011.

[4] Z.S. Abdallah, M.M. Gaber, B. Srinivasan, and S. Krishnaswamy, "Adaptive mobile activity recognition system with evolving data streams," Neurocomputing, vol.150, no.Part A, pp.304–317, Feb. 2015.

[5] N. Raman and S.J. Maybank, "Activity recognition using a supervised non-parametric hierarchical hmm," Neurocomputing, vol.199, pp.163–177, July 2016.

[6] A. Veenendaal, E. Daly, E. Jones, Z. Gang, S. Vartak, and R.S. Patwardhan, "Sensor tracked points and hmm based classifier for human action recognition," Computer Science and Emerging Research Journal, vol.5, 2016.

[7] G.M. Weiss, J.W. Lockhart, T.T. Pulickal, P.T. McHugh, I.H. Ronan, and J.L. Timko, "Actitracker: A Smartphone-Based Activity Recognition System for Improving Health and Well-Being," 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp.682–688, Oct. 2016.

[8] M.A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.P. Tan, "Deep Activity Recognition Models with Triaxial Accelerometers," TheWorkshops of the Thirtieth AAAI Conference on Artifi. Intelli., pp.1–8, Nov. 2015.

[9] M. Zeng, L.T. Nguyen, B. Yu, O.J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," Applications and Services (MobiCASE), 2014 6th International Conference on Mobile Computing, pp.197–205, IEEE, Nov. 2014.

[10] B. Barshan and M.C. Yüksek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," The Computer Journal, vol.57, no.11, pp.1649–1667, Nov. 2014.

[11] T. Plötz, N.Y. Hammerla, and P. Olivier, "Feature Learning for Activity Recognition in Ubiquitous Computing," IJCAI'11 Proc. Twenty-Second international joint conference on Artifi. Intelli., vol.2, pp.1729–1734, 2011.

[12] A. Wang, G. Chen, J. Yang, S. Zhao, and C.Y. Chang, "A Comparative Study on Human Activity Recognition Using Inertial Sensors in a Smartphone," IEEE Sensors J., vol.16, no.11, pp.4566–4578, 2016.

[13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp.437–442, April 2013.

[14] D.M. Karantonis, M.R. Narayanan, M. Mathie, N.H. Lovell, and B.G. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," IEEE Trans. Inf. Technol. Biomed., vol.10, no.1, pp.156–167, Jan. 2006.

[15] D. Mizell, "Using gravity to estimate accelerometer orientation," Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings., pp.252–253, 2003.

[16] Y. Mohammad and T. Nishida, "On comparing ssa-based change point discovery algorithms," 2011 IEEE/SICE International Symposium on System Integration (SII), pp.938–945, 2011.

[17] M. Christ, A.W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," arXiv preprint arXiv:1610.07717, 2016.

[18] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol.40, no.1, pp.16–28, Jan. 2014.

[19] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," The annals of statistics, vol.28, no.2, pp.337–407, 2000.

[20] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, "Energy efficient smartphone-based activity recognition using fixed-point arithmetic," J. Universal Computer Science, vol.19, no.9, pp.1295–1314, 2013.

[21] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol.7657, LNCS, pp.216–223, 2012.

[22] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine learning, vol.63, no.1, pp.3–42, April 2006.

[23] L. Breiman, "Random forests," Machine learning, vol.45, no.1, pp.5–32, Oct. 2001.

[24] C.M. Bishop, Pattern Recognit., 2006.

[25] J.W. Lockhart, G.M. Weiss, J.C. Xue, S.T. Gallagher, A.B. Grosner, and T.T. Pulickal, "Design considerations for the wisdm smart phone-based sensor mining architecture," Proc. Fifth International Workshop on Knowledge Discovery from Sensor Data, pp.25–33, ACM, 2011.

[26] J.L. Reyes-Ortiz, L. Oneto, A. Ghio, A. Samá, D. Anguita, and X. Parra, "Human activity recognition on smartphones with awareness of basic activities and postural transitions," Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol.8681 LNCS, pp.177–184, 2014.

[27] Z. Wu and S. Zhang, "Human Activity Recognition using Wearable Sensors," tech. rep., Stanford University, 2015.

[28] G.M. Weiss and J.W. Lockhart, "The impact of personalization on smartphone-based activity recognition," AAAI Workshop on Activity Context Representation: Techniques and Languages, pp.98–104, 2012.

[29] A.D. Ignatov and V.V. Strijov, "Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer," Multimedia tools and applications, vol.75, no.12, pp.7257–7270, June 2016.

[30] S. Dharia, V. Jain, J. Patel, J. Vora, S. Chawla, and M. Eirinaki, "PRO-Fit: A personalized fitness assistant framework," pp.386–389, July 2016.

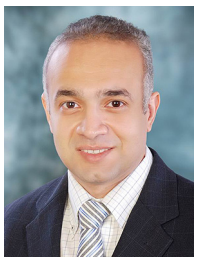
[31] G.M. Weiss and J.W. Lockhart, "The Impact of Personalization on Smartphone-Based Activity Recognition," AAAI Workshop on Activity Context Representation: Techniques and Languages, pp.98–104, 2012.

[32] M. Zeng, L.T. Nguyen, B. Yu, O.J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors," Proc. 6th International Conference on Mobile Computing, Applications and Services, pp.197–205, ICST, 2014.

[33] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," Pattern Recognit., vol.43, no.10, pp.3605–3620, Oct. 2010.

[34] K. Altun and B. Barshan, "Human Activity Recognition Using Inertial / Magnetic Sensor Units," Human Behavior Understanding, pp.38–51, 2010.

[35] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," Advances in Neural Information Processing Systems, pp.1135–1143, 2015.



Yasser Mohammad is a Researcher at KDDI Research Inc., Japan and an Associate Professor at Assiut University, Egypt. He received his Ph.D. from Kyoto University, Japan in 2009 in the area of intelligence science and technology after receiving a master and bachelor degrees in Electrical Engineering from Assiut University, Egypt. Recipient of four best paper awards from IEA/AIE 2008, IEA/AIE 2009, IEEE/SICE SII 2011, and ICCAS 2012. Author of Conversational Informatics: A data intensive approach (Springer, 2015) and Data mining for Social Robotics (Springer, 2016). His research focuses on applications of data mining techniques to time-series data, robotics, and HRI. Published over 75 international publications in these areas with an h-index of 15. Co founder of Ayonix Inc., Japan and Tebex IT, Egypt and a senior IEEE member.

Kazunori Matsumoto received the B.E. and M.E. degree in Information Science from Kyoto University in 1984 and 1986, and Ph.D. degree from Ritsumeikan University in 2009, respectively. He has been working at KDDI R&D Laboratories Incorporated since 1984 and now is a research manager of the Intelligent Media Group. He has engaged in the research areas of multimedia search and content delivery. He was awarded the JSAI Incentive Award from the Japanese Society for Artificial Intelligence in 1998, and the Best Paper Award from IEICE in 2000.

Keiichiro Hoashi received the B.E. and M.E. degree in Information Science from Waseda University in 1995 and 1997, and Ph.D. degree from Waseda University in 2007. He has been working at KDDI Research Inc. since 1997, and is currently the Senior Manager of the Intelligent Media Laboratory. He has been mainly engaged in the research areas of multimedia information retrieval. He was awarded the FIT 2004 Young Researcher award, and also worked as a part-time lecturer at Waseda University from 2001 to 2005. He was also selected as a member of the Sido 2015 Next Innovator program sponsored by METI (Ministry of Economy, Trade and Industry).