Online Personalization of Cross-Subjects based Activity Recognition Models on Wearable Devices

Timo Sztyler
University of Mannheim
Mannheim, Germany
timo@informatik.uni-mannheim.de

Heiner Stuckenschmidt
University of Mannheim
Mannheim, Germany
heiner@informatik.uni-mannheim.de

***Abstract*—Human activity recognition using wearable devices is an active area of research in pervasive computing. In our work, we address the problem of reducing the effort for training and adapting activity recognition approaches to a specific person. We focus on the problem of cross-subjects based recognition models and introduce an approach that considers physical characteristics. Further, to adapt such a model to the behavior of a new user, we present a personalization approach that relies on online and active machine learning. In this context, we use online random forest as a classifier to continuously adapt the model without keeping the already seen data available and an active learning approach that uses user-feedback for adapting the model while minimizing the effort for the new user. We test our approaches on a real world data set that covers 15 participants, 8 common activities, and 7 different on-body device positions. We show that our cross-subjects based approach performs constantly $+3\%$ better than the standard approach. Further, the personalized cross-subjects models, gained through user-feedback, recognize dynamic activities with an F-measure of 87% where the user has significantly less effort than collecting and labeling data.**

I. Introduction

Human Activity Recognition is an active field of research in pervasive computing [1]–[3]. For more than a decade, researchers have been focusing on inertial sensors to recognize physical human activities to support people in everyday life. Due to the development of wearable devices such as smartphones, smart-watches, and smart-glasses new opportunities and challenges arise. People wear these devices all day long which enable continuous monitoring for the purpose of personal assistant and health care management. Hence, the focus shifted from laboratory to real-world applications. However, most of the existing works target subject-specific activity recognition which requires labeled training data of the target person. In context of real world applications, for example in healthcare, it is often infeasible to ask users to collect the required amount of data and where such a system should be usable as fast as possible. For instance, the study in [3] relied on data that was generated by a set of subjects that had to perform physical activities for more than an hour based on a fixed protocol. In order to reduce this effort, we can use a cross-subjects activity recognition approach which considers labeled data of several other people and adapts the resulting model to new subjects at hand. This adaptation is of general use as also movement patterns of the same person can change over time, e.g. due to age, fitness level, and injuries. We believe that our approach significantly reduces the effort for patients and elders but also lowers the barriers. In a healthcare scenario, an initial set of labeled activity data for base models could be acquired from a representative sample of people with different physiologies in the course of a clinical study. These models could be adapted to real patients using the adaptation methods described in this paper on the fly while their activities are tracked.

Several studies already investigated activity recognition in a real world scenario relying on an acceleration sensor. They considered activities like climbing stairs, walking, and standing and achieved in context of subject-specific approaches good results [1], [4]. In contrast, cross-subjects approaches perform often significantly worse [5], [6] where researchers mainly focused on the leave-one-subject-out method, i.e., they consider all available labeled data except data of the target person. Researchers stated that the lower performance is due to the different acceleration patterns, e.g., children walk in a different way as elders. They hypothesized that physical characteristics could be reliable indicators to build meaningful groups [2], [6]. However, so far this was not entirely investigated and is still an open issue.

In this paper, we present an evolving cross-subjects based activity recognition approach which is based on an online learning version of a random forest classifier[1]. Compared to other classifiers, previous works stated that the random forest performs the best in this scenario [3], [7]. To build the initial activity recognition model, we rely on labeled sensor data of people that have similar physical characteristics as the target person. Afterwards, this model is personalized. In this context, online learning enables to adapt the model without retraining or keeping the processed training data available. The information to personalize the model, i.e., to improve the recognition rate, is gathered by analyzing the classified data. These classification results enable to perform active learning, i.e., to query the user. In our scenario, we consider the activities climbing stairs up and down, jumping, lying, running, standing, sitting, and walking where we rely on acceleration sensors. The accelerometer is the most interesting sensor due to low power consumption as well as the already

[1]In this paper, we distinguish between *online* and *incremental* learning. The former does not store the processed data where the latter keeps all data available.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

presented results in previous works [1], [2], [4]. During the experiments, we evaluate single and multi-sensor setups as well as distinguish between the relevant on-body device position, i.e., head, upper arm, forearm, chest, waist, thigh, and shin [3].

The main contributions of our work are the following:

- We perform comprehensive experiments regarding cross-subjects models in context of offline and online learning with single and multi-acceleration sensor setups including all common activities and on-body positions.
- We show that our group-based cross-subjects approach performs constantly 3% better than leave-one-subject-out where we are able to achieve an F-measure of 78%.
- We present an activity recognition approach that personalize with online and active machine learning cross-subjects based models and achieves a recognition rate of 84% but dynamic activities even with 87%.

The paper is structured as follows: In Section II, the related work concerning evolving and adaptive activity recognition is summarized followed by the description of our data set. Section IV introduces the online random forest classifier. Afterwards, we describe the method and strategy of our approach in detail. In Section VI, we present our experimental results and discuss the effect of online and active machine learning. Finally, Section VII covers the conclusion and future work of this paper.

II. Related Work

Subject-specific activity recognition has been extensively investigated by many researchers [1], [3], [4], [8]. They achieved reliable recognition rates in many different scenarios but required for each subject a labeled training set. Further, changes in user's motion patterns are often not considered which leads to worse recognition rate over time.

As a first approach to reduce the need of labeled data, researchers have investigated cross-subjects approaches. Especially, the leave-one-subject-out approach was evaluated comprehensively and researchers stated that it performs significantly worse compared to a subject-specific classifier [6], [9], [10]. This even holds if several acceleration sensors are considered simultaneously [10]. The researchers conclude that this is due to differences in the physical characteristics of the considered subjects, i.e., fitness level, gender, and body structure. Indeed, researchers hypothesize that these kinds of characteristics could be reliable indicators to identify subjects with similar acceleration data [2], [6]. So far, this assumption was only considered in few works. Maekawa et al. [11] applied this concept successfully but also state that a minimum number of subjects are required. However, the authors relied on five acceleration sensors and also considered activities of daily life (e.g., play pingpong) which makes it difficult to interpret the aggregated results. Besides, in some works models were trained on one person and used on another without considering any characteristics [5], [12], [13]. They state that such a model often cannot yield accurate results if it is used on different subjects and that a personalization is required. In our work, we will focus on this hypothesis but also investigate cross-subjects approaches concerning their performance in context of all relevant on-body device positions and combinations.

Instead of using labeled training data across subjects, several researchers also investigated semi-supervised approaches, e.g., active learning, to reduce the labeling effort [14], [15]. These works aim at extracting the most informative unlabeled samples to minimize the user interaction. By using active learning, the user could be queried regarding these samples to gain new knowledge. Their results show that active learning does improve the learning performance and also that it is possible to achieve comparable recognition rates with respect to a supervised approach [14]. In this context, the most informative unlabeled samples could be identified by interpreting the classifiers confidence values [16]. However, this approach still requires a small, initial labeled data set of the target user.

Indeed, using labeled data across subjects and active learning do not exclude each other but are complementary. Hence, the labeled data could be used across subjects to build a base model which could be personalized by information which was gathered by active learning. So far, personalization of an existing activity recognition model was realized by updating parameter of an existing model [17], [18], or incremental learning [19]–[21]. In this context, researchers evaluated neural network [20], [22], support vector machine [23], and fuzzy rule [24], [25] based approaches and even if the results of these works are difficult to compare due to the different setups, the results show that the concept of personalization is feasible. Besides, to gather additional information to personalize a model, researchers also focused on the unlabeled sensor data and applied successfully the concept of co-training [14], [26].

So far, concerning our scenario, nobody combined all of these techniques where in addition especially the mentioned personalization approaches have limitations. Concerning parameter adaption, the structure of the model is almost fixed where incremental learning has to keep all data available and commonly also does not distinguish between newer and older information. Indeed, some of these works also performed retraining to process new gathered data which is often unfeasible. For that purpose, we will rely on and investigate an online version of the random forest classifier which overcomes all of these limitations. In this context, the influence and performance concerning the users' effort that goes along with active learning or the relation concerning the number of uncertain samples, queries, and achieved improvement is also unclear.

III. Data Set

In this paper, we investigate cross-subjects based models that should be personalized by relying on online and active machine learning. In this context, we consider the physical characteristics of our subjects to identify people with similar acceleration patterns. We focus on all common on-body device positions and activities and evaluate their influence on the quality of activity recognition.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

Fig. 1. Sensor placement. Each subject wore the wearable devices on the head, chest, upper arm, waist, forearm, thigh, and shin (top down).

In a previous work [3], we created a data set[2] where we focused on a real world scenario. The data set covers, among others, the acceleration data of the activities climbing stairs down (A_1) and up (A_2), jumping (A_3), lying (A_4), standing (A_5), sitting (A_6), running/jogging (A_7), and walking (A_8) of fifteen subjects (age 31.9 ± 12.4, height 173.1 ± 6.9, weight 74.1 ± 13.8, eight males and seven females). For each activity, we recorded simultaneously the acceleration of the body positions chest (P_1), forearm (P_2), head (P_3), shin (P_4), thigh (P_5), upper arm (P_6), and waist (P_7). Each subject performed each activity roughly 10 minutes expect for jumping due to the physical exertion ($\sim$1.7 minutes). In detail, we recorded for each position and axes 1065 minutes where in addition everything was also captured by a video camera (third-person). Our data set covers several different groups of people, e.g., young, old, slim, strong, athletic, and nonathletic. Concerning male and female, the amount of data is equally distributed. These allow to investigate the physical characteristics in relation to the recorded acceleration data.

The data was collected by using customary smart-phones and a smart-watch[3] which were attached to the mentioned positions (see Figure 1). The devices were synchronized with the time service provider and the inertial sensors were sensed with a sampling rate of 50 Hz where the data was stored on a local SD card. The sampling rate was chosen with consideration of battery life as well as with reference to previous studies [4], [12]. The recording of the data was performed using a self-developed sensor data collector and labeling framework[4].

To attach the devices to the mentioned body positions, common objects and clothes were used such as a sport armband case, trouser pocket, shirt pocket, or the bra. There was no further fixation of the devices to closely resemble their use in everyday life. In case of the head we used a belt to avoid that the subject had to hold this device during the performance of the activities. This simulates that the subject phones during the activities, alternatively it could be considered as smart-glasses.

The data collection took place under realistic conditions, i.e., the subjects walked through a city, jogged in a forest, or climbed up the stairs of a guard tower of an old castle. The order of the activities was left to the subjects but they were instructed to stand idle for a few seconds before and after an activity was performed. There were no further instructions concerning how the activities should be performed or how fast they have to walk. For instance, while sitting, the subjects typically used their smart-phone, talked with somebody else, or were eating and drinking something.

Beside the accelerometer, we also recorded the GPS, gyroscope, light, magnetic field, and sound level sensors. However, in this work, we focus only on the acceleration data. To facilitate the usage of our data set also by other people, we recorded each movement also by a video camera. Our data set is free available[2] and covers also a detailed description of each subject including images of the attached devices.

Compared to the well known data sets *OPPORTUNITY* [27] and *COSAR* [28], we did not focus on activities of daily living but physical activities. Indeed, it would be possible to derive the physical activities from the activities of daily living and both data sets also cover acceleration data from on-body devices, however, several aspects and activities are not covered. On the one hand, *OPPORTUNITY* covers several different on-body positions but provides only one single dynamic activity (walking) where on the other hand the *COSAR* data set covers several different physical activities but provides only acceleration data for two on-body positions. Besides, both data sets cover significant fewer subjects (four and six) which are too few to analyze and compare physical characteristics.

IV. Online Random Forest

Random Forests [31] are usually used in context of computer vision and machine learning applications where they achieve state-of-the-art results [3], [7]. They are ensembles of randomized decision trees combined with bagging for reducing variance and boosting to handle the bias. The result of a random forest classifier is made robust against overfitting and outliers by ensuring that the individual trees whose results are combined are uncorrelated.

A random forest R consists of T individual decision trees t_i where T is predefined. Considering a training set $D = \{(x_1, c_m), ..., (x_n, c_o)\}$ where $d \in D$ is a sample consisting of a feature vector x_i and a corresponding class $c \in C$. Each tree is initialized with a randomly selected subset of tests $g(x) > \theta$ where g is a function that maps the sample to a scalar value (a feature) and θ is a threshold for deciding whether the sample will go to the left or the right subtree. Subsequently, the entropy of each of the selected features is calculated (e.g., Information Gain or Gini Index) and the node is split on the feature that maximizes the quality measurement. The threshold for each feature at each node is chosen by

[2]http://sensor.informatik.uni-mannheim.de#dataset_realworld
[3]Samsung Galaxy S4 an LG G Watch R
[4]https://github.com/sztyler/sensordatacollector

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

Algorithm 1 OnlineBagging(R, L_o, d) [29]

for each base model $t_i \in R$, $i \in \{1, 2, ..., T\}$ **do**
  Set k according to $Poission(1)$.
  do k times
    $t_i = L_0(t_i, d)$.
end for

Algorithm 2 OfflineBagging(T, L_b, D) [30]

for each $i \in \{1, 2, ..., T\}$ **do**
  D_i = Sample_With_Replacement(D,$|D|$)
  $t_i = L_b(D_i)$
end for
Return $\{t_1, t_2, ..., t_T\}$

analyzing the samples that belong to the node or can also be chosen at random. In case of a random choice, R is named an extremely randomized forest [32]. The resulting set of uncorrelated decision trees is used to determine the class of an unseen feature vector x'_i. In particular the result is determined by averaging over the predicted results of all individual decision trees as follows:

$$p_R(c|x'_i) = \frac{1}{T}\sum_{k=1}^{T} p_{t_k}(c|x'_i) \tag{1}$$

where the resulting class is $C(x'_i) = \arg\max_{c \in C} p_R(c|x'_i)$.

Considering the random forest classifier in online mode, the main differences are the implementation of bagging, i.e. the generation of subsamples used for constructing the individual trees, and the growing of the individual random decisions trees.

It has been proven that bagging improves the predictive power of random forests by generating replicated bootstrap samples of the training set D [30]. Hence, for each decision tree t_i, the training set is sampled with replacement, so, the set keeps the same size but some instances that occur in the original training set may not appear where others could appear more than once. This requires that the whole training set has to be available at once. Oza [29] introduced an online version of bagging (see Algorithm 1) where the number of occurrences of a sample for training an individual tree is drawn from a Poisson distribution with a constant parameter. This means that the subsample for a tree can be determined on the fly as a new sample becomes available. Oza provides both theoretical and experimental evidence that the results of online bagging converges towards the results of offline/batch bagging (see Algorithm 2).

The growing of an online decision tree based on the concept of an extremely randomized tree. As in the beginning, the complete data set is not available, split decision are postponed until enough information is available. This is guided by two parameters: the minimal number of samples that have to be seen before deciding and the minimal quality measurement that has to be achieved by the split. In order to be able to further construct the decision tree, statistics about class membership of new samples are propagated through the tree and provides the basis for computing the quality measurement of possible splits. As these statistics can easily be updated on the fly, the trees are refined as new samples arrive. In order to compensate for changes in the distribution of arriving information, the results can be adapted by deleting trees whose performance degrade with new information.

Saffari et al. combined the introduced concepts, i.e., online bagging, online decision trees, and random feature selection and developed the first publicly available version of an online random forest [33]. They presented experiments which show that the random forest in online mode converged to the results that were achieved in offline mode. Besides, the authors implemented this classifier in C++. As we want perform the activity recognition on wearable devices, i.e. on an Android platform, we reimplemented this classifier in Java. We repeated the experiments performed by Saffari et al. [33] and achieved the same results. Further, we enhanced the original implementation by implementing threading, incremental learning, and information gain as a quality measurement to split nodes. Our implementation is also publicly available[5].

V. Method

In the following, we introduce the feature generation from the acceleration sensor data and subsequently our proposed approach for cross-subjects based activity recognition and personalization.

A. Feature Extraction

The essential idea behind generating features from time depended data streams is to segment the recorded data into windows and compute a feature vector for each window. Several preceding studies in the context of activity recognition already examined different settings regarding the window size [34] and meaningful features [35]. They state that overlapping windows are more suitable because they can handle transitions more accurately. Further, the window size depends on the kind of activities which should be recognized. In our context, most of the existing studies considered a size between one and three seconds [34].

In [3], we investigated the most common time- and frequency-based features that were considered in previous works (see Table I). Our earlier experiments showed that especially in context of the random forest it is to prefer to consider all of these features because this classifier is robust against outlier and also considers the entropy of the individual features. Hence, meaningless features do not affect negatively the result where a reduction of the considered feature set could lead to lower recognition rates. Besides, comparing feature sets of related work is difficult due to the different setups. However, some researchers have compared different groups of features and also state explicitly that frequency-based features improve the accuracy of the recognition [35].

In addition to these features, we also computed gravity-based features that provide information of the device orientation. The gravity component was extracted from the recorded

[5]http://sensor.informatik.uni-mannheim.de/#onlineforest

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

TABLE I
SUMMARY OF CONSIDERED STATISTICAL FEATURES.

	Methods
Time	Correlation coefficient (Pearson), entropy (Shannon), gravity (roll, pitch), mean, mean absolute deviation, interquartile range (type R-5), kurtosis, median, standard deviation, variance
Frequency	Energy (Fourier, Parseval), entropy (Fourier, Shannon), DC mean (Fourier)

accelerometer data. We applied a *low-pass filter*[6] to separate the acceleration and gravitational force to derive the gravity vectors. These vectors allow to determine the orientation of the device by computing the angles between them, also known as *roll* and *pitch*. The *azimuth* angle, however, cannot be calculated because the direction of north is required. This means that it is not possible to derive if the device is back-to-front. Further, we only consider absolute values of the acceleration so that we do not distinguish if the device is upside down. Hence, we consider these four cases as the same position. To be more flexible and avoid overfitting, we also transform the roll and pitch angles in one of sixteen predefined discretized orientations. The gravity-based features are only considered in the context of static activities to distinguish between standing and lying.

Summarizing, we use windows which overlap by half and have a length of one second. Further, we consider the most common time- and frequency-based features where time-based feature values are transformed into frequency-based ones by applying *Discrete Fourier transform*[7].

B. Cross-Subjects Activity Recognition

Cross-subjects activity recognition focuses on building an initial classification model for a specific user by relying on labeled acceleration data of other users. The most common approach is leave-one-subject-out where the classifier is trained on all labeled training data except the labeled data that correspond to the target user. In contrast, we aim to build a cross-subjects model for a user that relies only on labeled data of users that might have similar acceleration patterns. We believe that this is promising concerning the recognition rates.

To determine which users have similar acceleration patterns, we rely on physical characteristics, i.e., gender, fitness, and physique. In preliminary experiments, we identified that people who have the same gender and physique or fitness also have similar acceleration patterns. However, typically people do not have exactly the same physical characteristics but only some characteristics are similar. As a result, these people have comparable acceleration patterns for some activities but not for all. The idea is that people with the same fitness level also have similar acceleration patterns regarding running whereas gender and physique could be characterizing concerning walking.

[6] A low-pass filter passes values which have a lower frequency as the specified *cutoff frequency* and attenuates values that have a higher frequency.

[7] The Fourier transformation can be applied with different scaling factors. We use the *JTransforms* implementation (https://github.com/wendykierp/JTransforms) which scales by 1.

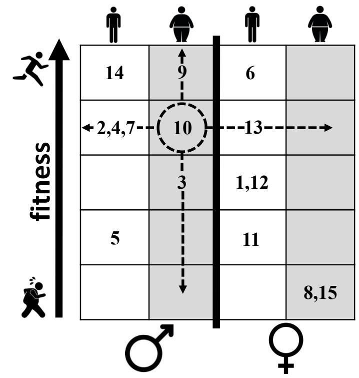


Fig. 2. Our cross-subjects based activity recognition approach. A subject has similar acceleration patterns to people in the same row and column.

Figure 2 shows how we classified and grouped our subjects. For instance, if we want to build a cross-subjects based activity recognition model for subject 10 then we consider the labeled data of all subjects that are in the same row (same fitness level: 2, 4, 7, 13) or column (same gender and physique: 3, 9). This means that we follow a group-based approach, where the groups are dynamically determined and can overlap for different subjects. The gender and physique were derived from the recorded images where we distinguished between a strong and slim physique. The fitness level results from the distance that the subject covered in 10 minutes running. We focused on a practical and feasible classification system for lower barriers and easy adoption.

C. Personalization: Online and Active Learning

Online learning enables to evolve an existing model without keeping the whole data set available. The model is adapted over time to the behavior of a user where recent received information is more weighted than older. In this context, we use online learning to adapt a cross-subjects model by new information that is gathered from the classified windows. In the following, we introduce the techniques *smoothing* and *user-feedback* which we apply to gather this information. Both techniques are applied separately (see Figure 3).

We apply *smoothing* if a single classified window is surrounded by windows that belong to another activity. More

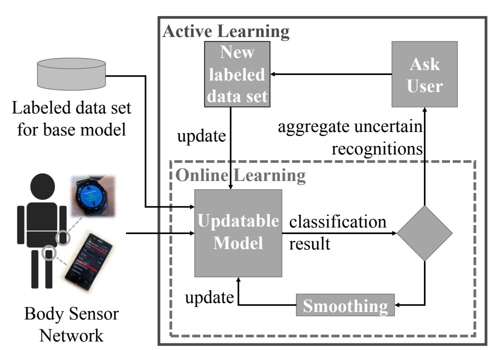


Fig. 3. Personalization of a cross-subjects based model by online and active machine learning. This approach analyzes the classified windows regarding their uncertainty to gather new information.

precisely, if two preceding and two succeeding windows have the same class but another than the surrounded then the label is adjusted. The sample of the adjusted record is also used to update the model. Concerning *user-feedback* (active learning), we ask the user for feedback on certain samples that have been classified with a low confidence. As it is unfeasible to ask the user for a specific window we analyze and cluster the classified windows for a specific time interval. If several classified windows with a low uncertainty occur close to each other, we ask the user for that specific time interval. Based on preliminary experiments, we decided that a sequence of uncertain classified windows is interrupted if the distance between two uncertain windows is ≥ 5 seconds. Further, we only asked the user for feedback if a sequence was longer than 30 seconds. This value was chosen in regard to the amount of the available testing data. Figure 3 shows our approach in detail. The initial model classifies the acceleration data of the target user. Subsequently, the classified windows are analyzed to identify uncertain classified windows. These windows are used to gather new knowledge by *user-feedback* and *smoothing*.

The idea is that *smoothing* provides information regarding minor classification errors where *user-feedback* targets major classification errors. Hence, the resulting information from *user-feedback* and *smoothing* is combined to create a new small labeled data set to update the initial model. To maximize the information gathering, we focused on classified windows with a low uncertainty. Of course, the number of uncertain windows depends on a predefined threshold. Hence, during our experiments, we also consider several different confidence value thresholds and analyze the relation between uncertainty, user interaction, and gained recognition rate.

To evaluate the improvement of our recognition model over time, we perform five iterations of this approach. In this context, an iteration comprises that first the model has to process a certain amount of acceleration data where subsequently *user-feedback* and *smoothing* are performed separately. Afterwards, the model is updated with the gathered data and the new performance is measured. To avoid overfitting, we separated the data set of the target user in two equally sized parts where the classes are equally distributed. The one half is used to perform the introduced approach where the other half is considered to evaluate the performance of the evolving model. Hence, in each iteration, the model classifies new unseen acceleration data where the evolving model is always evaluated with the same data set. We repeat our experiments several times where we also consider other splits of the data sets to make the results more stable. For these experiments, we rely on the introduced online random forest classifier.

VI. Results

In the following, we present our results and outline the conducted experiments[8] to show the performance of the proposed cross-subjects activity recognition approach but also the effect of smoothing and active learning to personalize the model. The introduced methods were evaluated for each individual subject. Due to lack of space, we only present the aggregated results of all subjects. However, the individual results of each dimension (i.e., subject, activity, device position, and number of accelerometers) are available[9]. Unless otherwise specified, the provided results are aggregated over all two-part acceleration combinations (21).

A. Cross-Subjects Activity Recognition

In a first experiment, we compare our group-based recognition model with the standard leave-one-subject-out approach that is typically used in the literature. For each subject, we trained a single classifier on the data of all other people. To have an additional reference, we also considered a cross-subjects approach where number and people were chosen at random (*randomly*). We repeat the random approach ten times and present the average recognition rates.

TABLE II
Recognition rates (F-measure) of the introduced cross-subjects based approaches.

	Number of Accelerometers					
	1	**2**	**3**	**4**	**5**	**6**
Randomly	0.61	0.69	0.75	0.77	0.79	0.80
Leave-one-out	0.65	0.74	0.79	0.82	0.83	0.85
Our Method	**0.68**	**0.78**	**0.82**	**0.85**	**0.87**	**0.88**

Table II shows the results of the corresponding experiments and indicates that our group-based approach consistently performs better than the other approaches ($+3\%$) where *randomly* produces the worst results (-4.5%). Indeed, with an increasing number of accelerometers the gap between the recognition rates seems to remain stable. The results also show that the recognition rates are far worse than a subject-specific classifier (compare results on the same dataset from [3]). At least a four-sensor setup seems to be necessary to achieve even satisfying recognition rates. This is not feasible in a real world scenario and underlines the necessity for adapting the model to new individuals.

Considering the individual activities, Table III shows the recognition rates of the different approaches. We can see that our approach performs satisfying concerning all activities. Focusing on static (77.7%) and dynamic (78.2%) activities separately, points out that their recognition rates are similar but the rates for climbing stairs (A_1 and A_2, 69%) and walking (A_8, 70%) are lower. Varying movement speed and patterns of these activities cause these lower recognition rates. In contrast, running (A_7) and jumping (A_3) have significantly higher recognition rates because the strong acceleration is a reliable indicator. Indeed, considering the confusion matrix (not presented), climbing stairs and walking are activities that are often confused among each other. This problem occurs independently of the number of accelerometers.

[8]We provide our preprocessed files, i.e., the computed windows and features for easier verification of our results on our result website.

[9]http://sensor.informatik.uni-mannheim.de/#results2017online

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

TABLE III
RESULTS (F-MEASURE) SHOW THE RECOGNITION RATES FOR THE INDIVIDUAL ACTIVITIES OF THE CROSS-SUBJECTS APPROACHES.

	Offline Learning		
Class	**Randomly**	**Leave-one-out**	**Our approach**
A_1	0.62	0.66	**0.69**
A_2	0.63	0.67	**0.69**
A_3	0.79	**0.88**	0.87
A_4	0.81	0.83	**0.86**
A_5	0.71	0.73	**0.79**
A_6	0.59	0.63	**0.68**
A_7	0.88	0.90	**0.96**
A_8	0.60	0.67	**0.70**
$avg.$	0.69	0.74	**0.78**

TABLE IV
RECOGNITION RATES OF INTERESTING ACCELEROMETER/POSITION COMBINATIONS (OUR APPROACH).

	Offline Learning					
	P_2-P_5 (Watch & Phone)			P_3-P_5 (Glasses & Phone)		
Class	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
A_1	0.61	0.58	0.59	0.44	0.61	0.51
A_2	0.56	0.74	0.64	0.65	0.72	0.69
A_3	0.99	0.87	0.93	0.99	0.75	0.85
A_4	0.64	0.39	0.48	0.83	0.77	0.80
A_5	0.84	0.80	0.82	0.77	0.79	0.78
A_6	0.48	0.70	0.57	0.64	0.67	0.66
A_7	0.98	0.97	0.98	0.96	0.93	0.94
A_8	0.77	0.61	0.68	0.74	0.48	0.58
$avg.$	0.71	0.69	0.69	0.74	0.72	0.72

Subsequently, we investigated changes in the recognition rate for different combinations of sensors that are realistic in a real world setting, in particular thigh and forearm (smart-phone and smart-watch) and thigh and head (smart-phone and smart-glasses). Table IV summarizes these results. As we can see, these interesting combinations (smart-phone and smart-watch (69%) and smart-phone and smart-glasses (72%)) perform significantly worse than the best two-sensor combination, which is waist and shin. This is additional evidence that a cross-subjects based model needs personalization to be applicable in a real-world setting. In this context, it also points out that it depends on the set of activities that should be recognized which combination is most suitable. Besides, these results provide also evidence that the considered physical characteristics are reliable properties to identify which people can be considered for a group-based cross-subjects model. We analyzed the individual activities concerning all on-body device positions and combinations and in each case our approach performs equal or better. Certainly, due to the size of our data set, it is likely that there are further meaningful characteristics which we could not identify. However, these results confirm the hypothesis of previous works [2], [6].

B. Personalization: Online and Active Learning

The core idea of is this work is that feedback concerning the classification results improves the cross-subjects based activity recognition model. To confirm this thesis, we performed a series of experiments in improving the group-based recognition models using online and active learning. More precisely, first we analyze the difference in performance regarding offline and online learning. Subsequently, we investigate our introduced information gathering methods, i.e., *user-feedback* and *smoothing*, to personalize the model. Finally, we focus on the obtained activity recognition rate concerning certain aspects.

TABLE V
IMPROVEMENTS OF THE RECOGNITION RATE CONCERNING PERSONALIZATION OF THE BASE MODEL.

	Online & Active Learning			
Class	**Our method (Base)**	**+ Smoothing**	**+ User-Feedback**	**+ Smoothing & User-Feedback**
A_1	0.65	0.67	0.80	0.80
A_2	0.66	0.68	0.80	0.81
A_3	0.82	0.87	0.89	0.90
A_4	0.86	0.86	0.88	0.88
A_5	0.77	0.77	0.79	0.79
A_6	0.66	0.66	0.70	0.70
A_7	0.95	0.96	0.97	0.97
A_8	0.71	0.74	0.86	0.87
$avg.$	0.76	0.78	0.83	0.84

Table III and V illustrate the activity recognition rate for our approach in offline and online mode. It points out that in online mode the recognition rate is slightly worse (-2%). This is due to fact that in online mode the classifier does not know the whole data set a priori. Therefore, the chosen internal thresholds of the classifier concerning the node splits and features are coarser. In turn, this ensures that the trained classifier is not fitted to a specific data set. Further, the lower initial recognition rate of the base model is the drawback to enable updating of the model on the fly without knowing or storing preceding data.

Applying our personalization approach (*smoothing* & *user-feedback*) improves the recognition rate of the base model by $+8\%$ (see Table V). Considering the individual activities show that the recognition rate improves for all activities (up to $+16\%$). If we examine static and dynamic activities separately (see Table VI), it strikes that the recognition rate improves especially for dynamic activities ($+11\%$) where the performance concerning static activities increases slightly ($+3\%$). This means that the dynamic activities are much better characterized by acceleration data and that even the gravity-based features that we took into account for static activities did not resolve this issue. The corresponding confusion matrix (see Table VII) confirms this statement. Hence, the static activities lying (A_4), standing (A_5), and sitting (A_6) are often confused

TABLE VI
DISTINCTION BETWEEN STATIC AND DYNAMIC ACTIVITIES CONCERNING ONLINE AND OFFLINE TRAINING.

	Static			Dynamic		
Method	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
Our approach (offline)	0.78	0.77	0.78	0.79	0.78	0.78
Our approach (online)	0.77	0.76	0.76	0.76	0.75	0.76
+ Smoothing	0.79	0.79	0.79	0.88	0.85	0.86
+ User-Feedback	0.80	0.79	0.79	0.86	0.86	0.86
+ Smoothing & U-F	0.80	0.79	0.79	0.88	0.86	0.87

978-1-5090-4327-9/17/$31.00 ©2017 IEEE

TABLE VII
CONFUSION MATRIX AFTER THE PERSONALIZATION OF THE BASE MODEL (OUR APPROACH) WITH ONLINE AND ACTIVE LEARNING. THE PRESENTED VALUES ARE DIVIDED BY 100 AND ROUNDED.

	Predicted							
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	1878	378	0	0	0	0	16	188
A_2	189	2481	0	0	0	0	21	230
A_3	2	1	378	0	0	0	57	0
A_4	0	0	0	1768	66	462	0	0
A_5	0	0	0	22	2546	544	0	0
A_6	0	0	0	175	719	2259	0	0
A_7	42	30	23	0	0	0	3660	8
A_8	101	354	0	0	0	0	12	2881

TABLE VIII
AFTER PERSONALIZATION OF THE BASE MODEL (OUR APPROACH): RECOGNITION RATES OF INTERESTING ACCELEROMETER/POSITION COMBINATIONS.

	Online & Active Learning					
	P_2-P_5 (Watch & Phone)			P_3-P_5 (Glasses & Phone)		
Class	Precision	Recall	F_1	Precision	Recall	F_1
A_1	0.80	0.72	0.76	0.79	0.77	0.78
A_2	0.77	0.81	0.79	0.82	0.84	0.83
A_3	0.98	0.87	0.92	0.97	0.83	0.89
A_4	0.83	0.61	0.70	0.90	0.79	0.84
A_5	0.82	0.82	0.82	0.77	0.89	0.83
A_6	0.59	0.75	0.66	0.73	0.72	0.73
A_7	0.98	0.98	0.98	0.97	0.97	0.97
A_8	0.83	0.86	0.85	0.87	0.88	0.87
$avg.$	0.81	0.80	0.80	0.84	0.84	0.84

among each other. Even *user-feedback* only improves the recognition of these activities slightly. In contrast, the dynamic activities also cover activities that are confused (climbing down (A_1), climbing up (A_2), and walking (A_8)) but the *user-feedback* mostly resolves this problem.

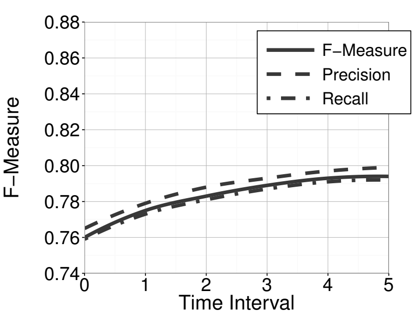
Evaluating these two techniques separately and together showed that they improve different parts of the activity recognition model thus complementing each other (see Table V and VI). Focusing only on *smoothing*, the performance of the base model improves by ∼1-2% where mostly the recognition rate of dynamic activities increased. This indicates that this kind of minor errors occur less frequency. Indeed, the more acceleration data was processed by our updatable model, the less frequently such errors occurred.

Focusing on the same specific device position combinations as in the previous section (see Table IV and VIII), it points out that also for these combinations the recognition rate improved significantly (watch & phone (+11%), glasses & phone (+12%)). Considering the individual activities, especially walking (A_8) achieves a satisfying recognition rate (85% and 86%). As in the preceding results, jumping (A_3) and running (A_7) have the highest and sitting (A_6) the lowest recognition rates.

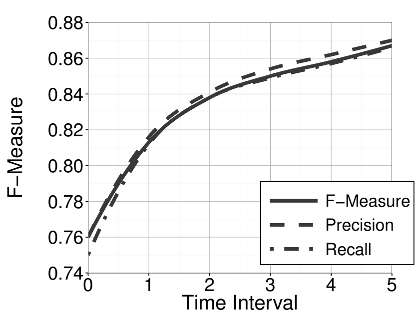
The personalization of a cross-subjects model is a continuous process. Figure 4 shows how the performance evolves over time and clarifies that especially the recognition rate of dynamic activities improves significantly (87%). Each time interval covers acceleration data for each activity and also the same amount of data across the intervals that are classified by our model. For both activity types, we can observe that the recognition rate increased mostly during the first two time intervals. This indicates that the number of windows with a low confidence classification decreases with each iteration. The recognition rate of static activities seems to converge which is an indicator that the acceleration data is not sufficient. Nevertheless, the recognition rate of the base model improves after the first iteration by +4% and after five iterations by +8% (84%).

We also evaluated different thresholds for the confidence value of the classified windows. Figure 5 shows the ratio between additional obtained recognition rate (first derivative, slope) and the number of questions that has to be answered by the target person. It depicts that a higher confidence value results in a larger number of classified windows that are considered as uncertain so the number of questions increases. Of course, the number of questions depends on the number of considered activities, more precisely, the number of activity instances that are covered by the considered data set. During our experiments, we assumed that all considered activities occurred exactly once during a *time interval*. For our presented results, we considered a threshold of 0.5 to keep the number of questions small but cover the turning point of the slope. Hence, in average each user had to answer ∼10 questions to improve the base recognition model by +8%. Besides, if the threshold is high, the slope function converges to zero, i.e., windows with a high confidence value are correct classified.

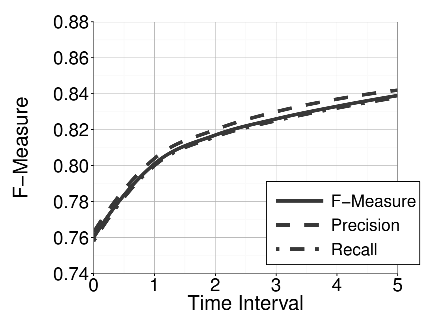
Finally, we examined the relation between the activity recognition rate and the number of trees of an online random



(a) Static activities



(b) Dynamic activities



(c) All activities

Fig. 4. Static vs. dynamic activity recognition: Improvement due to active learning of the base recognition model (our approach) over time.

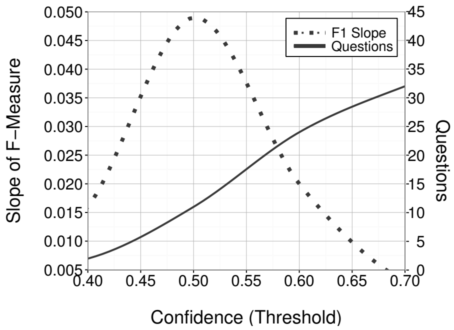


Fig. 5. Progression of the activity recognition rate dependent on the confidence threshold concerning uncertain windows.

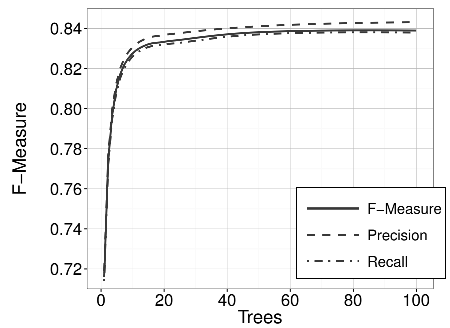


Fig. 6. Influence of the size of the random forest concerning the activity recognition rate.

forest classifier (see Figure 6). It points out that already a forest with 10 trees performs comparable to a forest with 100 trees. Indeed, their recognition rate differs only by $\sim$1-2% where precision and recall are close to each other. The advantages which result from a small forest are less computational power, lower memory usage, and a shorter computation time. This result shows the feasibility of online learning on wearable devices.

All of these results are a strong evidence for the feasibility that cross-subjects based models can be personalized by online and active machine learning. The personalized models achieve recognition rates of 84% and for dynamic activities even 87%. Concerning static activities, gravity-based features enable to decrease the confusion between standing and lying where sitting is still often confused with these two activities. Further, instead of collection a labeled data set, the personalization of an existing base model is significantly less effort for the target user and also feasible for elderly and patients. Besides, the achieved recognition rates are comparable to subject-specific approaches of previous works [3], [4].

VII. Conclusion and Future Work

In this paper, we investigated the feasibility to personalize cross-subjects activity recognition models using an online random forest as a classifier to improve the model over time using user-feedback (active learning) and smoothing. In this context, we examined different kinds of cross-subjects based model, i.e., leave-one-subject-out, randomly selected subjects, and relying on subjects with similar physical characteristics. Further, we considered all common on-body device positions and combinations and focused on common activities. For the experiments, we considered a large real world data set that covers 15 subjects, 8 different activities, and 7 on-body device positions. Besides, the online random forest classifier was self-implemented since there was no Java implementation available which is preferable for most wearable devices.

The results show that our group-based recognition model performs the best (78%). Subsequently, the personalization experiments were conducted showing that the recognition rate for a new subject can be improved to 84% while dynamic activities which are normally of higher interest could be recognized with 87% (F-measure). These results show on the one hand that physical characteristic (fitness level, body structure, and gender) enable to build promising cross-subjects activity recognition models as a basis for personalization while on the other hand online and active learning are a suitable way for increasing significantly the recognition rate of such a model. The resulting effort for the target user that goes along with the personalization is limited to 10 questions, i.e., significantly less effort than collecting and labeling a new data set. Thus, the benefits for the user are evident and make an application in a real world situation more feasible.

In previous works, Weiss et al. [6] and Lara et al. [2] already hypothesized that common physical characteristics could be reliable indicators to build cross-subjects models. In our work, we analyzed and evaluated this hypothesis and provide evidences for its correctness. With respect of personalization, we can state that our approach achieves a higher improvement than a combination of neural networks and fuzzy clustering [20] or online parameter optimization [17], [18]. Further, related work also suggests that an extension of our approach by co-training could be a promising idea [14].

So far, we have shown that activity recognition based on wearable devices can be reliably executed in a real world setting and the necessary training effort can be reduced significantly using online and active learning. However, in our work, we have only considered rather basic activities like walking and running. Many interesting scenarios, however, require the recognition of activities on a higher level of abstraction. For a healthcare setting for example it would be highly beneficial to be able to recognize high level activities such as working, performing sports, or eating. In future work, we will investigate how our work so far can be extended towards the recognition of such higher level activities. We expect that this will require the use of background knowledge about the nature of and relations between high level activities. In previous works, we have already proposed purely unsupervised methods for recognizing high level activities [36], [37], these approaches however, were only tested in a highly restricted setting. Trying to recognize high level activities in more realistic open world settings will come with significant challenges both with respect to acquiring background knowledge and developing robust recognition methods.

REFERENCES

[1] L. Bao and S. S. Intille, *Activity Recognition from User-Annotated Acceleration Data*. Berlin, Heidelberg: Springer, 2004, pp. 1–17.

[2] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[3] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," in *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE Computer Society, 2016, pp. 1–9.

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

[5] Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu, "Cross-people mobile-phone based activity recognition," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Three*. AAAI Press, 2011, pp. 2545–2550.

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

[7] D. Coskun, O. D. Incel, and A. Ozgovde, "Phone position/placement detection using accelerometer: Impact on activity recognition," in *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*, 2015, pp. 1–6.

[8] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.

[9] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman, "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor," in *2007 11th IEEE International Symposium on Wearable Computers*. IEEE Computer Society, 2007, pp. 37–40.

[10] K. Altun and B. Barshan, *Human Activity Recognition Using Inertial/Magnetic Sensor Units*. Berlin, Heidelberg: Springer, 2010, pp. 38–51.

[11] T. Maekawa and S. Watanabe, "Unsupervised activity recognition with user's physical characteristics data," in *2011 15th Annual International Symposium on Wearable Computers*. IEEE Computer Society, 2011, pp. 89–96.

[12] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*, ser. IAAI'05. AAAI Press, 2005, pp. 1541–1546.

[13] W.-Y. Deng, Q.-H. Zheng, and Z.-M. Wang, "Cross-person activity recognition using reduced kernel extreme learning machine," *Neural Networks*, vol. 53, pp. 1–7, 2014.

[14] M. Stikic, K. V. Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," in *2008 12th IEEE International Symposium on Wearable Computers*. IEEE Computer Society, 2008, pp. 81–88.

[15] R. Liu, T. Chen, and L. Huang, "Research on human activity recognition based on active learning," in *2010 International Conference on Machine Learning and Cybernetics*. IEEE Computer Society, 2010, pp. 285–290.

[16] H. Alemdar, T. L. M. van Kasteren, and C. Ersoy, *Using Active Learning to Allow Activity Recognition on a Large Scale*. Berlin, Heidelberg: Springer, 2011, pp. 105–114.

[17] J. Prkk, L. Cluitmans, and M. Ermes, "Personalization algorithm for real-time activity recognition using PDA, wireless motion bands, and binary decision tree," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 5, pp. 1211–1215, 2010.

[18] A. Reiss and D. Stricker, "Personalized mobile physical activity recognition," in *Proceedings of the 2013 International Symposium on Wearable Computers*. ACM, 2013, pp. 25–28.

[19] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, "Streamar: Incremental and active learning with evolving sensory data for activity recognition," in *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*. IEEE Computer Society, 2012, pp. 1163–1170.

[20] Z. Wang, M. Jiang, Y. Hu, and H. Li, "An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 691–699, 2012.

[21] Z. S. Abdallah, M. Gaber, B. Srinivasan, and S. Krishnaswamy, "Adaptive mobile activity recognition system with evolving data streams," *Neurocomputing*, vol. 150, Part A, pp. 304–317, 2015.

[22] Y. A. Jeroudi, M. A. Ali, M. Latief, and R. Akmeliawati, "Online sequential extreme learning machine algorithm based human activity recognition using inertial data," in *Control Conference (ASCC), 2015 10th Asian*. IEEE Computer Society, 2015, pp. 1–6.

[23] Q. V. Vo, M. T. Hoang, and D. Choi, "Personalization in mobile activity recognition system using-medoids clustering algorithm," *International Journal of Distributed Sensor Networks*, vol. 2013, p. 12, 2013.

[24] J. Andreu, R. D. Baruah, and P. Angelov, "Real time recognition of human activities from wearable sensors by evolving classifiers," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE Computer Society, 2011, pp. 2786–2793.

[25] J. Andreu and P. Angelov, "An evolving machine learning method for human activity recognition systems," *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 2, pp. 195–206, 2013.

[26] D. Guan, W. Yuan, Y. K. Lee, A. Gavrilov, and S. Lee, "Activity recognition based on semi-supervised learning," in *13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2007)*. IEEE Computer Society, 2007, pp. 469–475.

[27] D. Roggen, K. Forster, A. Calatroni, T. Holleczek, Y. Fang, G. Troster, A. Ferscha, C. Holzmann, A. Riener, P. Lukowicz, G. Pirkl, D. Bannach, K. Kunze, R. Chavarriaga, and J. d. R. Millan, "Opportunity: Towards opportunistic activity and context recognition systems," in *2009 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops (WoWMoM)*. IEEE Computer Society, 2009, pp. 1–6.

[28] D. Riboni and C. Bettini, "Cosar: hybrid reasoning for context-aware activity recognition," *Personal and Ubiquitous Computing*, vol. 15, no. 3, pp. 271–289, 2011.

[29] N. C. Oza, "Online bagging and boosting," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, 2005, pp. 2340–2345.

[30] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[31] ——, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[33] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line random forests," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE Computer Society, 2009, pp. 1393–1400.

[34] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, p. 6474, 2014.

[35] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2009.

[36] R. Helaoui, M. Niepert, and H. Stuckenschmidt, "Recognizing interleaved and concurrent activities: A statistical-relational approach," in *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, 2011, pp. 1–9.

[37] D. Riboni, T. Sztyler, G. Civitarese, and H. Stuckenschmidt, "Unsupervised recognition of interleaved activities of daily living through ontological and probabilistic reasoning," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 1–12.

978-1-5090-4327-9/17/$31.00 ©2017 IEEE