Towards Robust Activity Recognition for Everyday Life: Methods and Evaluation

Attila Reiss, Didier Stricker
Department of Augmented Vision
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
firstname.lastname@dfki.de

Gustaf Hendeby
Department of Electrical Engineering
Linköping University
Linköping, Sweden
hendeby@isy.liu.se

***Abstract*—The monitoring of physical activities under realistic, everyday life conditions — thus while an individual follows his regular daily routine — is usually neglected or even completely ignored. Therefore, this paper investigates the development and evaluation of robust methods for everyday life scenarios, with focus on the task of aerobic activity recognition. Two important aspects of robustness are investigated: dealing with various (unknown) other activities and subject independency. Methods to handle these issues are proposed and compared, a thorough evaluation simulates usual everyday scenarios of the usage of activity recognition applications. Moreover, a new evaluation technique is introduced (leave-one-other-activity-out) to simulate when an activity recognition system is used while performing a previously unknown activity. Through applying the proposed methods it is possible to design a robust physical activity recognition system with the desired generalization characteristic.**

I. Introduction

Many health benefits have been associated with regular physical activity in the past decades, from maintaining or even enhancing physical fitness to reducing the risk of different diseases. The authors in [2] argue that being physically active is — apart from not smoking — the most powerful lifestyle choice individuals can make to improve their health. There exist recommendations — *e.g.* from the American College of Sports Medicine and the American Heart Association, *cf.* [9] — of how much and what type of (aerobic, muscle-strengthening, etc.) physical activity individuals should do. Monitoring these performed activities is important to ensure the right quality and quantity. However, the non-obtrusive and accurate monitoring of physical activity, while an individual follows his regular daily routine, is a difficult task. Especially the performance of aerobic activities can be highly integrated into the daily routine. Therefore, it is essential to investigate the monitoring of aerobic activities in the realistic conditions of everyday life. This paper focuses on the task of aerobic activity recognition, with the goal to develop and evaluate robust methods for everyday life scenarios. The robustness of the methods is investigated on two different levels: dealing with various (unknown) other activities and subject independency, both explained in more detail in the next subsections.

A. Problem statement: other activities

The recognition of basic aerobic activities (such as walk, run or cycle) and basic postures (lie, sit, stand) is well researched, and is possible with just one 3D-accelerometer [5], [12]. However, since these approaches only consider a limited set of similar activities, they only apply to specific scenarios. Therefore, current research in the area of physical activity recognition focuses amongst others on increasing the number of activities to recognize. For example 11 different activities are recognized in [14], 16 different activities of daily living (ADL) in [10], 19 different activities (with focus on locomotion and sport activities) in [3], and 20 different everyday activities are distinguished in [4], etc. However, there are countless number of activities (*e.g.* 605 different activities are listed in [1]), thus it is not feasible to recognize all of them — not only due to the highly increased complexity of the classification problem, but also due to the fact that collecting data from those hundreds of different activities is practically not possible.

In practice, activity monitoring systems usually focus on only a few activities of interest. Therefore, the main goal is to recognize only these few activities, but as part of a classification problem where all other activities are included as well. Thus the other activities do not need to be recognized, but should not be completely ignored either. One possible way to handle non-interesting other activities is to add a null-class rejection stage at the end of the activity recognition chain, thus discard instances of classified activities based on the confidence of the classification result [19]. Another possibility is to handle them as sub-activities clustered into the main, basic activity classes, *e.g.* ascend/descend stairs considered as walk [12]. The drawback of this solution is that there still remain many activities which can not be put into any of the basic activity classes (*e.g.* vacuum clean or rope jump). The concept of a null-class (or so called background activity class) has been successfully used in the field of activity spotting *e.g.* in [13], and applied in [18] for aerobic activity recognition: apart from the few activities to be recognized, all other activities are part of this null activity class in the defined problem. It was shown that the inclusion of the other activities increases the applicability of the system, but also significantly increases the complexity of the classification problem.

The above mentioned previous works represent a first important step towards dealing with various other activities. However, they only handle a given set of other activities (the entire set of other activities is known when developing the system), thus neglect to simulate the — in practice important — scenario when the user of the system performs an activity previously unknown to the system. Therefore, it remains an open question what happens with all the activities not

considered during the monitoring system's development. To give a concrete example, assume that an activity monitoring system has the goal to recognize 5 basic physical activities (walk, run, etc). When developing this system, in addition to the activities to recognize, 10 other activities are considered as well (vacuum clean, play soccer, etc). The system is specified so that if a user performs any of these other activities, it is not recognized as a basic activity but as an other activity or is rejected. Furthermore, assume that the activity 'rope jump' is neither included in the basic, nor in the set of other activities. Therefore, it is undefined how the system handles the situation when a user performs this 'rope jump' activity. By not dealing with this issue, existing work leaves basically two possibilities: either the user is limited to a scenario where only the considered activities occur (even if $20-30$ different activities are included in the development of a system, this still is a significant limitation for the user), or the user is permitted to perform any kind of physical activity, but it is not specified how the monitoring system handles an activity not considered during the system's development phase (*e.g.* whether it is recognized as one of the basic activities). Either way, by neglecting this issue, the applicability of an activity monitoring application is significantly limited.

B. Problem statement: subject independency

Another important aspect of robustness is the subject independency of an activity monitoring system. In [14], a comparison of subject dependent and independent validation is shown, and a large difference of classifier performance is reported between the two validation techniques ($1.26-5.92\%$ misclassification *vs.* $12.09-29.47\%$ misclassification for different classifiers, respectively). Moreover, [18] also compared subject dependent and independent evaluation, and argues that — unless the development of personalized approaches is the explicit goal — subject independent validation techniques should be preferred. This best simulates the common scenario that such systems are usually trained on a large number of subjects and then used by a new subject (similar to the concept of unknown other activities as discussed above, here the user of the system is unknown during the development phase), while subject dependent evaluation leads to highly "optimistic" performance results. However, many research works even recently still use subject dependent validation techniques (*e.g.* in [11], [20]), neglecting that although they present high performance using their approach, these results might not have as much practical meaning as if subject independent validation would have been applied.

C. Overview and main contributions

This paper introduces and analyzes different approaches to overcome the currently existing limitations in respect of the two presented issues. A large number of experiments are carried out, for which Section II defines the necessary tools and Section III presents and discusses the results. The main contributions of the paper are the following. a) It introduces and strengthens a usually neglected point of view in physical activity recognition: the robustness in everyday life scenarios. Two important aspects of this issue are investigated: dealing with (unknown) other activities and subject independency. Methods to handle these issues are proposed and compared, a thorough evaluation simulates everyday scenarios of the usage of activity monitoring systems. b) A new evaluation technique is introduced: leave-one-other-activity-out (LOOAO). This simulates one of the most commonly neglected scenarios: when an activity recognition system is used while performing a previously unknown other activity. c) Through comparing various different methods, the experiments show that the approach of using an other activity class (referred to as 'bgClass' throughout this paper) has the best generalization characteristic for designing a robust activity recognition system.

II. Methods

This section defines the classification problem used within this work, presents data processing methods and classification algorithms applied, and defines the performance measures to quantify results. Moreover, 4 different models are proposed to deal with other activities. Finally, the evaluation methods used in Section III are presented, including a new evaluation algorithm to simulate everyday life scenarios for using activity recognition applications.

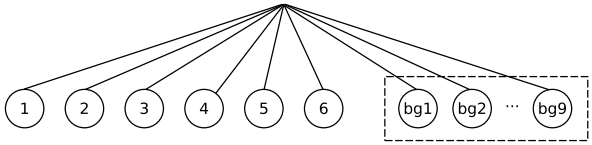
A. Defining the classification problem

The experiments performed within this work are all based on the PAMAP2 dataset, a physical activity monitoring dataset created and released recently [16], [17], and included in the UCI machine learning repository [7]. This dataset is used since it not only includes the basic physical activities (walk, run, cycle, Nordic walk) and postures (lie, sit, stand), but also a wide range of everyday (ascend and descend stairs, watch TV, computer work, drive car), household (iron clothes, vacuum clean, fold laundry, clean house) and fitness activities (rope jump, play soccer). The dataset was recorded from overall 18 physical activities performed by 9 subjects, wearing 3 inertial measurement units (IMU) and a heart rate monitor. The IMUs were placed at 3 different positions on the test subjects' bodies: on the chest, over the wrist on the dominant arm and on the dominant side's ankle. The subjects were aged $27.22\ \pm 3.31$ years and were having a BMI of $25.11\ \pm 2.62$ kgm^{-2}. A more detailed description of the dataset can be found in [16].

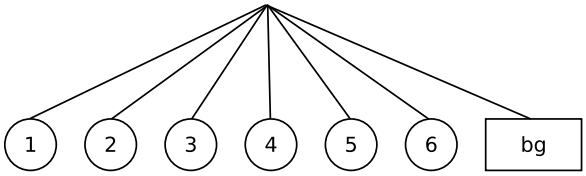
The overall goal of this paper is to develop a physical activity monitoring system which can recognize a few, basic activities and postures of interest, but is also robust in everyday situations. In their daily routine, users of activity monitoring systems perform a large amount of different activities, many of them are not of interest from the activity recognition point of view. Therefore, to simulate this common usage of activity monitoring systems, the activity recognition task is defined as follows.[1] There are 6 different basic activity classes to recognize: lie, sit/stand[2], walk, run, cycle and Nordic walk. In addition, 9 different activities are regarded as other/background activities: iron clothes, vacuum clean, ascend stairs, descend stairs, rope jump, fold laundry, clean house, play soccer and drive car. These other activities should not be recognized as one of the basic activities, but as part of an other activity class

[1]The defined classification task uses 16 different activities from the PAMAP2 dataset. The remaining 2 activities (computer work and watch TV) are discarded here due to their high resemblance to the basic postures.

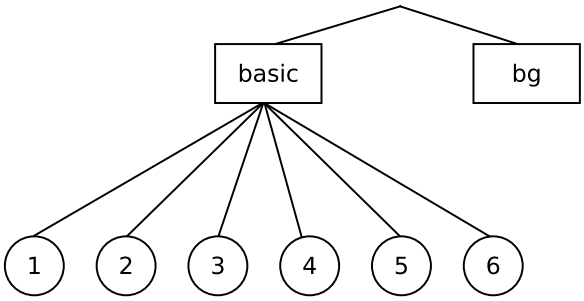
[2]It is a common restriction made in activity recognition (*e.g.* in [6]) that the postures sit and stand form one activity class, since an extra IMU on the thigh would be needed for a reliable differentiation of them.



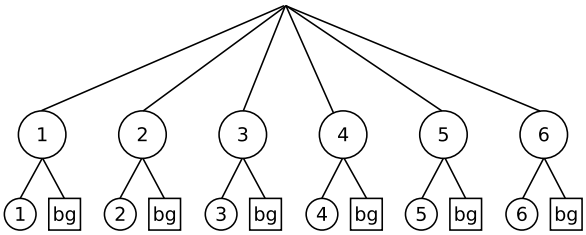
(a) The ‘allSeparate’ model.



(b) The ‘bgClass’ model.



(c) The ‘preReject’ model.



(d) The ‘postReject’ model.

Fig. 1. The 4 proposed models for dealing with the other activities.

or should be rejected. The additional activities will be used to simulate the scenario when users perform other activities than the few basic ones, and are also used to simulate the scenario when users perform an activity unknown to the system. This defined classification problem will be referred to as ‘extended’ activity recognition task throughout this paper. Moreover, for comparison reasons, the classification problem only including the 6 basic activity classes will also be used, and will be referred to as ‘basic’ activity recognition task.

The defined activity recognition task focuses on the monitoring of traditionally recommended aerobic activities (walk, run, cycle and Nordic walk), and can thus be justified by various physical activity recommendations — as given *e.g.* in [9]. Especially patients with diabetes, obesity or cardiovascular disease are often required to follow a well defined exercise routine as part of their treatment. Therefore, the recognition of these basic physical activities is essential to monitor the progress of the patients and give feedback to their caregiver. Moreover, a summary of resting activities (lie, sit and stand still) gives also a feedback on how much sedentary activity the patients “performed”. However, in other use cases the focus of an activity monitoring application could be different, thus the definition of the classification problem (the definition of the basic and other activity classes) would differ. Nevertheless, the methods presented in this work could be applied on those other classification tasks as well.

B. Data processing and classification

The PAMAP2 dataset provides raw sensory data from the 3 IMUs and the heart rate monitor. A data processing chain is applied on the raw data including preprocessing, segmentation and feature extraction steps (these data processing steps are further described in [16]). In total, 137 features are extracted: 133 features from IMU acceleration data (such as mean, standard deviation, energy, entropy, correlation, etc.) and 4 features from heart rate data (mean and gradient). These extracted features serve as input for the classification step, together with the activity class labels provided by the dataset. Previous work in activity recognition showed that decision tree based classifiers, especially boosted decision trees, usually achieve high performance [16]. Moreover, decision tree based classifiers have the benefit to be fast classification algorithms with a simple structure, and are thus also easy to implement. These benefits are especially important for activity recognition applications since they are usually running on mobile, portable systems for everyday usage, thus the available computational power is limited [18]. Therefore, the C4.5 decision tree classifier [15] and the AdaBoost.M1 (using C4.5 decision tree as weak learner) algorithm [8] are used and compared in the experiments on the defined classification problem.

C. Modeling other activities

As discussed above, the focus is on the recognition of the basic activity classes, but all the other activities should not be completely neglected either. Therefore, 4 different models are proposed for dealing with these other activities. The main goal of these solutions is the high recognition rate of the basic activities, but they should also show robust behaviour concerning unknown activities, thus should have good generalization characteristic. The 4 proposed methods are listed below, and are visualized — by means of the concrete example of the defined activity recognition task: 6 basic activity classes and 9 other activities — in Fig. 1.

The ‘allSeparate’ model: for each of the other (or also called background) activities a separate activity class is defined (‘bg1’ . . . ‘bg9’), and all these classes are regarded as activities not belonging to the 6 basic activity classes (‘1’ . . . ‘6’). The concept of the method is visualized in Fig. 1(a). This model

TABLE I. CONFUSION MATRIX USED FOR THE ADJUSTED DEFINITION OF THE PERFORMANCE MEASURES.

Annotated activity	**Recognized activity** 1	2	...	C	0	
1	$P_{1,1}$	$P_{1,2}$	...	$P_{1,C}$	$P_{1,C+1}$	S_1
2	$P_{2,1}$	$P_{2,2}$	...	$P_{2,C}$	$P_{2,C+1}$	S_2
...						
C	$P_{C,1}$	$P_{C,2}$	...	$P_{C,C}$	$P_{C,C+1}$	S_C
$C+1$	$P_{C+1,1}$	$P_{C+1,2}$	...	$P_{C+1,C}$	$P_{C+1,C+1}$	S_{C+1}
...						
$C+B$	$P_{C+B,1}$	$P_{C+B,2}$	...	$P_{C+B,C}$	$P_{C+B,C+1}$	S_{C+B}
	R_1	R_2	...	R_C	R_{C+1}	

refers to the nowadays common approach of dealing with a large number of activities: most research work is focused on increasing the number of recognized activities, thus to have a high number of separate activity classes.

The 'bgClass' model: in addition to the basic activity classes a background activity class ('bg' in Fig. 1(b)) is defined, containing all the other activities. This approach of a null-class for physical activity recognition was proposed in [18] to increase the applicability in everyday life scenarios.

The 'preReject' model: it basically inserts a null class rejection step before the actual classification. The concept of this two-level model is visualized in Fig. 1(c). On the first level the basic activities are separated from all the other activities ('bg' class). The second level — only on the 'basic' branch of the first level — distinguishes the 6 different basic activity classes. When constructing a classifier based on this model, all training samples are used to create the sub-classifier of the first level, while for the second level only training samples from the basic activity classes are used.

The 'postReject' model: similar to the 'preReject' model, this is also a two-level model, as shown in Fig. 1(d). However, the null class rejection step is applied here after classifying the basic activities. This solution is similar to *e.g.* the decision filtering step applied after activity classification in the activity recognition chain of [19]. Only samples of the basic activity classes are used to create the first level of this classifier, while the second level consists of 6 sub-classifiers: each created using the respective basic activity class and all samples from the other activities.

D. Performance measures

The common measures are used to describe the classification performance of the different approaches: precision, recall, F-measure and accuracy.[3] However, since the focus of this work is on the recognition of the basic activity classes, these performance measures are adjusted so that they also focus on the basic activities. The adjusted definition of the 4 measures uses the following notation (*cf.* also the confusion matrix in Table I). Assume that a confusion matrix is given by its entries $P_{i,j}$, where i refers to the rows (annotated activities), and j to the columns (recognized activities) of the matrix. Let S_i be the sum of all entries in the row i of the matrix (referring to the number of samples annotated as activity i), and R_j the sum of all entries in the column j of the matrix (referring to the number of samples recognized as activity j). Let N be the total number of samples in the confusion matrix. Let the classification problem represented in the confusion matrix have C basic activity classes: $1, \ldots, C$ and B other activity classes: $1, \ldots, B$. Let the activity classes ordered so in the confusion matrix that the background activity classes follow the basic activity classes (*cf.* the order of the annotated activity classes in Table I). Since the classification of the samples belonging to the other activities is not of interest, this is represented as a null activity class in the confusion matrix (*cf.* the column referred to as $P_{i,C+1}$ in Table I). Samples classified as one of the background activity classes ('allSeparate' model), or classified into the other activity class ('bgClass' model), or rejected before or after the classification of the basic activities ('preReject' or 'postReject' model, respectively) are counted into this null class. Using this notation, the performance measures used in this paper are defined as following (the correct classification of only the basic activities is of interest in the definition of the metrics):

$$precision = \frac{1}{C} \sum_{i=1}^{C} \frac{P_{i,i}}{R_i} \tag{1}$$

$$recall = \frac{1}{C} \sum_{i=1}^{C} \frac{P_{i,i}}{S_i} \tag{2}$$

$$F\text{-}measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

$$accuracy = \frac{1}{N - \sum_{j=C+1}^{C+B} P_{j,C+1}} \sum_{i=1}^{C} P_{i,i} \tag{4}$$

Concrete confusion matrices on the defined 'basic' and 'extended' classification problems are shown as results in Section III. Moreover, those confusion matrices are used to understand and compare the results in more detail, thus *e.g.* which activity classes are confused more frequently by different approaches.

E. Evaluation techniques

The goal of the evaluation of the created classifiers is to estimate their behaviour in everyday life scenarios, thus to simulate how they would perform in named situations. The commonly used standard *k*-fold cross-validation (CV) is not adequate for this task, since it only estimates the behaviour on the scenario in which the classifier was trained, thus on a limited and known set of users and physical activities. Nevertheless, standard 10-fold CV is also applied as an evaluation technique in the experiments of this work for comparison reasons. These results will show how "optimistic" *k*-fold CV is for validation, how unrealistic the so achieved performance is in real life scenarios.

The simulation of everyday life scenarios means concretely to simulate how the created system behaves when used by a previously (in training time) unknown person, or when a previously unknown activity is performed. To simulate subject independency the evaluation technique leave-one-subject-out (LOSO) CV is applied. Since the used PAMAP2 dataset provides data from 9 subjects, LOSO 9-fold CV is applied in this paper. This means that data from 8 subjects is used for training

[3]Recently new error metrics were introduced for continuous activity recognition, *e.g.* insertion, merge, overfill, etc. [21], [22]. However, contrary to activity recognition in *e.g.* home or industrial settings, for physical activity monitoring the frame by frame metrics are sufficient, as discussed in [16].

Algorithm 1 LOSO_LOOAO

Require: **S** is the set of S different subjects, $s : 1, \ldots, S$
C is the set of C different basic activities, $c : 1, \ldots, C$
B is the set of B different other activities, $b : 1, \ldots, B$
A is the set of all different activities: **A** = **C** $\cup$ **B**, an arbitrary activity is referred to as a
N is the set of N different samples, where each sample consists of subject and activity information and a feature vector, thus $\underline{n}$: $< s, a, features >$
$s(\underline{n})$ refers to the subject of the sample $\underline{n}$
$a(\underline{n})$ refers to the activity of the sample $\underline{n}$

procedure LOSO_LOOAO(**S,C,B,A,N**)
  for $i \leftarrow 1, S$ **do**
    $\mathbf{P}_{train} = \{\forall \underline{n} \in \mathbf{N} \mid s(\underline{n}) \neq i\}$
    $\mathbf{P}_{test} = \{\forall \underline{n} \in \mathbf{N} \mid s(\underline{n}) = i\}$
    $\mathbf{P}_{test\_basic} = \{\forall \underline{n} \in \mathbf{P}_{test} | a(\underline{n}) \in \mathbf{C}\}$
    Train classifier using $\mathbf{P}_{train} \rightarrow F_i$
    Use F_i on $\mathbf{P}_{test\_basic}$ % *LOSO on basic activities*
    for $j \leftarrow 1, B$ **do**
      $\mathbf{P}_{train\_other} = \{\forall \underline{n} \in \mathbf{P}_{train} | ((a(\underline{n}) \in \mathbf{C}\ )\ or\ ((a(\underline{n}) \in \mathbf{B}\ and\ (a(\underline{n}) \neq j)))\}$ % *thus the sample does not belong to the jth other activity*
      $\mathbf{P}_{test\_other} = \{\forall \underline{n} \in \mathbf{P}_{test} | ((a(\underline{n}) \in \mathbf{B}\ )\ and\ (a(\underline{n}) = j))\}$ % *thus the sample belongs to the jth other activity*
      Train classifier using $\mathbf{P}_{train\_other} \rightarrow F_{i,j}$
      Use $F_{i,j}$ on $\mathbf{P}_{test\_other}$ % *LOOAO on jth other activity*
    end for % *LOOAO with all B other activities is finished here*
  end for % *The LOSO results with the basic activities and the LOOAO results with the other activities together return the LOSO_LOOAO result*
end procedure

and data from the remaining subject for testing, repeating this procedure 9 times leaving always another subject's data for testing. Moreover, to simulate the scenario of performing unknown other activities a new evaluation technique is introduced: leave-one-other-activity-out (LOOAO). It has a similar concept to the LOSO technique: if the classification problem includes B other activities, data from $B - 1$ other activities is used for training and data from the remaining other activity for testing, repeating this procedure B times leaving always another activity's data for testing.

To receive the best possible estimation of the developed system's behaviour in everyday life scenarios, the newly introduced LOOAO evaluation technique is combined with the LOSO technique. This combined evaluation method will be referred to as LOSO_LOOAO throughout the paper, the procedure is formally described in Algorithm 1. With LOSO_LOOAO evaluation the following practical scenarios are evaluated:

- The system is trained with a large amount of subjects for the 'extended' task. Then the system is deployed to a new subject (thus for this subject no data was available during the training phase of the system), and the new subject performs one of the basic activities (estimated through the LOSO component).
- The system is trained with a large amount of subjects for the 'extended' task. Then the system is deployed to a new subject, who performs one of the known other activities (estimated through the LOSO component). This is the first step in testing the robustness of the system in situations when the user performs activities other than the few basic recognized ones.
- The system is trained with a large amount of subjects for the 'extended' task. Then the system is deployed to a new subject, who performs a previously unknown activity — thus an activity neither belonging to the basic activity classes, nor to one of the other activities available during the training phase (estimated through the LOOAO component). This scenario simulates basically the generalization characteristic of the classifier's other activity model, estimating how robust the system is in the usually neglected situation when unknown activities are performed.

III. Results and discussion

Table II shows the confusion matrix on the 'basic' classification task using the C4.5 decision tree classifier and standard CV as evaluation technique (the results are an average of 10 test runs). Almost no misclassifications can be observed, all performance measures are clearly above 99%. Therefore, this result could indicate that physical activity recognition is an easily solvable classification problem, even with a simple classifier such as a decision tree. However, the result of Table II has two main drawbacks: it is subject dependent (thus does not tell anything about the performance of the system when used by a new subject), and only applies to the specific scenario of these 6 basic activity classes. Therefore, an extension of this result is required to increase the applicability of the system concerning both limitations. Further results in this section will show that the performance of activity recognition is much lower under realistic, everyday life conditions.

A. The 'basic' task

The 'basic' classification task serves only for comparison, thus to see the baseline characteristic of physical activity recognition. Since all activities of the task are to be recognized, only the subject independency of the system can be simulated from the aforementioned two issues. The performance measures are shown in Table III for both standard CV and LOSO

TABLE II. CONFUSION MATRIX ON THE 'BASIC' TASK USING THE C4.5 DECISION TREE CLASSIFIER AND STANDARD CV EVALUATION TECHNIQUE. THE TABLE SHOWS HOW DIFFERENT ANNOTATED ACTIVITIES ARE CLASSIFIED IN [%].

Annotated activity	**Recognized activity**					
	1	2	3	4	5	6
1 lie	100	0	0	0	0	0
2 sit/stand	0	99.87	0	0	0.13	0
3 walk	0	0.05	99.66	0	0.02	0.27
4 run	0	0	0	100	0	0
5 cycle	0	0.29	0.25	0.11	99.29	0.06
6 Nordic walk	0	0	0.60	0	0	99.40

evaluation. Each of the tests is performed 10 times, the table shows the mean and standard deviation of these 10 test runs.

The results of Table III show the significant difference between using standard CV or LOSO as evaluation method, for both classifiers. An interesting result is that the AdaBoost.M1 classifier only slightly outperforms the C4.5 classifier on the 'basic' task (the difference between the two classifiers on the 'extended' task is much more significant, as shown in the next subsection). This can be explained by the fact that the 'basic' task is a rather simple classification problem where even base-level classifiers can reach the highest possible accuracy. Therefore, it is not necessarily worth using more complex classification algorithms here. The lower performance when using LOSO evaluation is due to the difficulty of the generalization in respect of the users, and not due to the difficulty of the classification task. Although using subject independent evaluation is the first step towards simulating the conditions of everyday usage of activity recognition applications, the 'basic' task only estimates the system's behaviour when activities of one of the 6 included activity classes are performed, thus the system's response is not defined when the user performs activities such as descend stairs or vacuum clean. This issue is investigated in the next subsection, by analyzing the results obtained on the 'extended' classification task.

B. The 'extended' task

The performance measures on the 'extended' task are presented in Table IV: for each of the 4 other activity models, by using the 2 classifiers and the 3 different evaluation techniques. The results are given in form of mean and standard deviation of the 10 test runs performed for every possible combination of the models, classifiers and evaluation methods. Overall it is clear that with the inclusion of the other activities the classification task becomes significantly more difficult (*cf.* the comparison of the results achieved with standard CV and LOSO to the respective results on the 'basic' task). This can be explained not only by the increased number of activities in the classification problem, but also by the fact that the characteristic of some of the introduced other activities overlap with the characteristic of some of the basic activity classes. For example, the other activity *iron* has a similar characteristic as when talking and gesticulating during *stand*, thus misclassifications appear between these two activities. Similarly it is nontrivial to distinguish running with a ball (during the other activity *play soccer*) from just *running*. Since the 'extended' task defines a complex classification problem, it is worth to apply more complex classification algorithms here — contrary to the 'basic' classification task. For example when considering the 'allSeparate' model and LOSO evaluation, the C4.5 decision tree only achieves an F-measure of 83.30% while with the AdaBoost.M1 classifier 92.22% can be reached.

From the results of Table IV it is trivial that the performance measures achieved with LOSO evaluation are significantly lower than results obtained with standard CV, as already seen in Table III. If only considering subject independency the 'allSeparate' model performs best, closely followed by the models 'preReject' and 'bgClass'. However, on the 'extended' task it is also simulated when the user of the system performs unknown other activities (LOOAO). The results of applying the evaluation method of Algorithm 1 are shown in Table IV in the respective rows of LOSO_LOOAO. Considering this combined evaluation technique the 'bgClass' model performs best, followed by the models 'preReject' and 'allSeparate'. From all the 4 other activity models the 'allSeparate' model shows the largest decrease in performance from LOSO evaluation to LOSO_LOOAO evaluation. Especially the precision measure decreases largely, thus when the user performs unknown activities they are more likely recognized as one of the basic activity classes compared to the results of other models. This can be explained by the fact that for this model separate activity classes are created and trained for each of the known other activities, thus the generalization capability of the model is rather limited when a previously unknown activity is performed. On the other hand, the training instances belonging to the other/background activity class of the 'bgClass' model are scattered in the feature space, resulting in a large class with good generalization characteristic. Moreover, since much more instances are used for the creation of the background activity class during training than for the 6 basic activity classes, this class becomes more important, thus resulting in significantly higher precision than recall result with the 'bgClass' model.

The 'preReject' model performed second best for both LOSO and LOSO_LOOAO evaluation, justifying the idea of first recognizing whether a performed activity belongs to the basic activity classes or not. When analyzing the trained classifiers for the two levels of this model, it can be noticed that the classifier of the first level is much more complex: although representing only a binary decision, the separation of basic activities from other activities is a difficult task. The classification problem defined in the second level of the model is identical to the 'basic' classification task defined in this paper, and thus is — as discussed in the previous subsection — a rather simple task. Finally, the 'postReject' model performed worst with both LOSO and LOSO_LOOAO evaluation, resulting in the lowest F-measure and accuracy values. Since the basic activities are distinguished on the first level of this model (without any other activities concerned), this model has the least confusion between the basic activity classes. The confusion matrices belonging to the evaluation of this model — not shown in this paper due to the limited space — confirm this statement: except of some misclassifications of *Nordic walk* samples into the normal *walk* class, all confusion is done towards the other activity class. Moreover, due to the unbalanced classification tasks defined on the second level of the model (only one basic activity versus all other activities, thus these tasks are even more unbalanced than the classification task defined by the 'bgClass' model), the precision values are comparable with those of other models. Therefore, if the goal of an activity recognition application is only the precise recognition of activities of interest the

TABLE III. PERFORMANCE MEASURES ON THE 'BASIC' ACTIVITY RECOGNITION TASK

Classifier	Evaluation method	Precision	Recall	F-measure	Accuracy
C4.5	standard CV	99.71 ± 0.04	99.70 ± 0.02	99.71 ± 0.03	99.71 ± 0.03
	LOSO	96.05 ± 1.06	94.96 ± 1.40	95.50 ± 1.20	95.14 ± 1.10
AdaBoost.M1	standard CV	99.97 ± 0.02	99.97 ± 0.02	99.97 ± 0.02	99.97 ± 0.02
	LOSO	95.91 ± 1.45	95.47 ± 1.45	95.69 ± 1.40	95.43 ± 1.54

TABLE IV. PERFORMANCE MEASURES ON THE 'EXTENDED' ACTIVITY RECOGNITION TASK

Model	Classifier	Evaluation method	Precision	Recall	F-measure	Accuracy
'allSeparate'	C4.5	standard CV	98.17 ± 0.23	98.00 ± 0.09	98.09 ± 0.14	95.80 ± 0.25
		LOSO	89.77 ± 1.89	77.75 ± 3.08	83.30 ± 2.10	73.81 ± 2.21
		LOSO_LOOAO	81.84 ± 1.77	78.59 ± 3.43	80.16 ± 2.44	67.06 ± 2.71
	AdaBoost.M1	standard CV	99.94 ± 0.01	99.93 ± 0.04	99.93 ± 0.02	99.83 ± 0.05
		LOSO	95.42 ± 0.98	89.23 ± 2.00	92.22 ± 1.40	86.60 ± 2.09
		LOSO_LOOAO	86.80 ± 0.99	88.72 ± 1.28	87.75 ± 1.07	78.83 ± 1.29
'bgClass'	C4.5	standard CV	98.68 ± 0.17	98.66 ± 0.11	98.67 ± 0.12	96.85 ± 0.21
		LOSO	89.85 ± 1.35	85.83 ± 3.11	87.78 ± 2.11	80.63 ± 1.81
		LOSO_LOOAO	83.64 ± 2.46	85.56 ± 2.67	84.58 ± 2.39	73.76 ± 2.10
	AdaBoost.M1	standard CV	99.96 ± 0.02	99.88 ± 0.03	99.92 ± 0.02	99.77 ± 0.05
		LOSO	96.07 ± 0.99	85.76 ± 2.45	90.61 ± 1.72	84.14 ± 2.35
		LOSO_LOOAO	91.81 ± 0.82	86.82 ± 1.71	89.24 ± 1.17	80.97 ± 1.20
'preReject'	C4.5	standard CV	98.28 ± 0.14	97.83 ± 0.12	98.05 ± 0.07	95.46 ± 0.14
		LOSO	88.58 ± 1.40	78.66 ± 2.51	83.30 ± 1.36	71.78 ± 1.76
		LOSO_LOOAO	83.07 ± 1.68	78.83 ± 3.63	80.87 ± 2.53	67.32 ± 2.74
	AdaBoost.M1	standard CV	99.95 ± 0.04	99.89 ± 0.04	99.92 ± 0.04	99.82 ± 0.06
		LOSO	93.85 ± 1.57	88.46 ± 2.26	91.07 ± 1.83	85.20 ± 2.07
		LOSO_LOOAO	87.99 ± 1.47	87.98 ± 1.80	87.98 ± 1.58	79.11 ± 1.60
'postReject'	C4.5	standard CV	99.08 ± 0.09	98.21 ± 0.15	98.64 ± 0.10	96.89 ± 0.20
		LOSO	92.93 ± 0.93	77.65 ± 3.05	84.59 ± 2.11	74.89 ± 1.80
		LOSO_LOOAO	89.02 ± 0.62	78.96 ± 2.05	83.67 ± 1.23	71.59 ± 1.66
	AdaBoost.M1	standard CV	99.93 ± 0.04	99.82 ± 0.02	99.87 ± 0.03	99.75 ± 0.05
		LOSO	95.76 ± 1.38	81.18 ± 2.57	87.86 ± 1.87	80.92 ± 2.50
		LOSO_LOOAO	92.01 ± 1.80	80.65 ± 3.02	85.94 ± 2.40	77.78 ± 2.52

'postReject' model can also be considered, but otherwise one of the three other models should be used.

From the results of Table IV the performance measures obtained with LOSO_LOOAO evaluation should be regarded as most important, since this evaluation technique simulates the widest range of practical scenarios. The approach achieving the best performance results with LOSO_LOOAO can thus be regarded as the approach which is the most robust in everyday life situations. Therefore, overall the 'bgClass' model can be regarded as the model with the best generalization characteristic: the approach using the 'bgClass' model and the AdaBoost.M1 classifier achieves an average F-measure of 89.24% and an average accuracy of 80.97%. The confusion matrix obtained with this approach is shown in Table V (the results represent the average from the 10 test runs). It is obvious that most of the misclassifications occur due to the other activities: either a sample belonging to a basic activity class is classified into the background class, or a sample from an other activity is confused with one of the basic activities. For example, *drive car* and *iron* are in high percentage confused with the basic class *sit/stand*. This is due to the overlapping characteristic of some basic and other activities, as already discussed above. The strength of the 'bgClass' model is especially pointed out by the results obtained with other activities such as *ascend stairs*, *descend stairs*, *vacuum clean* or *rope jump*: although previously unknown to the system, these activities were basically not misclassified as a basic activity. Therefore, it can be expected that the proposed approach shows such robustness with most of other unknown activities as well. Only unknown activities similar to the target activities might be problematic for the 'bgClass' approach, as seen with *drive car* or *iron*, or is expected with activities such as *computer work* or *watch TV*. However, it is difficult to set the defining boundaries of some of the basic activity classes — *e.g.* if *computer work* should be regarded as *sitting* or as a separate other class. Deciding this question might highly depend on the actual application.

IV. CONCLUSION

This paper created the means for simulating everyday life scenarios and thus to evaluate the robustness of activity recognition — a usually neglected point of view in the development of physical activity monitoring systems. Experiments were carried out on a classification problem defined on the recently released PAMAP2 dataset, including 6 basic activity classes and 9 different other activities. The goal of the classification task was the accurate recognition and separation of the basic activities, while samples of the other activities should be recognized as part of an other activity class or should be rejected. Common data processing and classification methods were used to achieve this, comparing two — in previous work successfully applied — classification algorithms: the C4.5 decision tree classifier and the AdaBoost.M1 algorithm. Moreover, to deal with other activities, 4 different models are proposed: 'allSeparate', 'bgClass', 'preReject' and 'postReject'. Finally, the evaluation of the proposed methods was performed with

TABLE V. CONFUSION MATRIX ON THE 'EXTENDED' CLASSIFICATION TASK USING THE 'BGCLASS' MODEL, ADABOOST.M1 CLASSIFIER AND LOSO_LOOAO EVALUATION TECHNIQUE. THE TABLE SHOWS HOW DIFFERENT ANNOTATED ACTIVITIES ARE CLASSIFIED IN [%].

Annotated activity	**Recognized activity**						
	1	2	3	4	5	6	0
1 lie	96.66	2.62	0	0	0	0	0.72
2 sit/stand	0.15	90.05	0	0	0	0	9.80
3 walk	0	0	85.87	0	0	0.13	14.00
4 run	0	0	0.16	76.24	0	0.30	23.31
5 cycle	0	0	0.01	0	92.43	0.03	7.53
6 Nordic walk	0	0	8.71	0	0	79.69	11.60
7 drive car	0	39.10	0	0	0.06	0	60.84
8 asc. stairs	0	0	0.53	0	0	0.01	99.46
9 desc. stairs	0.08	0	1.97	0.02	0.84	0.06	97.04
10 vacuum clean	0	0	0	0	0.42	0	99.58
11 iron	0	20.02	0	0	0.01	0	79.97
12 fold laundry	0	3.70	0.01	0	0	0	96.29
13 clean house	0.26	7.10	0	0	0.06	0	92.58
14 play soccer	0	0	3.34	32.08	0	0.13	64.46
15 rope jump	0	0	0.11	0.11	0	0	99.78

different techniques, including standard CV, LOSO and the newly introduced LOOAO. Standard 10-fold CV was only included for comparison reasons: to underline how unrealistic the so achieved performance is in everyday life scenarios. The LOSO technique serves to simulate subject independency, while LOOAO simulates the scenario of performing unknown other activities. The results of the thorough evaluation process revealed that the 'bgClass' model has the best generalization characteristic, while the generalization capability of the widely used 'allSeparate' approach is rather limited in respect of performing previously unknown activities.

Developing physical activity monitoring systems while also taking *e.g.* subject independency or unknown activities into account has two important benefits compared to when standard CV evaluation is used only. First of all it is estimated how the developed system behaves in various everyday life scenarios, while this behaviour would be otherwise undefined. Moreover, the best performing models and algorithms can be selected when applying LOSO and LOOAO evaluation during the development phase of the system, hence creating the best possible system from the robustness point of view for everyday life. In future work it is planned to apply the proposed models and evaluation techniques also with other classification problems. It should be also investigated how well the developed approaches generalize with user groups (*e.g.* elderly) significantly differing from the subjects (all young, healthy adults) included in the PAMAP2 dataset. Moreover, it is also planned to investigate the effect of increasing the number of known (thus in the training included) other activities, with the goal to increase even more the robustness towards unknown other activities while keeping the high performance regarding the basic activity classes.

ACKNOWLEDGEMENTS

This work has been performed within the project Activity-Plus funded by the "Stiftung Rheinland-Pfalz für Innovation" under contract number 961 - 386261/1028, and the European project AlterEgo under contract number 600610.

REFERENCES

[1] B. E. Ainsworth, W. L. Haskell, M. C. Whitt, M. L. Irwin, a. M. Swartz, S. J. Strath, W. L. O'Brien, D. R. Bassett, K. H. Schmitz, P. O. Emplaincourt, D. R. Jacobs, and a. S. Leon. Compendium of physical activities: an update of activity codes and MET intensities. *Medicine and science in sports and exercise*, 32(9 Suppl):S498–504, 2000.

[2] L. Alford. What men should know about the impact of physical activity on their health. *International journal of clinical practice*, 64(13):1731–1734, 2010.

[3] K. Altun, B. Barshan, and O. Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

[4] L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. In *Proc. PERVASIVE*, Linz/Vienna, Austria, 2004.

[5] M. Ermes, J. Pärkkä, and L. Cluitmans. Advancing from offline to online activity recognition with wearable sensors. In *Proc. 30th Annual International IEEE EMBS Conference*, Vancouver, Canada, 2008.

[6] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):20–26, 2008.

[7] A. Frank and A. Asuncion. UCI machine learning repository, University of California, Irvine, School of Information and Computer Sciences, 2010. URL http://archive.ics.uci.edu/ml

[8] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[9] W. L. Haskell, I.-M. Lee, R. R. Pate, K. E. Powell, S. N. Blair, B. a. Franklin, C. a. Macera, G. W. Heath, P. D. Thompson, and A. Bauman. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Medicine and science in sports and exercise*, 39(8):1423–1434, 2007.

[10] T. Huynh, U. Blanke, and B. Schiele. Scalable Recognition of Daily Activities with Wearable Sensors. In *Proc. LoCA*, 2007.

[11] O. D. Lara, A. J. Pérez, M. a. Labrador, and J. D. Posada. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and Mobile Computing*, 8(5):717–729, 2011.

[12] X. Long, B. Yin, and R. M. Aarts. Single-accelerometer based daily physical activity classification. In *Proc. 31st Annual International IEEE EMBS Conference*, Minneapolis, MN, pp. 6107–6110, 2009.

[13] G. Ogris, T. Stiefmeier, P. Lukowicz, and G. Troster. Using a complex multi-modal on-body sensor system for activity spotting. In *Proc. ISWC*, Washington, DC, USA, 2012.

[14] S. Patel, C. Mancinelli, P. Bonato, J. Healey, and M. Moy. Using Wearable Sensors to Monitor Physical Activities of Patients with COPD: A Comparison of Classifier Performance. In *Proc. BSN*, Berkeley, CA, pp. 236–241, 2009.

[15] J. R. Quinlan. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.

[16] A. Reiss and D. Stricker. Creating and Benchmarking a New Dataset for Physical Activity Monitoring. In *Proc. 5th Workshop on Affect and Behaviour Related Assistance (ABRA)*, Crete, Greece, 2012.

[17] A. Reiss and D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. In *Proc. ISWC*, Newcastle, UK, 2012.

[18] A. Reiss, M. Weber, and D. Stricker. Exploring and Extending the Boundaries of Physical Activity Recognition. In *Proc. IEEE SMC Workshop on Robust Machine Learning Techniques for Human Activity Recognition*, Anchorage, AK, pp. 46–50, 2011.

[19] D. Roggen, S. Magnenat, M. Waibel, and G. Tröster. Wearable Computing: Designing and Sharing Activity Recognition Systems Across Platforms. *IEEE Robotics & Automation Magazine*, 18(2):83–95, 2011.

[20] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Supervised and unsupervised classification approaches for human activity recognition using body-mounted sensors. In *Proc. ESANN*, Bruges, Belgium, pp. 417–422, 2012.

[21] T. van Kasteren, H. Alemdar, and C. Ersoy. Effective Performance Metrics for Evaluating Activity Recognition Methods. In *Proc. ARCS*, Como, Italy, 2011.

[22] J. A. Ward, P. Lukowicz, and H. W. Gellersen. Performance Metrics for Activity Recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1), 2011.