Creating and Benchmarking a New Dataset for Physical Activity Monitoring

Conference Paper · June 2012

DOI: 10.1145/2413097.2413148

Didier Stricker
Technische Universität Kaiserslautern
322 PUBLICATIONS **4,266** CITATIONS

Creating and Benchmarking a New Dataset for Physical Activity Monitoring

Attila Reiss
German Research Center for
Artificial Intelligence (DFKI)
Department of Augmented Vision
Trippstader Str. 122
D-67663 Kaiserslautern, Germany
attila.reiss@dfki.de

Didier Stricker
German Research Center for
Artificial Intelligence (DFKI)
Department of Augmented Vision
Trippstader Str. 122
D-67663 Kaiserslautern, Germany
didier.stricker@dfki.de

ABSTRACT

Physical activity monitoring has recently become an important field in wearable computing research. However, there is a lack of a commonly used, standard dataset and established benchmarking problems. In this work, a new dataset for physical activity monitoring — recorded from 9 subjects, wearing 3 inertial measurement units and a heart rate monitor, and performing 18 different activities — is created and made publicly available. Moreover, 4 classification problems are benchmarked on the dataset, using a standard data processing chain and 5 different classifiers. The benchmark shows the difficulty of the classification tasks and exposes some challenges, defined by e.g. a high number of activities and personalization.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Information Systems Applications—*Miscellaneous*; I.5.4 [**Pattern Recognition**]: Applications—*Signal processing*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Health*; F.2.0 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity—*General*

General Terms

Algorithms, Experimentation, Measurement, Performance, Standardization

Keywords

Dataset, Benchmark, Physical Activity Monitoring, Activity Recognition, Intensity Estimation, Wearable Sensors, Performance Measures, Data Processing, Classification

PETRA'12, June 6 - 8, 2012, Crete Island, Greece.
Copyright ©2012 ACM ISBN 978-1-4503-1300-1/12/06 ... $10.00

1. INTRODUCTION

Most established research fields are characterized amongst others with standard, benchmarked datasets. Such datasets have many benefits: different and new approaches can be compared to each other, no research time has to be spent on laborious data collection, etc. Moreover, benchmarking a dataset is also important to show the difficulty of different tasks and to show where the challenges are in a research field.

Recently, datasets for different fields of activity and context recognition have become publicly available, e.g. the PlaceLab dataset [8] or a dataset using simple binary sensing nodes to observe activities performed in a home environment [25]. The Opportunity dataset [9, 19] provides a large recording of 12 subjects performing morning activities (activities of daily living, ADL), using numerous sensors on the body and in the environment. Moreover, [22] presents a benchmark of 4 classification techniques applied to multiple tasks defined on the latter dataset.

The goals of physical (aerobic) activity monitoring are to estimate the intensity and to recognize activities like sitting, walking, running or cycling. The focus and challenges in this field are — compared to activity recognition in e.g. ADL or industrial scenarios — different, due to differing conditions (e.g. sensor setup: only a few, wearable sensors can be used). Since the characteristic of the activities in this field also significantly differ from the specific activities of home or industrial settings, different approaches are required, e.g. features are calculated usually on longer intervals, etc. Therefore, there is a need for datasets specifically created for physical activity monitoring, and benchmarked on tasks defined in this field. However, only a few, limited datasets are publicly available in this research field. The DLR dataset[1] contains 4.5 hours of annotated data from 7 activities performed by 16 subjects, wearing one belt-mounted inertial measurement unit (IMU). [2] presents a data recording of 20 different activities with 20 subjects, wearing five 2-axis accelerometers, and shows results in activity recognition with 4 different classifiers. The Opportunity dataset contains 4 basic modes of locomotion, included as a task in the benchmark of [22]. A protocol of 10 different activities was followed by 44 subjects in [28], wearing one 3D-accelerometer. Benchmarking on the activity recognition task is presented in [28], using an

[1] available at `http://www.kn-s.dlr.de/activity/`

SVM classifier with different sets of features. The PAMAP dataset [16] was recorded on 14 activities with 8 subjects, wearing 3 IMUs and a heart rate monitor.

Data recording for physical activity monitoring faces some difficulties compared to data collection in e.g. home environments, resulting in less comprehensive and established datasets. For instance, a robust hardware setup (activities like running are highly stressing the setup) consisting of only wearable sensors is required. Moreover, only online annotation of the performed activities is possible for creating a reliable ground truth, the parallel video recording for the purpose of offline annotation — a widely used method in other fields — is not feasible if some outdoor activities are also included in the data collection. As a result, there is a lack of a commonly used, standard dataset and established benchmarking problems for physical activity monitoring.

The main contributions of this work are twofold. On the one hand, a new dataset for physical activity monitoring is created and made publicly available. The following specifications — based on the conditions and limitations of the public datasets briefly described above, and experience made in previous work [16, 18] — are defined for this: a wide range of everyday, household and sport activities should be performed by an adequate number of subjects, wearing a few 3D-IMUs and a HR-monitor (physiological sensing — missing in other public datasets — is especially useful for intensity estimation: inertial sensing alone can not reliably distinguish activities with similar movement characteristic but different energy expenditure, e.g. walking and ascending stairs, or even more difficult: walking and walking with some load). On the other hand, this work also presents an initial benchmarking on various defined tasks, showing the difficulty of common classification problems and exposing some challenges. The key facts of the created dataset, and the key results of an initial classification are presented in [17]. This paper describes both the data collection (hardware setup, data collection protocol) and the benchmarking (choices made for data processing and selected performance measures) in more detail. It also discusses lessons learnt from the data recording, and presents and discusses more comprehensive results of the created benchmark.

2. DATA COLLECTION

This section presents the creation of the dataset. It describes the hardware setup and the subjects participating in the data collection. The data collection protocol is also presented, and the selected activities are justified. The section concludes with a brief description of some lessons learnt from this data recording.

2.1 Hardware setup

3 IMUs and a heart rate monitor were used as sensors during the data collection. For the inertial measurements, the Colibri wireless IMUs from Trivisio [23] were used. The sensors are relatively lightweight (48 g with battery) and small (56 × 42 × 19 mm). Each IMU contains two 3-axis MEMS accelerometers (scale: ±16 g / ±6 g, resolution: 13-bit), a 3-axis MEMS gyroscope (scale: ±1500°/s, resolution: 13-bit), and a 3-axis magneto-resistive magnetic sensor (scale: ±400 μT, resolution: 12-bit), all sampled at 100 Hz. To obtain heart rate information, the BM-CS5SR HR-monitor from BM innovations GmbH [3] was used, providing heart rate values with approximately 9 Hz.

Table 1: Protocol of data collection

Activity	Duration [Min]	Activity	Duration [Min]
Lie	3	Descend stairs	1
Sit	3	Break	2
Stand	3	Normal walk	3
Iron	3	Break	1
Break	1	Nordic walk	3
Vacuum clean	3	Break	1
Break	1	Cycle	3
Ascend stairs	1	Break	1
Break	2	Run	3
Descend stairs	1	Break	2
Break	1	Rope jump	2
Ascend stairs	1		

The sensors are placed onto 3 different body positions. A chest sensor fixation includes one IMU and the heart rate chest strap. The second IMU is attached over the wrist on the dominant arm, and the third IMU on the dominant side's ankle, both are fixed with sensor straps. The reason why only 3 sensor placements are used is, that previous work (e.g. [13]) showed, that in the trade-off between classification performance and number of sensors, using 3 sensor locations is the most effective. In systems for physical activity monitoring, the number of sensor placements should be kept as low as possible for reasons of practicability and comfort — since users of such systems usually wear them for many hours a day. On the other hand, previous work also showed, that less than 3 sensor placements are not sufficient for an accurate activity recognition [15].

A Viliv S5 UMPC (Intel Atom Z520 1.33GHz CPU and 1GB of RAM [26]) was used as data collection companion unit. The main advantage of this device is a battery time of up to 6 hours. A custom bag was made for this companion unit and the 2 USB-dongles additionally required for the wireless data transfer — one for the IMUs and one for the HR-monitor, respectively — which was carried by the subjects fixed on their belt. Labeling of the currently performed activities was done via a GUI specifically developed for this purpose on the UMPC. The collection of all raw sensory data and the labeling were running in separate threads in an application running on the companion unit, to lighten the synchronization of all collected data.

2.2 Subjects

In total 9 subjects participated in the data collection, 8 males and 1 female. The subjects were mainly employees or students at our research institute, aged 27.22 ±3.31 years, and having a BMI of 25.11 ±2.62 kgm^{-2}. One subject was left-handed, all the others were right-handed. The data collection took place in autumn 2011.

2.3 Data collection protocol

A protocol containing 12 activities was defined for the data collection, shown in Table 1. A criterion for selecting activities was on the one hand that the basic activities (walking, running, cycling and Nordic walking) and postures (lying, sitting and standing), traditionally used in related work, should be included. On the other hand, everyday (ascending and descending stairs), household (ironing, vacuum cleaning) and fitness (rope jumping) activities were also included to cover a wide range of activities. Each of the subjects had to follow this protocol, performing all defined activities in the way most suitable for the subject.

Furthermore, a list of optional activities to perform was also suggested to the subjects. The idea of these optional activities was to further enrich the range of activities in the recorded dataset. Activities from this optional list were only performed by some of the subjects if different circumstances made it possible, e.g. if the subject had additional free time after completing the protocol, if there was equipment available to be able to perform an optional activity, and if the hardware setup made further data recording possible. In total, 6 different optional activities were performed by some of the subjects: watching TV, computer work, car driving, folding laundry, house cleaning and playing soccer.

The recorded dataset contains therefore in total data from 18 different activities. A brief description of each of these activities can be found attached to the published dataset. Most of the activities from the protocol were performed over approximately 3 minutes, except ascending/descending stairs (due to the limitatinos of the building where the indoor activities were carried out) and rope jumping (to avoid exhaustion of the subjects). The optional activities were performed as long as the subjects wished, or as long as it took to finish a task (e.g. arriving with the car at home, or completely finish dusting a bookshelf). Over 10 hours of data were collected altogether, from which nearly 8 hours were labeled as one of the 18 activities performed during data collection. The dataset is made publicly available, and can be downloaded from the PAMAP [10] project's website[2].

2.4 Data collection: lessons learnt

Attaching the sensors and the custom bag was straightforward, the entire setup time was not longer than 5 minutes. All subjects reported, that the sensor fixations were comfortable and did not restrict normal movements at all. Only the custom bag felt sometimes uncomfortable during intensive movements (e.g. running). A smaller solution for the companion unit — using e.g. a smartphone — would be recommendable for similar data collections. One aspect, which should not be underestimated, is the weather. Opposed to most of the datasets collected in the research field of activity recognition (recorded e.g. in home or industrial settings), a significant part of the dataset presented in this paper had to be recorded outdoors. Since most of the subjects preferred not to run or cycle in too hot, cold or rainy conditions, and the entire data collection took several days, careful planning and consulting the weather forecast was required when making the schedule for the subjects.

Problems occuring during such complex and long data recordings are inevitable. There were two main reasons for data loss in the dataset. The first reason is data dropping caused by wireless data transfer. However, this was not too significant: the 3 IMUs had a real sampling frequency (a calculated sampling rate corrected with overall data dropping occurrence) of 99.63 Hz, 99.78 Hz and 99.65 Hz on the hand, chest and ankle placements, respectively. Data dropping on the wireless HR-monitor appeared even more rarely, and is also less critical than on the IMUs. The second, more severe reason was the somewhat fragile system setup due to the additionally required hardware components: 2 USB-dongles, a USB-hub and a USB extension cable were added to the companion unit in the custom bag. Especially during activities like running or rope jumping, the system was exposed to a lot of mechanical stress. This sometimes caused loosing connection to the sensors, or even a system crash, when the data recording had to be restarted — and in a few cases the data collection could not be recovered even this way. As a conclusion, some activities for certain subjects are partly or completely missing in the dataset (cf. the respective table attached to the published dataset, showing a summary of how much data and from how many subjects was recorded per activity). For trying to minimize such problems, it is preferable to use the entire sensor setup from one company (so that no second dongle is needed), or even better is using sensors with standard wireless communication (although the Trivisio sensors use the 2.4 GHz ISM band, they use a specific communication protocol, and thus a USB-dongle is needed for wireless data streaming). As an alternative, local storage on the sensors should be considered for future data collection, made possible by new sensor solutions appearing recently on the market.

3. BENCHMARK: BASIC CONDITIONS

Benchmarking in this work is done on the dataset presented in the previous section, using different classification techniques. This section describes the basic conditions of the benchmarking process. This includes the definition of some classification problems and the decision for performance measures used during benchmarking.

3.1 Defining the classification problems

The benchmark created in this paper only focuses on the 12 activities performed during the data collection protocol. The definition of classification problems including the 6 optional activities, and benchmarking them, is left for future work. 4 different classification problems are defined for benchmarking:

Intensity estimation task 3 classes are defined for this problem: activities of light, moderate and vigorous effort. The ground truth for this rough intensity estimation is based on the metabolic equivalent (MET) of the different activities, and is provided by [1] — as further explained in [16]. Therefore, the 3 classes are defined as following: lying, sitting, standing and ironing are regarded as activities of light effort ($<$ 3.0 METs); vacuum cleaning, descending stairs, normal walking, Nordic walking and cycling as activities of moderate effort (3.0-6.0 METs); ascending stairs, running and rope jumping as activities of vigorous effort ($>$ 6.0 METs).

Basic activity recognition task 5 classes are defined for this problem: lying, sitting/standing, walking, running and

[2] More precisely from `http://www.pamap.org/demo.html`, called the "PAMAP2_Dataset".

cycling. All other activities are discarded for this task. This classification problem refers to the many existent activity recognition applications only including these, or a similar set of few basic activities. The ground truth for this task — and for the other two activity recognition tasks presented below — is provided by the labels made during data collection. The activities sitting and standing are forming one class in this problem. This is a common restriction made in activity recognition (e.g. in [5, 11]), since an extra IMU on the thigh would be needed for a reliable differentiation of these postures. The numerous misclassifications between these two postures appearing in the results belonging to task 'all' in the benchmark (cf. Section 5) confirm, that these two activities can not be reliably distinguished with the given set of sensors.

Background activity recognition task 6 classes are defined for this problem: lying, sitting/standing, walking, running, cycling and other (this latter class consists of the remaining 6 activities of the data collection protocol: ironing, vacuum cleaning, ascending stairs, descending stairs, Nordic walking and rope jumping). The idea behind the definition of this task is, that in physical activity monitoring, users always perform meaningful activities. However, there are countless number of activities, and — apart from a few, for the particular application relevant activities — an exact recognition is not needed. On the other hand, ignoring these other activities would limit the applicability of the application. Therefore, the introduction of a background activity class is justified. The idea of the background activity class, and how it significantly increases the complexity of the classification problem, is further explained in [18].

All activity recognition task 12 classes are defined for this problem, corresponding to the 12 activities of the data collection protocol.

3.2 Performance measures

For physical activity monitoring, data is usually collected following a given protocol (as in this work), and it is common practice to delete the beginning and the end of each labeled activity (10 seconds are deleted in this work, respectively). Therefore, opposed to e.g. activity recognition in home or industrial settings, the ground truth is much less fragmented, and there is less variability in event (activity) length. For continuous activity recognition, new error metrics were introduced recently, e.g. insertion, deletion, merge, fragmentation, overfill, etc. [24, 27]. However, the goals of physical activity monitoring — as justified above — are usually restricted to frame by frame recognition (thus not the events are important, but the time spent performing each of the activities). Therefore, the frame by frame evaluation methods describe the performance of the used classifiers well, and are regarded as sufficient for benchmarking in this work.

The commonly used performance measures are used for creating the benchmark: precision, recall, F-measure and accuracy. For the definition of these metrics assume, that a confusion matrix is given by its entries P_{ij}, 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 (refering to the number of samples annotated as activity i), and R_j the sum of all entries in the column j of the matrix (refering 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 classes: $1...C$. Using this notation, the performance measures used in this paper are defined as following:

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

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

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

$$accuracy = \frac{1}{N}\sum_{i=1}^{C} P_{ii} \tag{4}$$

It has to be noted, that from the defined metrics, accuracy only considers the total number of samples. For the other 3 metrics, class imbalance is taken into account, since normalization is done with using the total number of samples for each activity class separately. This different behaviour of the performance measures is important, since fewer samples from some activities in a dataset are not necessarily due to lesser importance of these activities, but could be caused by e.g. a more difficult data capture of these activities, as discussed in Section 2. Some results in Section 5 will also point out the difference between the performance metrics, and how these results should be interpreted.

4. DATA PROCESSING

This section describes the data processing used for the benchmarking process of this work. The data processing follows a classical approach, similar e.g. to the activity recognition chain presented in [20]. Standard methods are used in the different steps of data processing, e.g. for feature extraction or for classification. The goal of this work is not aiming for the best performance on the different classification tasks, but to provide a baseline characterization of the defined classification problems. The data processing chain presented in this paper is shown in Figure 1, the different processing steps are further described within this section.

4.1 Preprocessing

The data collection described in Section 2 provides timestamped raw sensory data from the 3 IMUs and the heart rate monitor, and timestamped activity labels. All this data is synchronized in the preprocessing step, after which synchronized, timestamped and labeled acceleration[3] and heart rate data is available. For dealing with wireless data loss, linear interpolation is used as a simple method, which could be replaced by more complex methods [21] in future improved approaches. To avoid dealing with eventual transient activities, 10 seconds from the beginning and the end of each labeled activity is deleted, respectively.

[3] Previous work shows (e.g. in [12]), that for different tasks in physical activity monitoring (e.g. activity recognition or intensity estimation), accelerometers outperform gyroscopes. Therefore, from all 3 IMUs, only data from the accelerometers is used in the subsequent data processing steps.

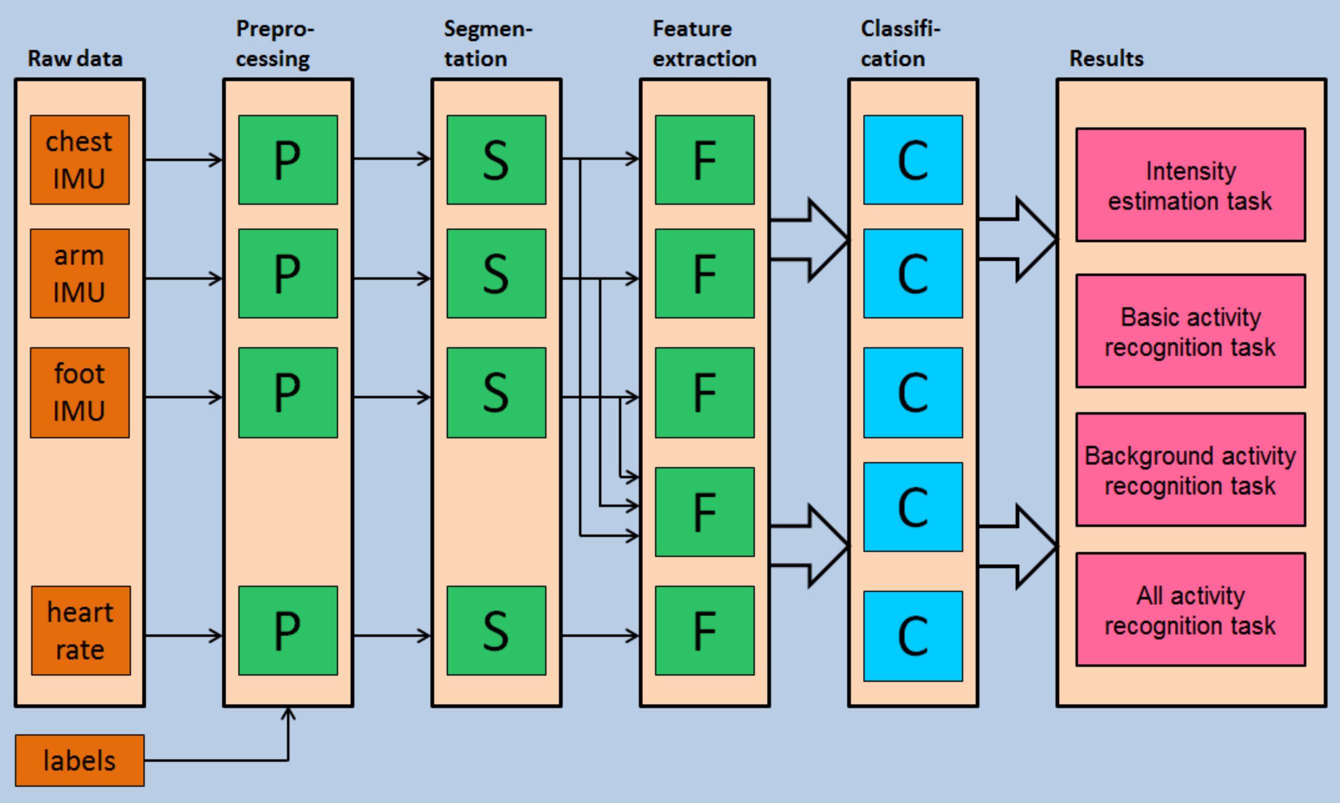
Raw data
Prepro-cessing
Segmen-tation
Feature extraction
Classifi-cation
Results
chest IMU
arm IMU
foot IMU
heart rate
labels
P
S
F
C
Intensity estimation task
Basic activity recognition task
Background activity recognition task
All activity recognition task

Figure 1: The data processing chain, consisting of the preprocessing step (P), the segmentation step (S), the feature extraction step (F), and the classification step. All 5 classifiers (C) are applied on all 4 defined classification problems.

4.2 Segmentation

Previous work shows (e.g. in [7]), that for segmentation there is no single best window length for all activities. To obtain at least 2 or 3 periods of all different periodic movements, a window length of about 3-5 seconds is reasonable. Therefore, and to assure effective FFT-calculation, a window size of 512 samples was selected. The preprocessed data is segmented using a sliding window with the defined 5.12 seconds of window size, shifted by 1 second between consecutive windows.

4.3 Feature extraction

From the segmented 3D-acceleration data, various signal features were calculated in both time and frequency domain. In addition to the most commonly used features in related work (mean, median, standard deviation, peak acceleration and energy), some other features — also proved to be useful in previous work — were calculated, too. The feature absolute integral was successfully used to estimate the metabolic equivalent in e.g. [12]. Correlation between each pair of axes is especially useful for differentiating among activities that involve translation in just one or multiple dimensions, e.g. walking, running vs. ascending stairs [14]. Power ratio of the frequency bands 0–2.75 Hz and 0–5 Hz proved to be useful in [15], while peak frequency of the power spectral density (PSD) was used for the detection of cyclic activities in e.g. [4]. Spectral entropy of the normalized PSD is a useful feature for differentiating between locomotion activities (walking, running) and cycling [4].

The above mentioned signal features, extracted from the 3D-acceleration data, are computed for each axis separately, and for the 3 axes together, too [16]. Moreover, since synchronized data from the 3 IMUs is available, combining sensors of different placements is possible. The features mean, standard deviation, absolute integral and energy are calculated on 3 axes of each of the IMUs pairwise (e.g. ankle + chest sensor placement) weighted accumulated, and a weighted sum for all the 3 sensors together is also added [16].

From the heart rate data, the features (normalized) mean and gradient are calculated. Normalization is done on the interval defined by resting and maximum HR, and proved to be useful in e.g. [15]. In total, 137 features were extracted from each data window of 512 samples: 133 features from IMU acceleration data and 4 features from the heart rate data. No feature selection or reduction of the feature space is applied on this feature set, thus all features are used for each of the classifiers presented in the next section.

4.4 Classification

The extracted features serve as input for the next processing step, the classification. The benefit of using the data processing chain (presented in Figure 1) is amongst others its modularity, which allows to easily test different classifier modules. Five different classifiers were selected from the Weka toolkit [6] for creating the benchmark in this work. These classification approaches are frequently used in related work, and represent a wide range of classifier complexity. The five classifiers are listed below, together with the parameters differing from the default values set in the Weka toolkit. These parameters were determined heuristically and used successfully in previous work [15, 18]. Moreover, for reproducibility and easier comparibility with future results, the exact definition (scheme) of each of the classifiers — as given in the Weka toolkit — is included in the following list of the five classifiers:

1. Decision tree (C4.5)
   - confidenceFactor = 0.15
   - minNumObj = 50
   - *Scheme:weka.classifiers.trees.J48 -C 0.15 -M 50*
2. Boosted C4.5 decision tree
   - confidenceFactor = 0.15 (in the decision tree)
   - minNumObj = 50 (in the decision tree)
   - *Scheme:weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 – -C 0.15 -M 50*
3. Bagging C4.5 decision tree
   - confidenceFactor = 0.15 (in the decision tree)
   - minNumObj = 50 (in the decision tree)
   - *Scheme:weka.classifiers.meta.Bagging -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 – -C 0.15 -M 50*
4. Naive Bayes
   - *Scheme:weka.classifiers.bayes.NaiveBayes*
5. kNN
   - KNN = 7 (number of neighbours)
   - *Scheme:weka.classifiers.lazy.IBk -K 7 -W 0*

5. RESULTS AND DISCUSSION

This section presents the 4 performance measures of all 5 classifiers applied to all 4 classification problems (Table 3 - Table 6). Moreover, both subject dependent and subject independent evaluation results are shown for all classifier/classification problem combinations. Subject independent evaluation is done with leave-one-subject-out (LOSO) 9-fold cross-validation, while subject dependent evaluation is done with standard x-fold cross-validation ($x = 9$ was choosen to have the same number of folds for both validation techniques). Although [18] showed, that usually subject independent validation techniques should be applied for the evaluation of activity monitoring systems, to create a widely used and comparable benchmark, results of the subject dependent evaluation are shown as well.

Although the main goal of this work is to give a benchmark on the various classification problems — and more advanced approaches of the data processing chain will outperform the presented results — some conclusions from the results are drawn and discussed further in this section.

Overall, the best performance was achieved by the kNN and the boosted decision tree classifiers. Furthermore, it is interesting to observe, how the Naive Bayes classifier performs on the different tasks. On classification problems having clear class boundaries (the problems 'basic' and 'all') it performs better, than the decision tree classifier. On the other hand, the decision tree classifier outperforms the Naive Bayes classifier on the other two tasks (the problems 'intensity' and 'background'): these tasks have classes containing multiple activities, thus it is difficult to define the class boundaries with the Naive Bayes classifier — opposed to the decision tree classifier. Furthermore, comparing the results of subject dependent and subject independent evaluation shows, that the former indicates highly 'optimistic' performance, confirming the results of [18].

Table 2: Confusion matrix on the 'background' task using boosted decision tree classifier and subject independent evaluation

Annotated	**Recognized activity**					
activity	Lie	Sit/Stand	Walk	Run	Cycle	Other
Lie	1510	218	0	0	0	0
Sit/Stand	56	3091	0	0	7	204
Walk	0	1	1891	0	0	301
Run	0	0	0	830	0	5
Cycle	0	7	2	0	1418	45
Other	0	106	69	0	26	7032

Analyzing the confusion matrices from the results (only one is shown in this paper due to limited space), more implications can be done. For instance, the best classifiers not only achieve $\sim$ 96% on the intensity estimation task, but misclassifications only appear into "neighbour" intensity classes, thus no samples annotated as light intensity were classified into the vigorous intensity class, and vice versa. Regarding the confusion matrices of the 'background' task (Table 2 shows the confusion matrix of the boosted decision tree classifier), a conclusion can be drawn on why the complexity of the classification problem increased so significantly compared to the 'basic' task: the characteristic of some of the introduced background activities overlap with some of the basic activity classes to be recognized. For instance, the background activity *ironing* has a similar characteristic as for example when talking and gesticulating during *standing*, thus the separation of these two activities is a difficult classification problem. Finally, the confusion matrices belonging to the task 'all' show many misclassifications between *sitting* and *standing*, confirming that these two activities (postures) can not be reliably distinguished with the given set of sensors.

In general looking at the benchmark results, very good performance is achieved on all 4 tasks ($\sim$ 90% and more, which is comparable to state of the art results on similar classification problems, e.g. in [15]). However, there are two important challenges defined by the benchmark, where more complex approaches in future work should improve the performance. On the one hand, by increasing the number of activities to be recognized — while keeping the same sensor set — the difficulty of the task exceeds the potential of standard methods. This not only applies for the task 'all', but for the 'background' task, too: by introducing an *other* activity class for all the background activities, the complexity of the classification problem significantly increases, and thus the performance drops using the same standard approaches. On the other hand, when comparing classification performance individually for the 9 subjects, a high variance can be observed. This strongly increases with the increase of task complexity: the individual performance on the 'basic'

Table 3: Performance measures on the 'Intensity estimation task'

	Standard 9-fold cross-validation				LOSO 9-fold cross-validation			
Classifier	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Decision tree (C4.5)	0.9796	0.9783	0.9789	0.9823	0.9490	0.9364	0.9426	0.9526
Boosted C4.5	0.9989	0.9983	0.9986	0.9988	0.9472	0.9564	0.9518	0.9587
Bagging C4.5	0.9853	0.9809	0.9831	0.9866	0.9591	0.9372	0.9480	0.9552
Naive Bayes	0.9157	0.8553	0.8845	0.9310	0.8986	0.8526	0.8750	0.9251
kNN	0.9985	0.9987	0.9986	0.9982	0.9488	0.9724	0.9604	0.9666

Table 4: Performance measures on the 'Basic activity recognition task'

	Standard 9-fold cross-validation				LOSO 9-fold cross-validation			
Classifier	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Decision tree (C4.5)	0.9968	0.9968	0.9968	0.9970	0.9349	0.9454	0.9401	0.9447
Boosted C4.5	0.9997	0.9994	0.9995	0.9995	0.9764	0.9825	0.9794	0.9785
Bagging C4.5	0.9971	0.9968	0.9970	0.9971	0.9346	0.9439	0.9392	0.9433
Naive Bayes	0.9899	0.9943	0.9921	0.9923	0.9670	0.9737	0.9703	0.9705
kNN	1.0000	1.0000	1.0000	1.0000	0.9955	0.9922	0.9938	0.9932

task, using the boosted decision tree classifier, varies between 93.99% and 100%, while on the 'all' task it varies between 74.02% and 100%. Therefore, especially on the more difficult classification problems defined in this paper, personalization approaches (subject dependent training) could significantly improve on the results of the benchmark, and are highly encouraged.

6. CONCLUSION AND FUTURE WORK

This work presented the creation and benchmarking of a dataset for physical activity monitoring. The dataset was recorded with 9 subjects performing 18 different activities. 3 sensor placements were used, over 10 hours of data were collected altogether. 4 classification problems of different complexity were defined on the presented dataset. Standard approaches of data processing were used during the benchmarking process, including 5 different classifiers. The presented results mainly serve to characterize the difficulty of the different tasks. Improved approaches should outperform these results with e.g. applying feature selection or reduction of the feature space (with e.g. PCA), or by using more advanced classifiers.

The definition and benchmark of classification problems including the 6 optional activities remains for future work. Moreover, the developed data collection system of this work is further used by elderly in the project PAMAP during organized clinical trials. Therefore, it is planned to enlarge the current dataset with various physical activities performed by elderly subjects.

7. ACKNOWLEDGEMENTS

This work has been performed within the project PAMAP [10] funded under the AAL Joint Programme (AAL-2008-1). The authors would like to thank the project partners, the EU and national funding authorities for the financial support. For more information, please visit the project's website. Many thanks to Benjamin Schenkenberger for his important contribution in developing the data collection application for the Viliv companion unit, and to Gabi Felske for helping producing the fixations for the hardware setup. The authors also would like to thank the DFKI employees and students participating in the data collection.

8. 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):498–504, Sept. 2000.

[2] L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. In *Proc. 2nd Int. Conf. Pervasive Comput*, pp. 1–17, 2004.

[3] BM-innovations. http://www.bm-innovations.com.

[4] M. Ermes, J. Pärkkä, and L. Cluitmans. Advancing from offline to online activity recognition with wearable sensors. In *30th Annual International IEEE EMBS Conference*, pp. 4451–4454, Jan. 2008.

[5] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Trans. Inf. Technol. Biomed.*, 12(1):20–26, Jan. 2008.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1), 2009.

[7] T. Huynh and B. Schiele. Analyzing features for activity recognition. In *sOc-EUSAI '05*, pp. 159–163. ACM Press, 2005.

[8] S. Intille, K. Larson, E. Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson. Using a live-in laboratory for ubiquitous computing research. *Proc. Int. Conf. on Pervasive Computing*, pp. 349–365, 2006.

[9] P. Lukowicz, G. Pirkl, D. Bannach, F. Wagner, A. Calatroni, K. Förster, T. Holleczek, M. Rossi, D. Roggen, G. Tröster, and Others. Recording a complex, multi modal activity data set for context

Table 5: Performance measures on the 'Background activity recognition task'

	Standard 9-fold cross-validation				LOSO 9-fold cross-validation			
Classifier	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Decision tree (C4.5)	0.9784	0.9701	0.9743	0.9709	0.8905	0.8635	0.8768	0.8722
Boosted C4.5	0.9991	0.9979	0.9985	0.9980	0.9559	0.9310	0.9433	0.9377
Bagging C4.5	0.9881	0.9766	0.9823	0.9787	0.9160	0.8937	0.9047	0.9042
Naive Bayes	0.8905	0.9314	0.9105	0.8508	0.8818	0.8931	0.8874	0.8308
kNN	0.9982	0.9966	0.9974	0.9957	0.9428	0.9458	0.9443	0.9264

Table 6: Performance measures on the 'All activity recognition task'

	Standard 9-fold cross-validation				LOSO 9-fold cross-validation			
Classifier	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Decision tree (C4.5)	0.9554	0.9563	0.9558	0.9546	0.8376	0.8226	0.8300	0.8244
Boosted C4.5	0.9974	0.9973	0.9974	0.9969	0.8908	0.8947	0.8928	0.8796
Bagging C4.5	0.9660	0.9674	0.9667	0.9666	0.8625	0.8489	0.8556	0.8554
Naive Bayes	0.9419	0.9519	0.9469	0.9438	0.8172	0.8561	0.8362	0.8365
kNN	0.9946	0.9937	0.9942	0.9925	0.9123	0.9097	0.9110	0.8924

recognition. In *23rd International Conference on Architecture of Computing Systems (ARCS)*, pp. 1–6. VDE, 2010.

[10] PAMAP (Physical Activity Monitoring for Aging People). http://www.pamap.org.

[11] J. Pärkkä, L. Cluitmans, and M. Ermes. Personalization algorithm for real-time activity recognition using PDA, wireless motion bands, and binary decision tree. *IEEE Trans. Inf. Technol. Biomed.*, 14(5):1211–5, Sept. 2010.

[12] J. Pärkkä, M. Ermes, K. Antila, M. van Gils, A. Mänttäri, and H. Nieminen. Estimating intensity of physical activity: a comparison of wearable accelerometer and gyro sensors and 3 sensor locations. *29th Annual International IEEE EMBS Conference*, pp. 1511–4, 2007.

[13] 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 *Body Sensor Networks*, pp. 236–241, 2009.

[14] N. Ravi, N. Dandekar, P. Mysore, and M. Littman. Activity recognition from accelerometer data. In *17th Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pp. 1541–1546, 2005.

[15] A. Reiss and D. Stricker. Introducing a Modular Activity Monitoring System. In *33rd Annual International IEEE EMBS Conference*, pp. 5621–5624, 2011.

[16] A. Reiss and D. Stricker. Towards Global Aerobic Activity Monitoring. In *4th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA)*, 2011.

[17] A. Reiss and D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. In *16th IEEE International Symposium on Wearable Computers (ISWC)*, 2012.

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

[19] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, and Others. Collecting complex activity datasets in highly rich networked sensor environments. In *Seventh Int. Conf. on Networked Sensing Systems (INSS)*, pp. 233–240. IEEE, 2010.

[20] 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.

[21] M. Saar-Tsechansky and F. Provost. Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research*, 8:1625–1657, 2007.

[22] H. Sagha, S. T. Digumarti, R. Chavarriaga, A. Calatroni, D. Roggen, and G. Tr. Benchmarking classification techniques using the Opportunity human activity dataset. In *IEEE SMC Workshop on Robust Machine Learning Techniques for Human Activity Recognition*, pp. 36–40, 2011.

[23] Trivisio. http://www.trivisio.com.

[24] T. van Kasteren, H. Alemdar, and C. Ersoy. Effective Performance Metrics for Evaluating Activity Recognition Methods. In *ARCS 2011 - 24th International Conference on Architecture of Computing Systems*, 2011.

[25] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp)*, pp. 1–9. ACM Press, 2008.

[26] Viliv-S5. http://www.myviliv.com/ces/main_s5.html.

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

[28] Y. Xue and L. Jin. A Naturalistic 3D Acceleration-based Activity Dataset & Benchmark Evaluations. In *International Conference on Systems, Man and Cybernetics (SMC)*, pp. 4081–4085, 2010.