Let's (not) Stick Together: Pairwise Similarity Biases Cross-Validation in Activity Recognition

Nils Y. Hammerla, Thomas Plötz
Open Lab
Newcastle University
Newcastle upon Tyne, UK
{firstname.lastname}@newcastle.ac.uk

ABSTRACT

The ability to generalise towards either new users or unforeseen behaviours is a key requirement for activity recognition systems in ubiquitous computing. Differences in recognition performance for the two application cases can be significant, and user-dependent performance is typically assumed to be an upper bound on performance. We demonstrate that this assumption does not hold for the widely used cross-validation evaluation scheme that is typically employed both during system bootstrapping and for reporting results. We describe how the characteristics of segmented time-series data render random cross-validation a poor fit, as adjacent segments are not statistically independent. We develop an alternative approach – meta-segmented cross validation – that explicitly circumvents this issue and evaluate it on two data-sets. Results indicate a significant drop in performance across a variety of feature extraction and classification methods if this bias is removed, and that prolonged, repetitive activities are particularly affected.

Author Keywords

Activity Recognition; Evaluation; Cross validation; Model selection

INTRODUCTION

The ability to generalise towards new users and unforeseen behaviour is a crucial aspect of activity recognition systems in ubiquitous computing. From the onset of design, different components are typically selected through experimentation on (pilot) data-sets, in order to maximise the performance expected under real-life conditions. To allow reliable estimation of the performance of activity recognition systems, practitioners in ubiquitous computing have adopted evaluation schemes and performance metrics from other fields that employ machine learning, such as speech recognition or computer vision. The evaluation strategies typically employed in ubiquitous computing aim to investigate two different aspects of generalisation performance: i) the *user-independent* performance expected for new users of a system, and ii) the *user-dependent* performance expected for new behaviour of a known user. The evaluation strategy for a particular system depends on the envisioned use-case and aim of the study. For example, exploratory work that investigates a new sensing modality will likely evaluate user-dependent performance. A diagnostic system aimed for clinical use, on the other hand, will follow a user-independent evaluation approach, simply as the generalisation to new "patients" is of crucial interest.

Overall, recognition systems perform significantly better in the user-dependent scenario, where the difference in performance can be vast. This is well known, as for example Bulling et al. summarise: "Overall lower performance in the person-independent case is expected, as different users tend to perform activities differently" [7]. It is assumed that if more data from a more representative set of people could be obtained, this difference between user-dependent and independent evaluation would vanish. Effectively, the user-dependent performance is considered an upper bound for system performance, which is of particular interest for explorative studies that aim to demonstrate the feasibility of a novel technical approach.

In this paper we demonstrate that this assumption is wrong for random cross-validation, a popular evaluation scheme in activity recognition. We illustrate how (stratified) cross-validation, the standard approach to model selection in machine learning, is unfit for segmented time-series data, typical for ubiquitous computing in which subsequent samples are not fully independent. Results obtained with cross-validation are biased towards approaches that preserve similarity between adjacent segments of time-series data, which affects both feature extraction and classification approaches, and has a significant effect on the resulting performance estimates.

We illustrate the different types of evaluation schemes in activity recognition and discuss whether the type of data common for ubiquitous computing is a suitable basis for cross-validation experiments. We highlight how the recordings close in time are not statistically independent, and what effect this is likely to have for performance evaluation in activity recognition. In light of this *neighbourhood bias* we present a novel approach to cross-validation that explicitly circumvents this issue. In experiments on two publicly available data-sets typical for activity recognition in ubiquitous computing we demonstrate the effect of this bias empirically. Results

UbiComp '15, September 7-11, 2015, Osaka, Japan.
Copyright 2015 © ACM 978-1-4503-3574-4/15/09...$15.00.
http://dx.doi.org/10.1145/2750858.2807551

indicate that particularly physical activities such as walking or running benefit from the similarity of adjacent segments of time-series data, which leads to overly optimistic performance estimates.

MODEL SELECTION IN ACTIVITY RECOGNITION

Crucial for the development of a recognition system is the use of large collections of data recordings from a representative pool of participants. The nature of these data-sets depends on the application and may involve laboratory-based recordings that are driven by some form of (activity) protocol, may rely on longitudinal recordings under naturalistic conditions, or correspond to a combination of the two. Different data-sets afford different schemes for evaluation of recognition systems. While these schemes differ in approach they are similar in that the data-set is split into at least two disjoint sets, one of which is used during parameter selection (training) of the system while the other is used to test how the system generalises towards unseen recordings. The individual evaluation scheme reflects what aspect of generalisation is deemed most informative for future work or which best reflects the requirements of the envisioned use-case.

The simplest evaluation scheme corresponds to a *hold-out* validation. The data-set is split according to some heuristic into training and test-set, for example, to represent a split to allow 80% of the data for training and 20% of the data for testing. This can be performed on the complete data-set or specific subsets that result from the study protocol. For example, when participants perform the same routine three times, two may be used for training and one for testing. Depending on the specific setup, this evaluation scheme gives an insight into *user-independent* performance of the system, where one, or a subset of, participants is held back for testing; or corresponds to a *user-dependent* performance evaluation if following an approach similar to the one highlighted above (split based on protocol).

Hold-out validation, while simple and well established in, e.g. general machine learning, may not give sufficient insight into the performance of the system for robust model selection [5], as it is unsuitable for estimation of the variance in performance. Three schemes are common to evaluate the variability in performance of a system: i) repeated hold-out validation selects a number of different training/test sets and estimates the performance for each case; ii) the closely related leave-one-subject-out (LOSO) evaluation aims to investigate *user-independent* performance of a system by splitting the data based on the participants; and iii) (random) k-fold cross-validation (CV), where the data-set is split randomly at the lowest possible level into k equal sized disjoint sets (e.g. segments extracted from time-series, images in computer vision). In each case the system is trained and tested a number of times, where each time a different set is selected for testing and all remaining sets are used for training of the system. The mean over all evaluations represents the final performance figure. The mean, standard deviation, and abs. difference in recognition performance form the basis for model selection to reliably answer which classification approach, for example, performs best for a problem (e.g. through paired t-tests).

Not all evaluation schemes are suitable for every type of data-set. There are various reasons why a particular scheme may be unsuitable. If only data from few users is collected as, e.g. recordings are prohibitively expensive or time-consuming, it is unlikely that LOSO validation will give an adequate reflection of the performance, simply because the variability between the very few users is too large. LOSO validation may further be difficult to apply for data-sets that are very large (e.g. hundreds of participants), if training or testing the system is computationally expensive. In these cases, practitioners usually aim to estimate *user-dependent* performance of the system, e.g. by following a (repeated) hold-out validation or CV scheme.

It is best practice to select the evaluation scheme based on these practical considerations, as they are usually deemed to provide reliable estimates of the performance of the system, simply from different perspectives. While performance metrics used in ubiquitous computing have been subject to intensive study (see e.g. [35]), so far no work has investigated these different evaluation schemes for specific characteristics or biases. This is surprising, as the performance difference between the evaluation schemes can be significant.

EVALUATION STRATEGIES IN RELATED WORK

The sensor data in ubiquitous computing is usually analysed using a pipeline-based approach [7]. For each step of the pipeline a multitude of options exist that span from different preprocessing techniques or feature extraction approaches to sophisticated classification algorithms. When designing a recognition system, practitioners must select the most suitable components to maximise the generalisation performance while potentially abiding some additional constraints (e.g. resource availability in mobile settings, interpretability, etc.). This requires reliable estimation of the performance of the individual components at design time.

In order to investigate if there is a link between the evaluation scheme and design of activity recognition systems, we select recent works from our field that apply the different evaluation schemes outlined above. To limit the scope of this review, we focus on work that utilises body-worn movement sensors or mobile phones, and follow a sliding-window segmentation approach [7]. The papers selected are listed in Table 1. The evaluation scheme in each work is classified as (a combination of) i) random (stratified) cross-validation (CV), ii) Leave-one-subject-out validation (LOSO), and iii) Hold-out validation. Evaluation schemes described as CV by the respective authors, but which actually correspond to repeated hold-out validation (e.g. leave-one-*day*-out [3]) are classified as hold-out validation.

The main design decision we focus on here is the classification approach. We do so as classifiers are usually selected as an off-the-shelf component made available in common machine learning frameworks such as WEKA [13]. Other components of the pipeline, such as the feature extraction approach, are often hand-crafted for each problem setting and therefore difficult to compare between application scenarios. In some work the authors report performance figures for a variety of approaches. In work that reports the performance

Year	Ref.	Target	Sensor	# participants	CV	LOSO	Hold-out	Classifier
2014	Bogolomov et al. [6]	Stress levels	phone	117			x	Random Forest
2014	Gu et al. [11]	Sleep stage	phone	60			x	CRF
2014	Peterek et al. [26]	Phyiscal	acc	30			x	LDA
2013	Hirano et al. [16]	Physical	various	4			x	GMM/HMM
2012	Berlin et al. [3]	Leisure activities	acc	6			x	Motifs
2014	Bulling et al. [7]	Physical	acc	2		x	x	various
2014	Lane et al. [23]	Community activities	various	123		x		SVM
2013	Gjoreski et al. [10]	Physical	acc	10		x		Ensemble
2013	Reiss et al. [31]	Physical	acc, cardio	9		x		Custom
2013	Ladha et al. [22]	Climbing	acc	53	x	x		Log. Regr.
2013	Li et al. [24]	Oral	acc	8	x	x		SVM
2013	Reiss et al. [29]	Physical	acc, cardio	9	x	x		Adaboost
2013	Velloso et al. [34]	Weight-lifting	acc	6	x	x		Random Forest
2012	Park et al. [25]	Pose	acc	15	x	x		SVM
2014	Gupta et al. [12]	Physical	acc	2	x			KNN
2014	Alshurafa et al. [1]	Physical	acc	8	x			Stochastic
2013	Anjum et al. [2]	Physical	acc	10	x			DT
2013	Hung et al. [17]	Social actions	acc	32	x			HMM
2013	Ladha et al. [21]	Dog activities	acc	13	x			KNN
2013	Shoaib et al. [33]	Physical	phone	4	x			KNN
2012	Kose et al. [20]	Physical	phone	5	x			KNN
2012	Wu et al. [36]	Physical	phone	16	x			KNN

Table 1. List of recent work selected to reflect the different evaluation schemes typical for recognition systems in ubiquitous computing. Physical refers to physical activities such as walking, running, or descending stairs. Acc refers to the use of accelerometer data.

of only one classifier we assume that some form of experimentation has led to that decision, even if the results are not formally reported in the respective publication.

Virtually all suitable classification algorithms are applied in ubiquitous computing. *Discriminative learning* approaches are the most common family, which include e.g. Support Vector Machines (SVM), Decision Trees (DT) and Random Forests, artificial neural networks, logistic regression, linear discriminant analysis (LDA) and boosting. They require significant computational resources during training and potentially extensive experimentation to find suitable parameter settings [4]. The most common machine learning frameworks usually provide facilities to find those parameters that maximise performance measured in CV experiments. *Generative modelling* approaches are popular when modelling higher level activities and include Hidden Markov Models (HMMs), Conditional Random Fields (CRFs), Gaussian Mixture Models (GMM), or Naive Bayes (NB). Both discriminative learning and generative modelling have been selected in work that follows (repeated) hold-out or LOSO validation schemes, as can be seen in Table 1.

Work that relies on CV reports the use of instance-based learning algorithms like k-nearest neighbour classification (KNN), which includes work that compares multiple (more sophisticated) classification approaches. For example, in [36] KNN outperforms neural networks, decision trees and other methods in CV experiments on data collected from physical activities. Similarly, [33] reports superior performance figures for KNN compared to many other methods.

In work that applies CV as well as e.g. LOSO evaluation, the performance differences between the two are often striking. For example, [34] report a performance of $> 95\%$ accuracy in CV, which drops to 78.2% in the LOSO setting. This gap between the evaluation schemes can be much higher, as e.g. [24] report a CV performance of 93.8% which drops to 59.8% in LOSO. This difference in performance between user-dependent and user-independent performance is usually attributed to variability in behaviour of the participants [7]. People perform e.g. physical activities very differently, which makes generalisation towards new users difficult. User-dependent performance on the other hand, estimated using repeated hold-out or CV schemes, is seen as an upper bound on performance, practically giving an insight into how "difficult" the problem is. It is assumed that additional recordings from more participants would alleviate this gap between user-dependent and independent performance.

Good CV performance is therefore used as a selling point of a specific technical approach in exploratory work, even if the user-independent performance may be limited (as in [24]). However, this assumption relies on the condition that the performance estimate obtained with CV is unbiased and reliable. Below we demonstrate that there are significant issues with CV, which stem from the type of data and segmentation approach applied in activity recognition systems, which are likely leading to overly optimistic performance evaluations with CV.

CROSS-VALIDATION IN SEGMENTED TIME-SERIES

Cross-validation is widely used throughout machine learning and has been demonstrated to provide reliable performance figures [19]. It has even been shown theoretically that estimates with k-fold CV are superior to hold-out validation, if the "learning problem and learning algorithm [...] is insensitive to example ordering" [5]. In other words, CV assumes that the individual samples are statistically independent of each other, so that reordering the set will have no effect on the performance of a classification algorithm. But is this the case for the data typically analysed in activity recognition?

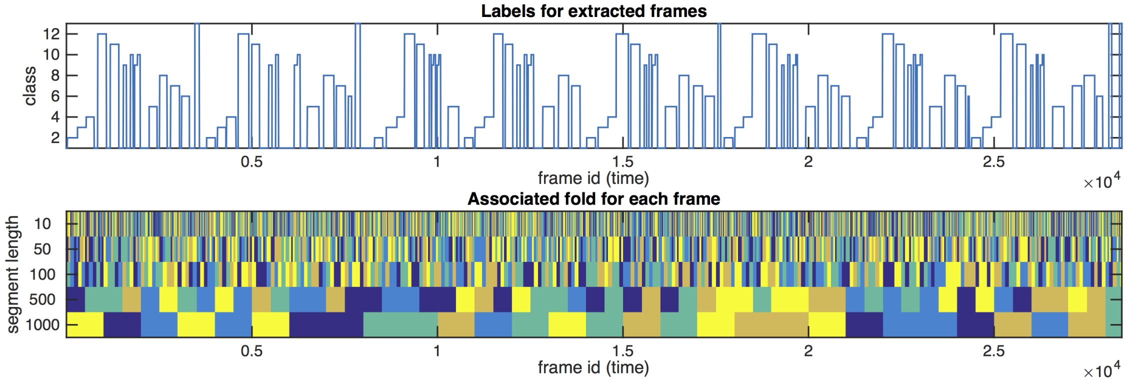

Figure 1. Associated folds for each frame in the PAMAP2 data-set for different settings of the segment length in algorithm 1. The top plot shows the labels associated to each extracted frame. The bottom plot shows the association of each frame to one of 5 folds in the cross-validation procedure outlined in Algorithm 1. See text for details [best viewed in colour].

A crucial part of the activity recognition pipeline is that of segmentation, where the sensor recordings are split into continuous segments, or frames, for further processing and analysis. There are different options for this segmentation step. Energy-based segmentation is based on simple signal characteristics and driven by empirically determined heuristics in order to estimate boundaries of segments that e.g. surround a peak in signal energy, as applied to e.g. assessment in Autism [27]. For this type of segmentation, which effectively splits the data into independent episodes based on some heuristic that are separated by non-segments, the assumption of statistical independence between segments is likely to hold.

However, the most popular segmentation approach in activity recognition in ubiquitous computing corresponds to a *sliding window segmentation* [18]: Segments (or *frames*) are extracted by "sliding" a window of fixed length over the sensor stream. Each position of the window corresponds to one segment. It is the de-facto standard that subsequent frames overlap with the last segment to a certain extent, such as 50% [24] or more, e.g. 80% in [30]. Work that extracts segments that do not overlap is rare, though an example can be found in [14]. Crucially the positioning and the size of the extracted segments are independent of the data. As adjacent frames overlap to a large extent, or in other words share many of the included measurements, can they be seen as statistically independent?

Even if consecutive frames do not overlap they are extracted "back-to-back" in sliding window procedures, without any empty spots in-between. This has implications for their (lack of) independence: activities such as walking or running are prolonged and easily span more than the frame-duration typically used in activity recognition systems (e.g. 1s). Adjacent frames extracted back to back from physical activities are therefore likely to capture the same activity, in the same context, which is likely to make them very similar.

In many, if not most application settings of activity recognition, the data does therefore *not* fulfil the requirements for reliable evaluation using cross-validation. It seems straightforward that segments that do occur close to each other in time, or those that overlap to an extent with the segment in question, will be not at all independent, but are likely to be very similar. Training and test-set in cross-validation can therefore not be seen as independent if adjacent frames are included in training and test-set at the same time.

For simple random k-fold CV, we can easily calculate the likelihood that one of the adjacent frames of frame i is in the training-set when frame i is being tested for. It is 1 minus the probability that both adjacent frames are in the same fold as frame i.

$$\begin{aligned} 1 - P(F_{i-1}, F_{i+1} \mid F_i) &= 1 - P(F_{i-1}, F_{i+1}) && (1)\\ &= 1 - P(F_{i-1})P(F_{i+1}) && (2)\\ &= 1 - \frac{1}{k^2} && (3) \end{aligned}$$

where F_i indicates that frame i is in fold F. For $k = 10$ this leads to a probability of 0.99 that surrounding frames are not in the same fold for each frame. Adjacent frames are in the training set for practically every frame extracted from the data-set. For segmentations that extract frames at an overlap of more than 50% the chance that a frame that overlaps would be in the training-set for a given frame would be significantly higher. Stratification of the frames with respect to the classes in the training set will further only have a limited effect, as subsequent frames often belong to the same class and are therefore likely to end up in different CV folds (as stratification aims to distribute frames of a class evenly between folds). The similarity between adjacent frames appears to be a significant issue and it is likely that approaches to feature extraction and classification that retain this similarity will easily (but maybe erroneously) outperform other methods in this evaluation scheme.

So far, this issue has not been explored in the field of ubiquitous computing. If this *neighbourhood-bias* affects cross-validation, we expect the apparent recognition performance in cross-validation:

- to decrease significantly if this bias is removed,
- to decrease independently of the preprocessing techniques,

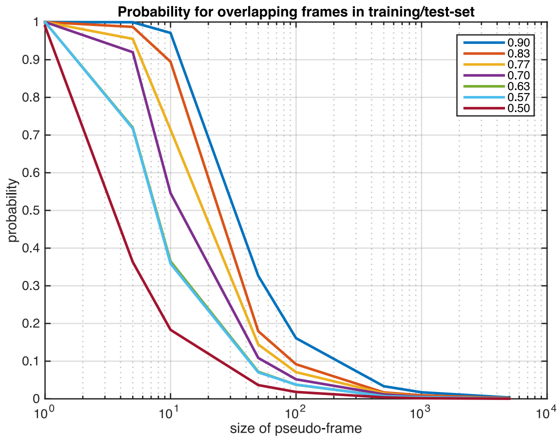

Figure 2. Likelihood of overlapping frames to be included in the same fold for different settings of the segment length in algorithm 1, for different degrees of overlap (see legend), in the PAMAP2 data-set. A segment length of 100 already reduced the chance to less than 10% for the most common degrees of overlap, which should be effective at alleviating the bias discussed in this work.

- to decrease independently of the classification approach,
- we expect the decrease in performance to be related to the similarity between extracted segments specific to the type of physical activity in each data-set.

In order to test these hypotheses we develop a novel approach to cross-validation which allows us to limit the effect of the neighbourhood-bias: Meta-segmented Cross-Validation, and evaluate it empirically on two data-sets typical for activity recognition.

META-SEGMENTED CROSS-VALIDATION

The simplest way to alleviate the bias outlined above is to follow hold-out or LOSO evaluation schemes. That is, however, not always possible as was discussed above. In this section we develop a novel approach to cross-validation that retains the overall nature of cross-validation in providing a user-dependent estimate of system performance. However, we alter the standard approach to alleviate the neighbourhood bias we find in random (stratified) CV.

Intuitively the approach can be summarised as follows: we avoid placing adjacent frames in different folds. We alter regular CV by introducing a higher level meta-segmentation on the level of extracted frames. We group adjacent frames into meta-segments of length d that do not overlap. Then we perform stratified CV on this set of meta segments. Algorithm 1 outlines the approach[1]: we first extract meta-segments and estimate the distribution of labels with some added Gaussian noise for randomness. We retain the label distributions in a matrix with one row for each meta-segment and one column for each class in the data-set. In order to estimate folds that are stratified with respect to the contained classes we need to order the matrix rows such that consecutive rows are similarly distributed. We apply lexicographic sort, to order the rows as if the unique entries in each column were letters in the alphabet. Next, we sequentially assign fold-ids to each row of the matrix in a circular manner using the modulo operator (see Algorithm 1). Rows that obtain fold-id k belong to the test-set of fold k. The procedure can be implemented efficiently and introduces a single new parameter: the length d of the meta-segments used in the procedure.

Figure 1 illustrates how frames, extracted from the PAMAP2 data-set, are split into stratified folds based on the assigned label and a number of different settings for the meta-segment length in Algorithm 1. The top graph shows the labels of each frame. The bottom graph illustrates the colour-coded association of each frame to one of five folds. The quality of the stratification in the proposed approach is explored towards the end of this work.

The chance for overlapping or adjacent frames to be included in the same fold decreases with increasing length of the meta-segments. This is illustrated in Figure 2. For the most common 50% overlap of subsequent frames the chance of including overlapping frames in the same fold drops rapidly towards 0 with increasing meta-segment length. The more frames overlap, the more we have to increase the meta-segment length to obtain similar results.

Decreasing overlap

An intuitive approach to reduce the neighbourhood bias is to reduce the size of the neighbourhood of each frame. This could be easily achieved by reducing the overlap between subsequent frames in the sliding window procedure, or to even remove the overlap completely (like in [14]). There are two reasons why we opted to keep the overlap fixed in this work: The overlap may be a requirement imposed from the outside, e.g. to allow responsiveness in a real-time application setting, where a system is bound to produce hypotheses at a

Algorithm 1 Calculate folds for segmented cross-validation

```
Input: frameLabels $\in \mathbb{N}^d$, numFolds, segLength $\in \mathbb{N}$
Output: frameFoldIds $\in \mathbb{N}^d$

  classes = $\mathbf{unique}$(frameLabels)
  numSegments = $\mathbf{length}$(frameLabels)/segLength
  segDist = $\mathbf{zeros}$(numSegments, $\mathbf{length}$(classes))

  for $i = 1$ to numSegments do
      # get distribution of labels within frame
      segDist$(i, :)$ = $\mathbf{getLabelDist}$(frameLabels $\in$ segment$_i$)
      # add noise for randomness
      segDist$(i, :)$ = segDist$(i, :)$ + $\mathbf{randn}() * 0.01$
  end for

  # get sorted list of indices (lexicographic sort of distributions)
  indices = $\mathbf{lexisort}$(segDist)
  # assign fold to each segment
  foldIds(indices) = $1 + \mathbf{mod}(\{1 \ldots$ numSegments$\}$, numFolds)

  for $i = 1$ to numSegments do
      # assign each frame within segment to fold
      frameFoldIds $\in$ segment$_i$ = foldIds$(i)$
  end for

  return frameFoldIds
```

[1] Source-code available at http://openlab.ncl.ac.uk/harcrossval

frequency higher than the frame duration. Secondly, a smaller overlap would decrease the number of frames extracted from a data-set and may make some classes appear less frequent or for a shorter duration, which may confound performance comparison.

EVALUATION

In order to study the bias that may result from cross-validation we developed an approach that allows us to reduce the effect of the neighbourhood of a frame in a controlled manner. We evaluate the effect of the neighbourhood bias in a number of different feature extraction and classification approaches on two publicly available data-sets typical for activity recognition in ubiquitous computing.

Data-sets

We selected two data-sets that are representative for the different applications of body-worn sensing in ubiquitous computing, yet differ significantly in the type of activities captured from the participants.

PAMAP2

Contains recordings from 9 individuals that follow a physical activity drill, in which they engage in a dozen different activities which include walking, running, rope-jumping, lying, among others [30]. The movement is captured with three inertial measurement units (IMUs) and an additional heart-rate monitor on the chest. The IMUs are worn on the wrist, chest, and ankle and record at a temporal resolution of 100Hz. To limit the dimensionality of this data-set we avoid the use of the heart-rate monitor, and focus solely on the raw tri-axial accelerometer data collected by the IMUs. For experiments on PAMAP2 we replicate the segmentation from the original work and extract segments that span 512 samples (5.12s) with a step-size between segments of 100 samples (1s).

Opportunity

Contains recordings from 4 participants in a sensor-rich (kitchen) environment [32]. Each subject engages in multiple "ADL runs", where they follow a loose activity protocol at their own pace, and a "Drill run" where each activity is repeated multiple times. Sensing includes object sensors, ambient sensors, and body-worn sensors. Each participant is equipped with 7 IMUs and 12 accelerometers at different places on the body and sample at a resolution of 30Hz. In this work we selected 3 tri-axial accelerometers worn on the arm of the subject to allow comparison to the PAMAP2 data-set in terms of input dimensionality. Further we select data from the drill run by each subject to replicate studies that follow a strict protocol such as PAMAP2. From this data we extract segments that span 30 samples (1s) with a step-size of 15 samples between segments.

Feature extraction

Three different feature extraction approaches are applied in this work. They were selected as they often show good performance in work that relies on CV experiments. They are applied "as is" to the frames extracted using a sliding window procedure.

ECDF features (ECDF)

These features were presented in [15] and aim to preserve crucial aspects of the underlying distribution of data within each frame in a compact representation. Basically they correspond to concatenated quantile functions for each axis of recording, e.g. the three axis of recording in an accelerometer. They have shown particularly good performance in stratified CV on a number of different data-sets (including the data-sets in this work). For all experiments we chose to set the feature parameter (number of interpolation points), to 15, which has shown good performance across different settings [15]. For each frame we extract 144 features using ECDF.

Fourier coefficients (FFT)

Particularly the PAMAP2 data-set contains repetitive behaviour such as walking and running. Fourier coefficients should be able to capture this repetitive nature and therefore constitute a suitable feature representation. FFTs are computationally efficient to extract and often used as a component in feature extraction approaches [9]. In order to remain comparable to the other features extracted in this work, we chose to include the first 15 Fourier coefficients for each sensing axis. For each frame we extract 135 features using FFT.

Statistical features (STAT)

For many applications, hand-crafted sets of statistical features such as mean, variance or momentum outperform other representations. We chose to replicate the feature representation from [28] which corresponds to 23 extracted features for each sensor used in the recordings. The STAT features include mean, standard deviation, energy, entropy, correlations, pitch, and roll. For each frame we extract 69 features using STAT.

Classification approaches

We evaluate common types of classification approaches from related work. Parameter tuning is kept to a minimum to simulate off-the-shelf use of each recognition approach.

k-Nearest Neighbour (KNN)

From the family of instance-based learning we apply k-nearest neighbour, with k set to 3. Before applying KNN we transform the features extracted from the frames to have zero mean and unit variance, using the mean and standard deviation estimated on each respective training set in CV.

Decision Trees (C4.5)

For each cross-validation fold we estimate a binary decision tree. Decision trees are popular in activity recognition, e.g. for the recognition of physical activities [2]. Decision trees are used in embedded settings due to their favourable computational properties at inference time.

Support Vector Machines (SVM)

We apply support vector machines with default parameters from libsvm [8] with RBF-kernel, a cost parameter of 1 and the kernel scaling parameter set to $1/\#$features. We chose not to optimise the hyper-parameters of the SVM for each specific experiment as it is computationally very intensive on one hand, and on the other hand to replicate off-the-shelf use of the approach in many applications.

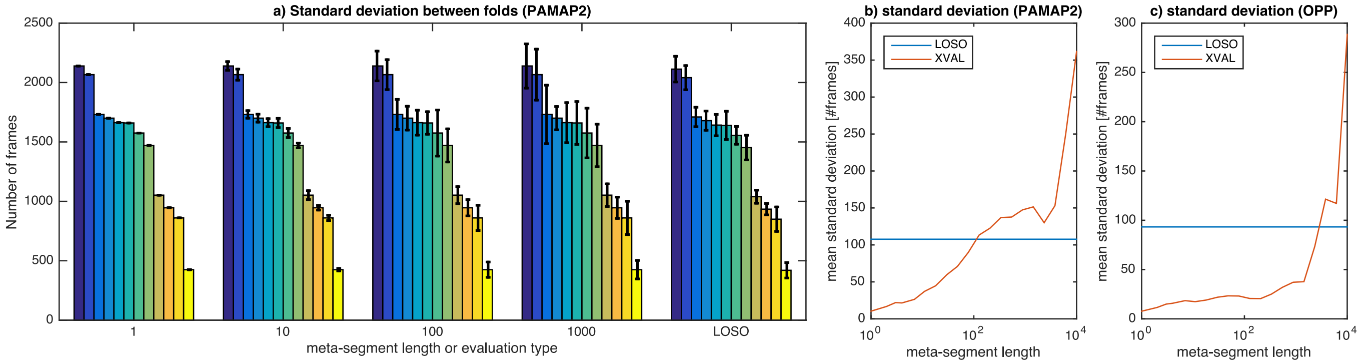

Figure 3. Stratification in the proposed approach on PAMAP2 and Opportunity (OPP) compared to standard CV and LOSO. a) Mean distribution of all objective classes in PAMAP2 (ordered by number of frames) across 10 training sets; error bars indicate one standard deviation. b) Standard deviation as a function of meta-segment length for PAMAP2; c) for OPP.

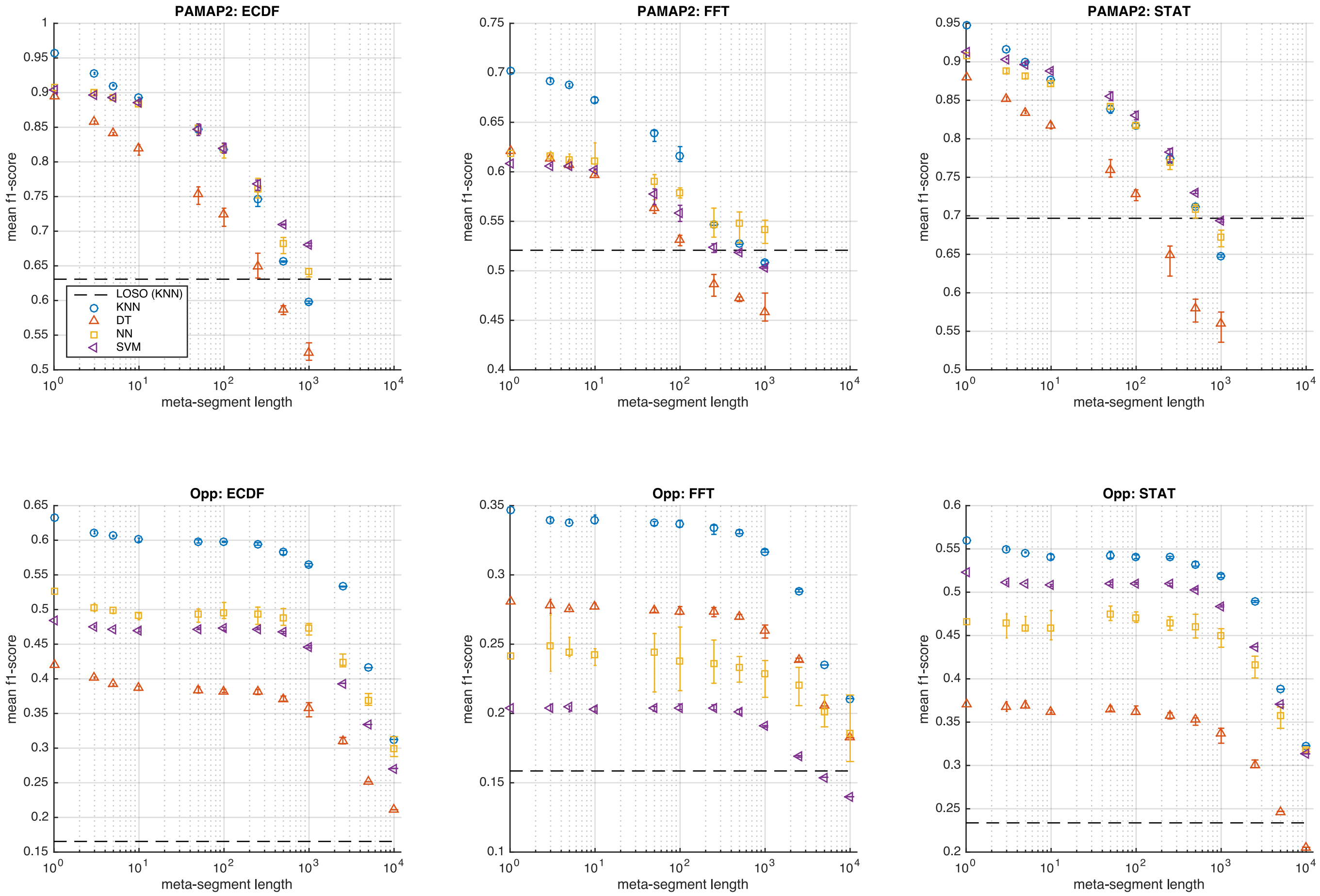

Figure 4. Prediction performance decreases for increasing meta-segment length in the proposed CV approach for different features and classifier combinations on a variety of data-sets. The top row shows results for systems trained on PAMAP2, the bottom row for Opportunity (Opp). The meta-segment length is increased on a logarithmic scale. A meta-segment length of 1 (10^0) indicates standard CV. Dashed black lines indicate LOSO performance in the respective setting. Error-bars indicate the minimum and maximum performance obtained across 5 repetitions of each experiment.

Feed-forward Neural Networks (NN)

The last approach we utilise is feed forward neural networks. We train networks with a single hidden layer that contains 100 hidden (sigmoid) units and one unit in the output layer for each activity class, arranged in a softmax group. Input data is scaled to have zero mean and unit variance. Training is performed using scaled conjugate gradients and early stopping [4].

Performance metric

As the activities of interest are not evenly distributed in both data-sets we opted for the mean f1-score as performance criterion. It is defined as the average geometric mean of precision and recall for each class:

$$f_1 = \frac{2}{c}\sum_{i=1}^{c}\frac{\text{prec}_i \times \text{recall}_i}{\text{prec}_i + \text{recall}_i}, \tag{4}$$

where $\texttt{prec}_i$ refers to the precision for class i, and c is the number of classes in the data-set.

RESULTS

Stratification in the proposed approach

The approach proposed in this work relies on a meta-segmentation where frames extracted from time-series data are grouped prior to the assignment to folds for training and testing. The approach was designed to produce folds that are stratified with respect to the classes of interest, as an unbalanced training set may have significant effect on classification performance of a recognition system. Figure 3 illustrates the quality of the stratification for a variety of meta-segment lengths in comparison to standard (stratified) CV (10 folds) and LOSO on extracted frames. The histograms in Figure 3 (a) show that for short meta-segments there is virtually no difference between the stratification in standard CV compared to the proposed approach in PAMAP2. For larger values (e.g. 1000) the standard deviation is larger than the one we observe for LOSO evaluation. This is further illustrated in Figure 3 (b) for PAMAP2 and (c) for Opportunity. We observe a similar standard deviation compared to LOSO at a meta-segment length of approximately 100 for PAMAP2 and approximately 2500 for Opportunity.

Effect on recognition performance

For each combination of feature representation, data-set and classification approach we ran a number of different experiments, whose results are reported in Figure 4. We increase the meta-segment length in the proposed CV approach (10 folds) from 1 (corresponding to standard CV) up to 10^4 (10^3 for PAMAP2) on a logarithmic scale. Dashed black lines show the LOSO performance of KNN for each setting. Markers indicate mean, error-bars the minimum and maximum performance across 5 repetitions. In each case, performance is estimated on the confusion matrix containing the predictions from all folds (or subjects in LOSO). This avoids issues in estimation of e.g. standard deviation if some classes are not included in each test-set.

The effect of an increase in meta-segment length on the recognition results is immediately apparent, but quite different for the two data-sets investigated here. In both cases we see a monotonic decrease in performance which approximates the LOSO performance for long meta-segments. However, the decrease is much more pronounced on PAMAP2. Where on Opportunity the performance largely remains constant up to a segment length of 100 we see a drop of more than 10% mean f1-score on PAMAP2. The decrease in recognition performance for both data-sets behaves similarly to the increase in standard deviation reported in Figure 3 (b) and (c). If we compare the recognition performance where the standard deviations are close to those observed in LOSO validation we see that the performance is still significantly different: At a meta-segment length of 100 we see a mean f1-score on PAMAP2 of 0.82 for KNN and ECDF features, which significantly better than the LOSO performance of 0.63. Similarly for Opportunity we see a performance of 0.53 at a meta-segment length of 2500, which is much better than the (poor) LOSO performance of 0.17. We believe that this performance gap, which is ultimately related to the *composition* of each training-set, illustrates the difference between the user-dependent and user-independent setting.

Beyond the impact of user-dependency and stratification on the recognition performance there appears to be an additional effect: Irrespective of the feature extraction or data-set we see that KNN outperforms all other methods by a large margin. Consider the results of ECDF on PAMAP2 (top left plot in Figure 4). The standard CV performance (meta-segment length of $1 = 10^0$) of DT, NN, and SVM, all fit naturally to an extrapolation of the results for longer meta-segments. KNN performance, however, appears to gain an advantage over the other methods when the meta-segment length approximates 1. We observe a similar behaviour for STAT on PAMAP2. Both ECDF and STAT capture statistical characteristics, which retain much of the similarity of subsequent frames if there is sufficient overlap in the segmentation. The advantage of KNN in regular CV therefore appears to be only an artefact, which, however, explains why it is probably the most popular classification approach for work relying on standard CV (see table 1). For longer meta-segments, more sophisticated approaches like SVMs or NNs show significantly better performance than KNN (e.g. on STAT/PAMAP2 for segment length ≥ 10), which is in line with results from other fields that apply machine learning, where KNN usually acts as a mere baseline for recognition performance.

Effect depending on activity type

Overall the two data-sets seem to be affected very differently by the proposed CV approach. Where on PAMAP2 the drop in performance is immense and relative performance of classification methods change, there is only a much smaller drop in performance on Opportunity for reasonably short meta-segments (e.g. below 1000). We believe this difference stems from the nature of the activities included in each data-set. Intuitively, activities that are very self-similar and occur over long periods (e.g. walking, running) would benefit most from adjacent frames, as they are inherently very similar. Activities that are short and rarely span more than one extracted frame of data (e.g. 1s), would benefit much less, as adjacent frames are very dissimilar to begin with.

In order to test this hypothesis, which is in line with our previously stated objectives, we estimated the intra-class distance between frames for each objective class in PAMAP2 and Opportunity. We do so on frames of raw recordings by estimating the mean difference in distribution for each sensing axis using the kolmogorov-smirnov test-statistic:

$$D_{ij} = \frac{1}{d}\sum_{a=1}^{d}\sup_x |C_{i,a}(x) - C_{j,a}(x)| \tag{5}$$

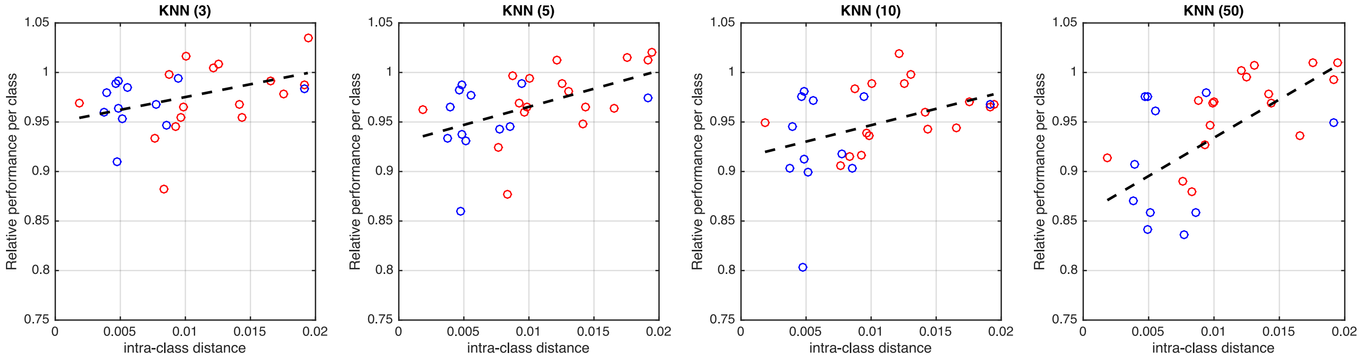

Figure 5. Relative performance between regular CV and meta-segmented CV for all objective classes in PAMAP2 (blue) and Opportunity (red), plotted against the intra-class distance, for a variety of meta-segment lengths (see plot title). Low intra-class distance corresponds to high self-similarity of simple repetitive physical activities such as walking. Opportunity mostly contains manipulative gestures (red) with low self-similarity, and suffers less decline in performance.

Where D_{ij} is the distance between frame i and frame j, d is the number of sensing axis within each frame, and $C_{i,a}$ is the cumulative distribution for axis a in frame i. The result of this procedure is one measure of intra-class similarity for each class of activity which reflects how similar adjacent frames of the same class are in each data-set.

According to our considerations about the neighbourhood bias in CV we should see a larger drop in performance for classes that are self-similar, because adjacent frames will contain data that is very similar. The relationship between intra-class distance and drop in performance between regular- and segmented cross-validation is illustrated in Figure 5.

The results seem to confirm our hypothesis. Activities with high self-similarity (low intra-class distance) in the PAMAP2 data-set (blue) show a significant drop in performance, while performance is largely retained for classes with low self-similarity in Opportunity (red). The drop in performance increases with increasing meta-segment length, even though the training-sets only show a small deviation in class distribution (see Figure 3). Intra-class distance is certainly not the only factor affecting this drop in performance, as it depends on the feature extraction to abstract from the raw recordings and on how well the classifier can model the boundaries between objective classes. There nevertheless seems to be a rather strong relationship, which is largely independent of the classification approach.

Meta-segment length

The cross-validation proposed in this work relies on an additional parameter: the meta-segment length. We have evaluated a large number of different settings and observed a monotonic decrease in recognition performance with increasing segment length. However, the choice of meta-segment length remains an open issue. It seems intuitive that a minimum meta-segment length could be obtained through the analysis of similarity between adjacent frames, in order to find the natural size of the neighbourhood of a frame or even of a single sample. However, so far we have not obtained sufficient insights that would enable an a-priori selection of the "correct" segment length for a specific data-set, which will be addressed in future work.

One possible avenue to obtain a meta-segmentation level for a particular data-set could be the insights obtained in Figure 3. By selecting a meta-segment length that shows a similar standard deviation compared to LOSO validation we can obtain what we believe to be a lower bound on user-dependent performance. This is, however, not suitable for use during early exploration where data from only one or a few users is accessible.

For practical applications of the proposed CV scheme the meta-parameter should be included into the optimisation procedure. Note that this is not a practical limitation per se as we propose using the new evaluation scheme already during system design. Very short meta-segments will already alleviate much of the neighbourhood bias discussed in this work. More realistic performance measures will lead to more robust recognition systems, even if the meta-parameter has not been chosen according to its theoretical optimum.

DISCUSSION

Random cross-validation for performance evaluation in activity recognition may provide overly optimistic performance figures. The empirical evaluation presented in this work confirms our concerns regarding the construction of the folds used for repeated evaluation. Cross validation requires that the learning problem is insensitive to the ordering of the input samples [5], and it is doubtful whether the type of data analysed in activity recognition fulfils this criterion. Adjacent frames may overlap to a large extent, are likely to record the same activity in the same context and are therefore intimately linked. This leads to a bias, as it is practically guaranteed that adjacent frames are included in the training set if we test for a specific frame.

Not all application settings are impacted equally by this *neighbourhood bias*. For activities with limited duration which are non-repetitive, we see significant changes in recognition performance if the bias is removed. However, the relative performance of classification and feature extraction approaches remains similar (see bottom row in Figure 4). For prolonged, repetitive behaviour such as walking or running the effect of the bias is much more significant. The recognition performance drops by 15% mean f1-score or more if we avoid this neighbourhood bias on PAMAP2 (see top row in Figure 4). Further it appears that for some applications the choice of feature representation and classification approach depends more on the evaluation approach than was previously anticipated, as the relative performance of classifiers changes if the neighbourhood bias is alleviated.

This is an important issue, as this application setting (recognition of repetitive physical activities) is seen as "easy" or trivial in the community. With the field of ubiquitous computing maturing it is those "low-hanging fruits" that are the first to be adopted by other fields, such as medical engineering. Overly optimistic performance estimates in regular cross-validation may motivate significant investments (e.g. into clinical trials), and a drop in performance of the magnitude observed in this work could significantly impair practical deployments.

Overall we demonstrated in this work that the performance difference between user-dependent and user-independent settings in activity recognition does not only stem from the differences in behaviour of the participants. It may further be influenced significantly by the evaluation approach. We hope that this work motivates practitioners in ubiquitous computing, and people that apply techniques from activity recognition in other fields, to rethink their evaluation strategy if they rely on cross-validation. This includes the authors themselves, as our own work was likely also affected by this bias in the past.

It is, however, not our intention to substitute for LOSO validation approaches with our proposed approach. We believe that LOSO gives the best insights into the performance of systems deployed in naturalistic conditions, which is a major goal in ubiquitous computing. Instead we aim to substitute regular CV approaches used in common tools and frameworks for parameter tuning of e.g. SVMs or for the training of neural networks, with a validation approach that is convenient to use but less affected by the neighbourhood bias discussed in this work.

REFERENCES

1. Alshurafa, N., Xu, W., Liu, J. J., Huang, M.-C., Mortazavi, B., Roberts, C. K., and Sarrafzadeh, M. Designing a robust activity recognition framework for health and exergaming using wearable sensors. *Biomedical and Health Informatics, IEEE Journal of 18*, 5 (2014), 1636–1646.
2. Anjum, A., and Ilyas, M. U. Activity recognition using smartphone sensors. In *Consumer Communications and Networking Conference (CCNC), 2013 IEEE*, IEEE (2013), 914–919.
3. Berlin, E., and Van Laerhoven, K. Detecting leisure activities with dense motif discovery. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM (2012), 250–259.
4. Bishop, C. M., et al. *Pattern recognition and machine learning*, vol. 4. springer New York, 2006.
5. Blum, A., Kalai, A., and Langford, J. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, ACM (1999), 203–208.
6. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., and Pentland, A. S. Pervasive stress recognition for sustainable living. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, IEEE (2014), 345–350.
7. Bulling, A., Blanke, U., and Schiele, B. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR) 46*, 3 (2014), 33.
8. Chang, C.-C., and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3 (2011), 27.
9. Figo, D., Diniz, P. C., Ferreira, D. R., and Cardoso, J. M. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing 14*, 7 (2010), 645–662.
10. Gjoreski, H., Kaluža, B., Gams, M., Milić, R., and Luštrek, M. Ensembles of multiple sensors for human energy expenditure estimation. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM (2013), 359–362.
11. Gu, W., Yang, Z., Shangguan, L., Sun, W., Jin, K., and Liu, Y. Intelligent sleep stage mining service with smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM (2014), 649–660.
12. Gupta, P., and Dallas, T. Feature selection and activity recognition system using a single tri-axial accelerometer.
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.
14. Hammerla, N. Y., Fisher, J. M., Andras, P., Rochester, L., Walker, R., and Plötz, T. Pd disease state assessment in naturalistic environments using deep learning.
15. Hammerla, N. Y., Kirkham, R., Andras, P., and Ploetz, T. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ACM (2013), 65–68.
16. Hirano, T., and Maekawa, T. A hybrid unsupervised/supervised model for group activity recognition. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ACM (2013), 21–24.
17. Hung, H., Englebienne, G., and Kools, J. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM (2013), 207–210.

18. Keogh, E., Chu, S., Hart, D., and Pazzani, M. Segmenting time series: A survey and novel approach. *Data mining in time series databases 57* (2004), 1–22.

19. Kim, J.-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis 53*, 11 (2009), 3735–3745.

20. Kose, M., Incel, O. D., and Ersoy, C. Online human activity recognition on smart phones. In *Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data* (2012), 11–15.

21. Ladha, C., Hammerla, N., Hughes, E., Olivier, P., and Plötz, T. Dog's life: wearable activity recognition for dogs. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM (2013), 415–418.

22. Ladha, C., Hammerla, N. Y., Olivier, P., and Plötz, T. Climbax: Skill assessment for climbing enthusiasts. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM (2013), 235–244.

23. Lane, N. D., Pengyu, L., Zhou, L., and Zhao, F. Connecting personal-scale sensing and networked community behavior to infer human activities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM (2014), 595–606.

24. Li, C.-Y., Chen, Y.-C., Chen, W.-J., Huang, P., and Chu, H.-h. Sensor-embedded teeth for oral activity recognition. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ACM (2013), 41–44.

25. Park, J.-g., Patel, A., Curtis, D., Teller, S., and Ledlie, J. Online pose classification and walking speed estimation using handheld devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM (2012), 113–122.

26. Peterek, T., Penhaker, M., Gajdoš, P., and Dohnálek, P. Comparison of classification algorithms for physical activity recognition. In *Innovations in Bio-inspired Computing and Applications*. Springer, 2014, 123–131.

27. Plötz, T., Hammerla, N. Y., Rozga, A., Reavis, A., Call, N., and Abowd, G. D. Automatic assessment of problem behavior in individuals with developmental disabilities. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM (2012), 391–400.

28. Plötz, T., Moynihan, P., Pham, C., and Olivier, P. Activity recognition and healthier food preparation. In *Activity Recognition in Pervasive Intelligent Environments*. Springer, 2011, 313–329.

29. Reiss, A., Hendeby, G., and Stricker, D. Confidence-based multiclass adaboost for physical activity monitoring. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ACM (2013), 13–20.

30. Reiss, A., and Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, IEEE (2012), 108–109.

31. Reiss, A., and Stricker, D. Personalized mobile physical activity recognition. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ACM (2013), 25–28.

32. Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Forster, K., Troster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., et al. Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, IEEE (2010), 233–240.

33. Shoaib, M., Scholten, H., and Havinga, P. J. Towards physical activity recognition using smartphone sensors. In *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*, IEEE (2013), 80–87.

34. Velloso, E., Bulling, A., Gellersen, H., Ugulino, W., and Fuks, H. Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference*, ACM (2013), 116–123.

35. Ward, J. A., Lukowicz, P., and Gellersen, H. W. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 1 (2011), 6.

36. Wu, W., Dasgupta, S., Ramirez, E. E., Peterson, C., and Norman, G. J. Classification accuracies of physical activities using smartphone motion sensors. *Journal of medical Internet research 14*, 5 (2012), e130.