Toward Personalized Activity Recognition Systems With a Semipopulation Approach

Article *in* IEEE Transactions on Human-Machine Systems · November 2015

DOI: 10.1109/THMS.2015.2489688

Julian Andres Ramos Rojas
Carnegie Mellon University

Anind K. Dey
Carnegie Mellon University

Manuscript received

Towards Personalized Activity Recognition Systems with a Semi-population Approach

Jin-Hyuk Hong
Carnegie Mellon University
hjinh7@gmail.com

Julian Ramos
Carnegie Mellon University
ingenia@andrew.cmu.edu

Anind K Dey
Carnegie Mellon University
anind@cs.cmu.edu

ABSTRACT

Activity recognition is a key component of context-aware computing to support people's physical activity, but conventional approaches often lack in their generalizability and scalability due to problems of: diversity in how individuals perform activities; overfitting when building activity models; and collection of a large amount of labeled data from end users. To address these limitations, we propose a semi-population-based approach that exploits activity models trained from other users so a new user does not need to provide a large volume of labeled activity data. Instead of relying on any additional information from users like their weight or height, our approach directly measures the fitness of others' models on a small amount of labeled data collected from the new user. With these shared activity models among users, we compose a hybrid model of Bayesian networks and support vector machines to accurately recognize the activity of the new user. On activity data collected from 28 people with a diversity in gender, age, weight and height, our approach produced an average accuracy of 83.4% (kappa: 0.852), compared to individual and (standard) population models that had accuracies of 77.3% (kappa: 0.79) and 77.4% (kappa: 0.743), respectively. Through an analysis on the performance of our approach and users' demographic information, our approach outperforms others that rely on users' demographic information for recognizing their activities, which may contradict the commonly held belief that physically similar people would have similar activity patterns.

I. INTRODUCTION

The ability to recognize one's physical activity has become a key component of context-aware computing. This can be used to support people in being more physically fit and healthier, as sensed in real-world settings. This is especially true for medical, military, and Activities of Daily Living (ADLs)-based applications [12, 15, 26, 27, 30, 43]. Advances in mobile sensing technology have led to the development of a number of human-activity recognition methods and systems, mostly using accelerometers [4, 11, 18, 22, 23, 47]. Despite research on sensor-based activity recognition, little work has been concerned with the applicability to users (and their environments) that are previously unknown to the system [11]. For a more practical activity recognition system that provides good performance and is easily deployable in the field, it is important to address how to build activity models for a new user. However, few research efforts have focused on methods for building an activity recognition system that will work in real-world environments [37, 41], and many current solutions are limited in their ability to generalize and scale to new users and/or new environments, requiring considerable effort and customization to achieve good performance [11, 28, 45].

One approach for building activity models is a single subject-independent, or *population*, model: a one-fits-all activity model that merges all the data from the available training population. It may be an ideal solution if we could obtain accurate models, but population models often exhibit weak performance when applied to test data for new users due to the high variation in how different people perform a number of different activities [37]. People are idiosyncratic and have their own particular styles or ability to perform activities. A single individual may even have a large variation when performing the same activity in different situations [11, 41]. The complexity of population models is usually very high.

To address the limitations of population models, researchers have investigated a personalization technique to build a subject-dependent, or *individual*, activity model for each user. When building such a model, there is an important choice to be made about how much data to collect for a new user. Individual models typically rely on the collection and annotation of large amounts of user-specific training data, which are often very burdensome to end users and limit the acceptance of such activity recognition systems [37, 38]. In contrast, trying to model a user's activity from a small amount of training data results in a limited ability to handle variations in how this user performs activities and is prone to overfitting [36, 44, 45].

An alternative approach is to use *calibration*. In calibration, a population or individual model is trained for a pool of existing subjects. Then, for each new user, this model is calibrated to the new user using the training data collected from her. Since it is very difficult to design an effective calibration method, it has been applied only to adjust for variations in raw sensor data caused by different sensor placements or orientations within a single user [7, 16], and not been applied across users.

To build an activity recognition model for a new user, in this paper, we present a novel calibration method that does not suffer from the problems of diversity in how individuals perform activities, of overfitting, and of needing to collect a large amount of labeled data from the new user. Our method avoids these problems by using a pool of activity models trained from existing subjects. In common situations where a new user may not be able to provide a large volume of labeled activity data, our proposed method uses a small set of labeled data collected from a new user and measures the performance (or fitness) of activity models trained from other people in the pool. The best performing activity recognition models on the new user's labeled data are aggregated to construct a hybrid activity model.

II. BUILDING AND EVALUATING SENSOR-BASED ACTIVITY MODELS

In the activity recognition community, there is an on-going open debate about the relative merits of subject-dependent (or individual) models and subject-independent (or population) models [26]. Some researchers believe that an activity model has to be personalized to an individual since different people have different styles of performing activities. Therefore, a new activity model has to be trained for a new user who has to provide labeled data. Other researchers argue that the activity model has to be flexible enough to work well on different users, where this activity model is trained on data from a number of users and applied to a new user without additional processing. From the perspective of machine learning, each approach has different issues to address for optimizing performance [13]. Population models suffer from problems of high variance in activity performance, while individual models suffer from problems of high bias. Both situations are to be avoided as they can contribute to poor performance [17]. A one-fits-all (*i.e.*, population) model is preferred as it does not require the collection of labeled data for each new user, but it is challenging to obtain a high-performing model in the field.

Some researchers have suggested new approaches for reducing the effort required for collecting and annotating training data [38, 44]. These approaches train models using a small amount of labeled data from a user. Zhang *et al*. focused on the use of incomplete data obtained from multiple inhabitants in a smart home environment. With a probabilistic activity learning algorithm using maximum-likelihood estimation, they learn the activity models directly from both complete and incomplete data [44]. However, their method relies on contact sensors instead of accelerometers so the type of activities their technique could recognize is limited. Also, they evaluated their method on synthetic data generated from 47 real labeled records, which may not sufficiently capture the variations in how activities are performed by users in the field. Stikic *et al*. studied annotation strategies to reduce the required amount of annotation. Two methods, multi-instance learning and graph-based label propagation, were explored to leverage sparsely labeled data with unlabeled data [38]. Despite the effort to reduce the annotation cost, their method is intended to recognize long-lived or very frequent activities, making it difficult to generate labels for short-lived or rarer activities. Also, they rely on assumptions such as that the sensor data for the same activity are similar when mapped in feature space, which allows the technique to propagate labels to unlabeled data; however this assumption depends on the variations in which the activity is performed, and on the nature of the activity itself.

Researchers have focused on building a good one-fits-all activity model. Riboni and Bettini exploited ontological reasoning combined with statistical inference [36]. In their method, statistical inference is used to recognize activities from raw sensor data, and then symbolic reasoning is applied to refine the results of the statistical inference based on a user's current context. By separating their recognition process into these two components, the statistical inference of their activity models is only asked to manage a limited amount of raw sensor data. It does so by constraining what activities can occur in which locations, *e.g.*, a user brushes his teeth in the kitchen or rest room so the brushTeeth model only deals with data collected in those places. Dalton and OLaighin compared individual models and population models by using standard machine learning methods [14]. Contrary to other findings, they argued that the population models could outperform the individual models if there were enough users for training the models. Still, the users needed for model training have to be quite diverse in order to get a similar result in the field.

Some approaches personalize one's activity models by incorporating them with data from others with similar characteristics. Maekawa *et al*. introduced the concept of supportive users, in which they trained activity models by including a new user's and others' data [29]. While this approach is very interesting, they did not show that the technique could generalize, with only a small performance improvement (1-2% accuracy improvement) and a limited population (three participants). Sun *et al*. personalized activity models by using online multitask learning, where similarities of data among users are calculated and training data for a new user are reconstructed from others' data based on the similarities [39]. They demonstrated an accuracy improvement of 1-4%, but their model complexity increases exponentially as more people and activities are considered. Lara and Labrador suggested a group-specific model, in which a population model is built from a subpopulation in which individuals have similar characteristics in age, gender, weight and height, *e.g.*, an activity model for tall, young males [26]. While a novel idea, this suggestion needs to be validated.

Zhao *et al*. present a transfer learning technique in which the model of a new user is adapted from the model of an existing user [46]. The advantage of the technique is that it boosts accuracy without requiring labeled data for the new user. It uses multiple iterations of decision trees and *k*-means clustering. Our approach is complementary, and uses the data for a new user to select existing models to use, rather than training a new model(s).

In the domain of social networking and crowd sourcing, Lane *et al*. [25] have shown that the physical conditions or lifestyles of users can be used to help identify activity (mobility) models for new users that share those characteristics. These characteristics are used to build similarity networks over the users. A classifier is trained over each network, and a personalized model is built for each user based on a combination of the individual network classifiers. This increases the amount of training data available, but assumes that the best models can be built from collections of similar users, based on observable characteristics. Our work builds upon this idea in a general sense, but we focus on the fitness of activity models (*i.e.*, how well the models perform) for similar users rather than the similarity of users' physical characteristics. Similar to our approach, Reiss and Stricker present a system that uses an already trained set of activity classifiers, and uses labeled data from a new user to learn weights for each classifier in a weighted majority vote [34]. Conceptually, we share the same idea with Reiss and Stricker as utilizing models built from other datasets for a new user. Technically, however, we approach the modeling of activities individually from many different users based on the proposed hybrid activity models, while Reiss and Stricker applied the weighted majority voting, in which an expert classifier is built for a *single* user to classify all his activities. Some activities of different users may be similar but their performing of other activities could be very different. By building a classifier for a *single* activity of a user, we model his performing the activity more precisely and flexibly, and our method should be able to better optimize activity models by selecting partial models from *different* users who are the most similar to a new user in performing the corresponding activities. Additionally, our approach has several potential advantages in scalability. When it introduces a new activity into the system, Reiss and Stricker's approach has to re-build the classifier but our approach only needs to build a model for the new activity. The more people are available in the pool, the more weights have to be retrained with their method. However, the complexity of a final model of our method for a new user is constant regardless of the number of people in the pool.

In addition, the evaluation of the performance of activity recognition systems can also depend on how activity models are built, *i.e.*, the evaluation of individual models rely on one's own data while the leave-one subject-out validation is a common evaluation strategy for population models [36, 45]. For a more accurate evaluation of population models, benchmark data should contain diversity in how people perform activities in real-world environments [45]. Technically, an overall dataset should be split up into training and test datasets with no overlap even with respect to their collection condition, *e.g*., in [14, 27]; otherwise, the evaluation leads to overoptimistic performance results that will not hold when applied in the field [35]. Previous work, relying on individual models, splits data from a single user data collection session into training and test data. Due to the likely lack of variance exhibited in a single session, actual performance on data from an additional collection session or data from the field will be different, and often worse. In different data collection sessions, there will be differences in practice, *e.g.*, the position of worn sensors may shift or varying health conditions may affect the performance of activities. Finally, a variety of data collection environments, *e.g.*, sitting on different kinds of furniture, may improve the quality of the evaluation in general and lead to more realistic and less over-optimistic activity recognition systems.

III. DATA COLLECTION

A Participants

We conducted a data collection with 12 males and 16 females. Their demographics appear in Table 1 (age: mean= 44, SD= 17.1; weight: mean= 74kg, SD= 15.1; height: mean= 170.7cm, SD= 11.7) .

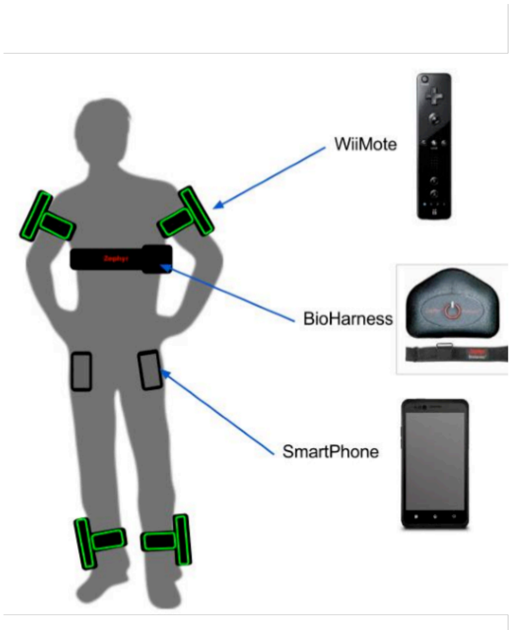
Table 1. Participants' demographic information.

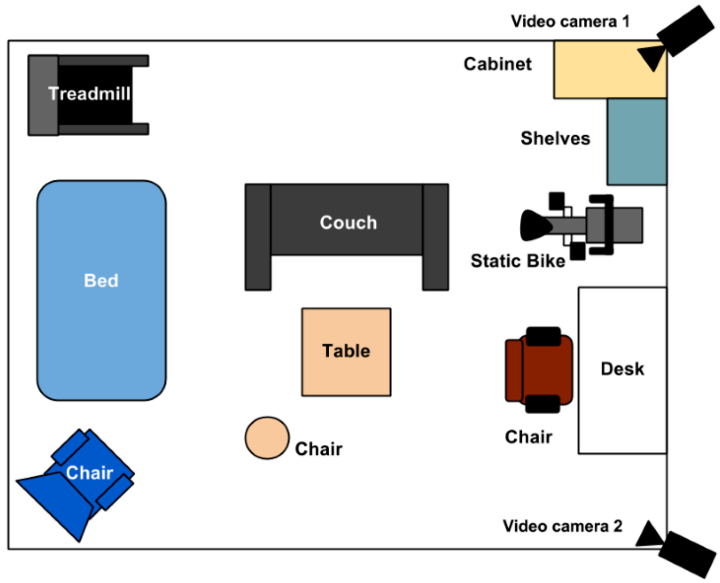
Participant	Age	Gender	Weight (kg)	Height (cm)

1	40	female	61.2	165.1
2	61	male	95.3	188.0
3	42	male	90.7	182.9
4	25	male	65.8	177.8
5	39	male	70.9	172.7
6	28	male	90.7	188.0
7	23	female	68.0	160.0
8	27	female	61.2	157.5
9	26	female	53.1	160.0
10	31	male	77.1	185.4
11	60	female	58.1	162.6
12	22	female	63.5	162.6
13	63	female	62.6	172.7
14	65	male	78.5	180.3
15	65	female	71.7	157.5
16	25	female	68.0	172.7
17	20	female	63.5	160.0
18	58	male	74.5	175.3
19	26	female	52.2	154.9
20	63	female	65.8	172.7
21	56	female	59.0	162.6
22	64	female	83.9	160.0
23	63	female	81.6	157.5
24	51	male	108.9	172.7
25	21	male	81.6	180.3
26	54	male	111.1	193.0
27	58	male	81.2	188.0
28	55	female	74.8	157.5

B Procedure

We attached a number of wearable sensors to each participant using custom, comfortable straps. Wii remotes were attached to the upper arms and calves, a BioHarness BT chestband was worn around the chest and two smartphones were placed inside the front pants pockets (Figure 1a). We chose the Wii remote as it is inexpensive and readily available. The participant was shown the testing area, furnished to resemble a one bedroom apartment (Figure 1b), and debriefed about the objective and details of the experiment and general activities to be performed. He was asked to perform a set of indoor activities as if he were in his own home. A list of thirty different activities was provided for participants to perform. None of these activities were the 7 target activities that we were interested in sensing: *sitting*, *walking (indoor or on a treadmill)*, *bicycling (on a stationary bike)*, *bending*, *lying*, and *falling*. We chose these activities because they correspond to the most basic and common activities in people's daily living and are useful for supporting assisted living environments for the elderly [1]. Instead, the suggested activities were high-level, composite, naturalistic activities such as moving books from one bookshelf to another, vacuuming, sitting on different chairs to read, and getting a drink of water from a water bottle across the room. While direct commands produce completely unnatural and robot-like performance that decrease the value of the data collected, we captured low-level activities in the wild during realistic task performance. For example, instead of asking participants to bend, we ask them to wipe a desk. Instead of standing we ask them to make a call using their own phone while standing or sitting. The one exception was the falling activity: we asked participants to fall onto an air mattress to reduce injury risks.





(a) **(b)**

Figure 1. a) Sensing platform. B) Sensing environment, including the view of the two cameras recording the study.

We captured video using two cameras in the environment (Figure 1) and used this to manually label our 7 target activities.

The position of sensors on the body can impact activity recognition accuracy. The optimal placements depend on the application and the type of activities being recognized [26]. In this work, we do not directly deal with the issue of sensor placement, but instead adopted a sensor placement approach commonly used [2, 3, 19, 21, 22, 24, 32, 40]. Three-dimensional acceleration data were collected from each worn device and breathing data were captured with the Bioharness BT device. All sensors streamed their sampled data to a laptop via Bluetooth or WiFi at 10Hz.

To capture a wider variety in how each participant performs activities, each was asked to complete 2 data collection sessions on different days, providing day-to-day variation. Each session lasted 1.5 hours on average. After extracting features and filtering out noise and missing values every second, 283,918 multidimensional data points (about 78.9 hours) of data were obtained from the 28 participants (standing: 13.6% of the samples, sitting: 35.8%, walking: 19.1%, bicycling: 7.3%, bending: 5.3%, lying: 17.7%, and falling: 1.1%)

IV SEMI-POPULATION-BASED PERSONALIZATION FOR ACTIVITY RECOGNITION

Instead of training a new activity model for a new user, our proposed activity recognition system constructs the model for him by exploiting activity models built from existing subjects. For this, our system has training, calibration, and test phases. In the training phase, we construct a pool of activity models, each of which is trained on the existing labeled data of users. From n users $U=\{u_1, u_2, \ldots, u_n\}$ and on m activities $A=\{a_1, a_2, \ldots, a_m\}$, we train sub-models that will be used to construct a hybrid activity model, including a group of Bayesian networks $BN=\{bn_{11}, bn_{12}, \ldots, bn_{1m}, bn_{21}\ldots, bn_{nm,} nb_{11}, nb_{12}, \ldots, nb_{1m}, nb_{21}\ldots, nb_{nm}\}$ and a group of support vector machines $SVM=\{svm_{11}, svm_{12}, \ldots, svm_{1m}, svm_{21}, \ldots, svm_{nm}\}$, where $bn_{ij,}$ nb_{ij} and svm_{ij} represent a Bayesian network, a naïve Bayes classifier and a support vector machine, respectively, that are trained on the labeled data of user u_i recognizing activity a_j. BNs were chosen because they provide interpretable probability estimates, and SVMs were chosen because they are very accurate for binary classification problems. Also, the first has lower variance but higher bias (underfitting) while the second has higher variance but lower bias (overfitting). By hybridizing the BNs and SVMs in the proposed structure, we aim to manage the overfitting issue of SVMs, which are often more accurate, with the underfitting characteristics of BNs. Moreover, we model activities individually as binary problems with SVMs and BNs to manage the highly unbalanced activity dataset. For the sum, we use a polynomial kernel function with degree=2 and cost=1. The training phase is completed before the system is deployed to a new user.

A Calibration Phase: Semi-population

For a new user who starts to use our system, we have two other phases of the system: calibration and test. In the calibration phase, we collect a *small* amount of labeled data from the new user, which represents her performance of the m activities. On this newly collected labeled data $N=\{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$ ($y_i \in \{1, 2, \ldots, m\}$), where x_i is the feature vector of the i^{th} sample and y_i is its label, we measure the fitness of all sub-models in the pool $F=\{f_{bn11}, f_{bn12}, \ldots, f_{svm11}, f_{svm12}, \ldots, f_{svmnm}\}$, constructed in the training phase (f_{bnij} is the fitness of the activity model bn_{ij}). Note that there are 3×7 models for each user from the training phase (*i.e.*, Bayesian network, naïve Bayes classifier and SVM for each of the 7 activities) Here, F is not limited only to accuracy. In this work, we use the F1 score [33] of a sub-model $sm_{ij} \in \{bn_{ij}, nb_{ij}, svm_{ij}\}$ as its fitness metric.

$Precision(sm_{ij}, N) = tp_j/(tp_j + fp_j)$, where tp_j and fp_j are the number of true positive and false positive predictions made by sm_{ij} for the j^{th} class of N.

$Recall(sm_{ij}, N) = tp_j/(tp_j+fn_j)$, where tp_j and fn_j are the number of true positive and false negative predictions made by sm_{ij} for the j^{th} class of N.

$$F1\ score(sm_{ij}) = 2 \times (Precision(sm_{ij}, N) \times Recall(sm_{ij}, N)) / (Precision(sm_{ij}, N) + Recall(sm_{ij}, N))$$

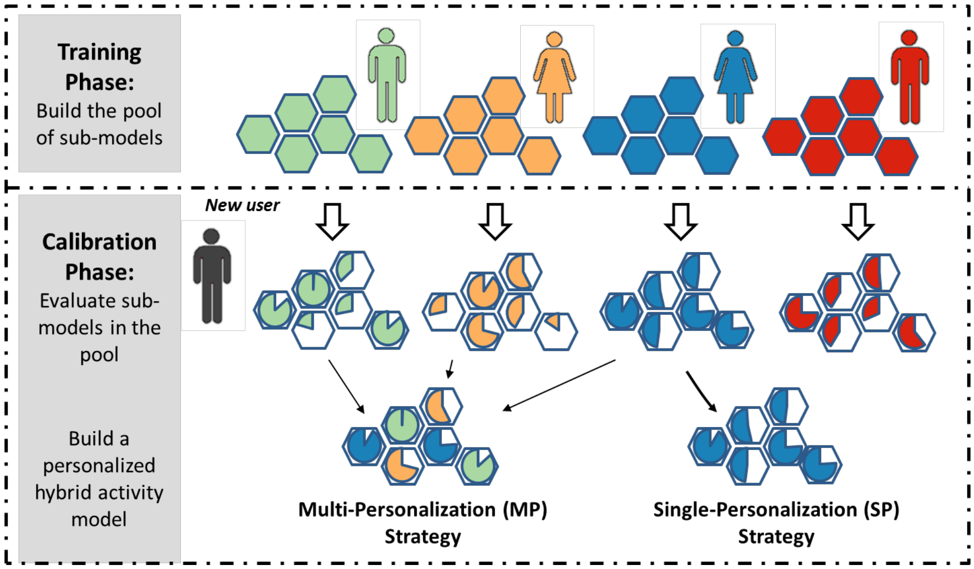


Figure 2. Semi-population based personalization. The first two stages are always the same the last two are mutually exclusive.

Based on the fitness of each sub-model, we select m BNs $\{b_1, b_2, \ldots, b_m\}$ and m SVMs $\{s_1, s_2, \ldots, s_m\}$ from the pool, and use them to construct a *hybrid* (*i.e.*, consisting of both BNs and SVMs) activity model for the new user. When selecting BNs and SVMs, we design two different strategies as shown in Figure 2. The first strategy, multi-personalization (MP), selects the b_i and s_i (BN and SVM) that have the highest fitness scores in recognizing the i^{th} activity. Therefore, the m BNs and m SVMs may be selected from a variety of different users in the pool. The second strategy, single-personalization (SP), instead restricts the selection of m BNs and m SVNs to be from a single user in the pool. For this, we first calculate sub-models' fitness as with MP. Then, for each user, we average the fitness scores of their sub-models. The user with the highest resulting score is selected and his/her models are used for the new user.

B Test Phase: Hybrid Activity Recognition

With m BNs $B=\{b_1, b_2, \ldots, b_m\}$ and m SVMs $S=\{s_1, s_2, \ldots, s_m\}$ selected by using either MP or SP in the calibration phase of a new user, a hybrid activity model is composed as shown in Figure 3 and can be used to recognize this new user's activities during the test phase. When a new feature vector (*i.e.*, additional data from the performance of an activity) from the user is input, our hybrid model first estimates the probability of each of the m activities by using the m BNs: $P=\{p_1, p_2, \ldots, p_m\}$. It then uses the SVMs, ordered according to their probabilities, in a subsumption architecture [7, 20], *i.e.*, the SVM of an activity with the highest probability checks first whether the input vector should be classified as the corresponding activity. The hybrid activity model continues through the ordered list of SVMs until an activity is recognized.

As a binary classifier, an SVM usually classifies a sample into its positive class if the prediction output (*i.e.*, distance to the hyperplane) is a positive number. When applying SVMs trained separately for each class to a multiclass problem, an additional process is often needed to choose among their individual outputs. For this, we define m thresholds θ={θ_1, θ_2, …, θ_m}, used to classify a sample as the i^{th} class if the output of s_i on the sample is over θ_i. θ was chosen to optimize accuracy on the training data.

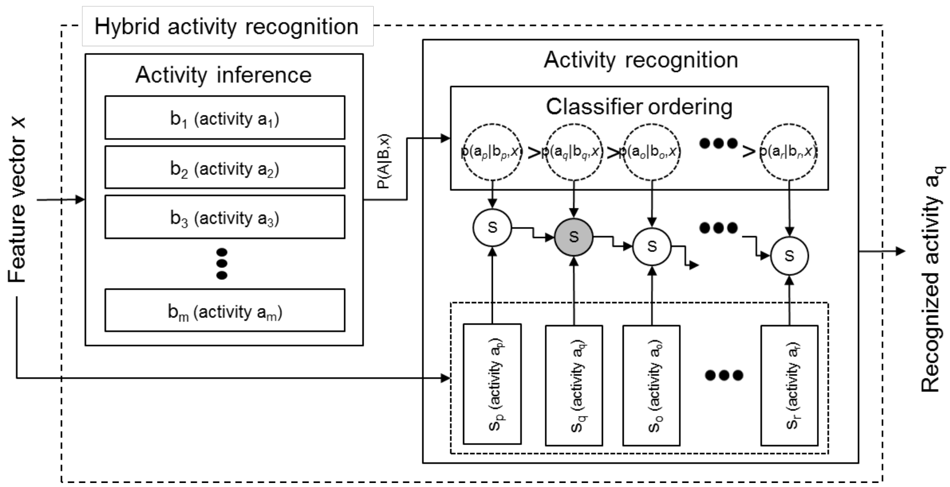


Figure 3. Structure of the proposed hybrid model. Our hybrid model uses a combination of Bayesian networks and SVMs.

For a new instance of activity data, *x*, from the new user, our hybrid model recognizes the activity as follows:

1. Calculate $P(A|B,x)$ = {$p(a_1|b_1,x)$, $p(a_2|b_2,x)$,…, $p(a_m|b_m,x)$}, where $p(a_i|b_i,x)$ is the probability of the i^{th} activity for x as inferred by b_i
2. Sort S={s_1, s_2, …, s_m} according to P and get $\acute{S}$={$\acute{s}_{1q}$, $\acute{s}_{2p}$, …, $\acute{s}_{mr}$}, where s_i corresponds to $p(a_i|b_i,x)$ and $\acute{s}_{1q}$ is s_q whose $p(a_q|b_q,x)$ is the highest among P while $\acute{s}_{mr}$ is s_r whose $p(a_r|b_r,x)$ is the lowest among P
3. For i=1 → m, *if* $\acute{s}_{iq}(x) > \theta_q$, return q as the prediction
4. If no s is satisfied, return *arg_max*(P), the activity with the highest probability

In this way, our hybrid activity model only generates an output from the outputs of the new user's m×2 sub-models.

In this work, we use the SMILE reasoning engine (developed by the Decision Systems Laboratory at the University of Pittsburgh and available at http://genie.sis.pitt.edu) to train Bayesian networks and naïve Bayes classifiers for each activity. The streaming sensor data from the worn sensors are segmented every second by using a 2-sec sliding window with 50% overlap. For each segment, a number of features are calculated to form a feature vector (Table 2). Based on a variety of features investigated in other work [14, 26, 27, 39], we extract the mean and standard deviation of each channel. In total, the dimensionality of the input feature space is 158.

Table 2. List of features used.

Features	Data streams	Sensors
mean, std	x, y, z, abs(x), abs(y), abs(z), x×y, y×z, x×z, x×y×z	Bioharness, smartphone A (left pocket), smartphone B (right pocket), Wii A (left arm), Wii B (right arm), Wii C (left ankle), Wii D (right ankle)
mean, std	peak acceleration, max(x), max(y), max(z), min(x), min(y), min(z), breathing rate, leaning angle	Bioharness

We train the Bayesian networks after selecting the 10 most informative features according to their information gain and discretizing them each into 5 states based on the training data. We train the SVMs using the SVM-Light library (available at http://svmlight. joachims.org) with a polynomial kernel function (degree=2, cost=1).

IV EVALUATION

A Methods

To evaluate our semi-population calibration approach, we use the leave-one subject/session-out cross validation [27, 28] that separates training and test data in terms of a data collection condition, and compare our approaches with common approaches: hybrid *vs.* non-hybrid, and semi-population *vs.* individual and population models. We use a single session of a user for training and use the other session for testing. We repeat this 56 (=28 participants×2 sessions) times and report the average.

For our sub-population calibration approach, we train models on all but one test user, and put these into the pool. Then, we take a small sub-set of the test user's data, to use in selecting appropriate models from the pool (from any combination of users for the multi-personalization (MP) strategy, or from a single user for the single-personalization (SP) strategy). This process is repeated for every user. Note that our sub-population approach uses the training data only for calibrating (*i.e.*, selecting) models for a new user, while the individual approach uses the user's own training data to learn his model and the population approach uses others' data to learn a model for him.

B Hybrid Activity Recognition results

We analyzed the performance of the single activity models (Table 3). In general, our single activity models, which only rely on standard machine learning methods with simple features like means and standard deviations of sensor signals (see Table 2), show moderate F1 scores. For some activities like sitting and lying, both BNs and SVMs achieved high accuracy even with the very simple features. However, similar accuracies were not achieved for bending and falling due to the lack of enough data for modeling (as these activities only constitute 5.3% and 1.1% of the collected data, respectively). Surprisingly, given the results from previous studies, standing was hard to recognize accurately. As our users engaged in natural activities which involved standing, as opposed to being told to stand still without much movement, we observed large variations in the way users stood, *e.g.*, a user swayed or turned his body while standing.

Table 3. Average F1 scores for each activity

F1 score	BN	SVM	Hybrid
Standing	0.526	0.626	0.688
Sitting	0.787	0.896	0.909
Walking	0.753	0.763	0.816
Bicycling	0.699	0.819	0.853
Bending	0.507	0.547	0.588
Lying	0.914	0.861	0.915
Falling	0.584	0.604	0.533

For the non-hybrid approach, base models, *i.e.*, BNs or SVMs, produce a final output by using the winner-takes-all strategy (*non-hybrid*) to aggregate multiple outputs from the 7 BNs or 7 SVMs, respectively. In recognizing each activity, as shown in Table 3, the hybrid approach shows an improvement of F1 scores for all activities except falling. For the general recognition performance, the BN-based non-hybrid model obtains 77.4% accuracy while the SVM-based non-hybrid model obtains 81% accuracy. The SVM-based non-hybrid model has better general performance than the BN-based non-hybrid model, which has a larger decrease in accuracy when going from the training data set to the test data set. The hybrid approach, which achieves 83.4% accuracy, performs better than either non-hybrid approach.

Table 5. Averaged confusion matrix of hybrid activity recognition using our semi-population MP approach. Column labels refer to the real class and rows to predicted class.

	Stand	Sit	Walk	Bike	Bend	Lie	Fall
Stand	490.5	49.5	100.1	4.5	37.0	5.8	4.4
Sit	45.3	1662.6	13.0	15.6	25.7	50.4	1.3
Walk	128.5	22.4	782.1	9.5	14.9	1.3	9.9
Bike	14.1	29.3	10.8	303.4	8.4	3.0	3.3
Bend	54.3	13.8	32.5	5.5	151.5	6.7	5.0
Lie	1.0	66.2	2.8	0.1	4.7	820.5	2.7
Fall	4.4	1.4	8.1	0.7	4.2	7.2	30.0

Table 5 shows the confusion matrix of the proposed hybrid activity recognition with our semi-population MP approach. On the validation data, it achieves an accuracy of 93.1% where there are some misclassifications between standing, walking and bending. Some instances of falling behaviors were confused with standing and lying, resulting in a small number but large portion of errors in classifying falling. Also, some errors of falling and bending activities being classified as walking can be explained by the fact that they often happen during walking. Note that we asked participants to fall as naturally as they could instead of specifying how to perform falling behaviors. The errors due to the ambiguity among activities increase on test data where standing is confused with sitting and bending, and bicycling with sitting.

C Semi-population Strategies results

1. MP and SP comparison

We compare the performance of the semi-population approaches for MP (select models from any combination of users in the pool) and SP (select all models from a single user) strategies. The MP strategy has an accuracy of 83.4% (kappa: 0.852), while the SP strategy has an accuracy of 80.7% (kappa: 0.745). The individual model has an accuracy of 77.3% (kappa: 0.79), while the population model has an accuracy of 77.4% (kappa: 0.743), on average. The MP strategy improves upon the individual and population models by 6%, a meaningful improvement. It achieved the increased accuracy without a complete labeled dataset to be acquired. Further, the MP strategy achieves not only the highest accuracy but also a high kappa value of 0.852, which means that our method is capable of accurately recognizing not only the majority activities (common output classes like sitting) but also the minority activities (*e.g.*, bending, falling). However, the SP strategy and the population model have similar accuracies but lower kappa values than the individual model, where the two methods might have more difficulty in recognizing the minority activities. The individual model has the lowest accuracy but a relatively higher kappa value. The population model works slightly better than the individual model on our dataset in terms of accuracy, since the individual model does not handle well the day variance in how users perform activities, for some users. Also, the population model often overperforms on the majority activities (and correspondingly, underperforms on the minority classes), and given the highly unbalanced activities distribution, this may lead to a higher accuracy but a relatively lower kappa value.

The MP and SP strategies had the highest performance with subject 2 (94.3%) and subject 21 (92.8%), respectively, while the individual and population models had the highest accuracy with subject 3 (92.8%) and subject 19 (93.2%), respectively. The poorest cases of the four methods were for subjects 10 (65.2%), 10 (60.3%), 22 (46.9%), and 10 (58.8%), respectively. For our 28 participants, the MP strategy performed the best among the four methods for 14 participants, the SP strategy for 6 participants, the individual models for 3 participants (and in all 3 cases was only slightly better) and population model for 5 participants. Interestingly, the individual and population models often exhibited different results. If one performed well then the other tended to perform poorly, *e.g.*, subject 3 had 94.1% and 81.3% accuracies with the individual and population models, respectively, while 19 obtained 81.2% and 93.2% accuracies. In general, the MP strategy performed better than the SP strategy. For the cases where it performed worse, it was due to overfitting of the models on some specific activities (*e.g.*, subjects 18, 24, 26). For some participants like 6, 9, 20, 22, and 28, the MP strategy was about 10% more accurate than the SP strategy, which means that it was hard to find a single subject who has similar patterns as the test user for all 7 activities. The SP strategy performs reasonably well, but for many subjects, it failed to find a good match among the subjects in the pool, *e.g*., subject 6 has 82.8% accuracy with the MP strategy, but only 66.3%, 69.3%, and 73.3% with the SP strategy, individual model, population model, respectively. Several participants (*e.g*., 2, 3, 11, 17) had high accuracies (close to 90%) over the four methods, while others (*e.g.*, 7, 10, 15, 22) were not able to achieve good performance (all lower than 70% accuracy) with any method.

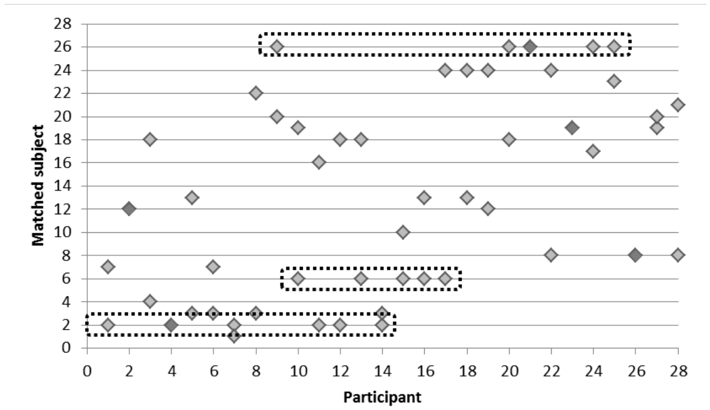
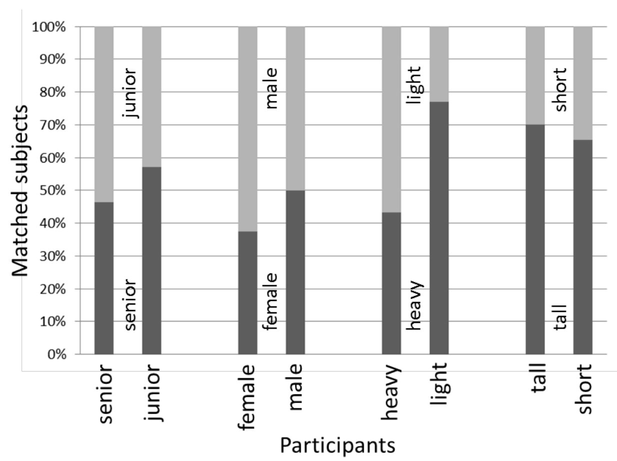


Figure 4. Users and their matched subjects by our semi-population SP approach.

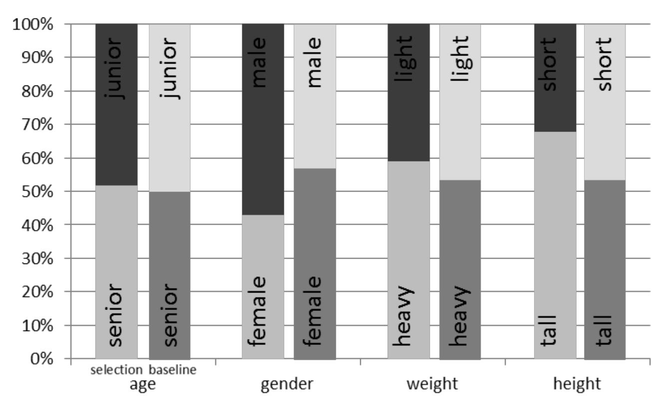
To further understand how people relate to each other through our semi-population SP approach and in the spirit of Lane *et al*. [25], we looked specifically at what types of users in the pool proved to be good matches for new users. Figure 4 shows which users in the pool were selected as matches during two runs for each participant (one run for each session of data) There was no one dominant subject that was commonly selected, but several subjects were selected multiple times, *e.g*., S02 (7), S26 (6), and S06 (5) as highlighted in the figure. The participants who matched one of these three subjects were different in their demographics compared to their matched subject, *e.g*., two young, light in weight and short females and a young, light in weight and average height male matched subject 2, a senior, heavy in weight, tall male.

In order to better understand the relationship between demographics and the likelihood of users matching, we grouped our users by age, senior (>50) as group 1 (50%) and younger (<50) as group 2 (50%); by gender, females as group 1 (57%) and males as group 2 (43%); by weight, heavy (>70kg) as group 1 (54%) and light (<70kg) as group 2 (46%); and, by height, taller (>170cm) as group 1 (54%) and shorter (<170cm) as group 2 (46%). Figure 5 shows what proportion of users in each group matched other users in that same group.

In contrast to the commonly held belief that people who have similar demographics would have similar activity patterns, users who matched each other were often different in their demographics. Figure 5(a) presents information on how our subject participants matched with other users, broken out by demographics. For example, we can see that only 37% of the females in our study population matched other female subjects.. Similarly, light in weight users (group 2) used the models of heavy in weight subjects (group 1) instead of those of users in the same group almost 80% of the time, while shorter users (group 2) matched taller users (group 1) 65% of the time. Looking at the averages across groups, we see that only 45%, 43% and 34% of users matched subjects in the same group by age, gender, and weight. Height was slightly more predictive, with 54% of users matching.



(a) demographic distribution of test users and matched subjects



(b) demographic distribution of subjects being matched

Figure 5. Influence of user demographics on matched subjects using the SP strategy.

Figure 5(b) presents an alternative view of these same data, showing the distribution of demographics for the baseline (subject) population in our study, and the demographics of the subjects in the pool that were selected or matched. For example, looking at age, slightly less than 50% of our subject population (baseline) were seniors (group 1), yet seniors were matched slightly more than 50% of the time (selection). This graph indicates that 1) it is *not true* that our similarly aged subjects have similar activity behaviors, and 2) male, heavy in weight, and tall users were disproportionately matched, so their models were most representative of the people in our pool. Especially, tall males have higher chance to be matched to any users in the pool (58% of users matched with male subjects (43% of users are male), 68% of users matched with tall subjects (54% of users are tall)). Note that there is likely a correlation between males being heavy in weight and tall, in the general population.

In addition, we investigated how physically similar people in a training pool could be used to help identify activity recognition models for new users who have similar characteristics in activity recognition. Instead of our mechanism that evaluated others' activity models based on the F1-score, we matched people whose demographics were the most similar in the pool and applied our semi-population SP approach. As shown in Table 7, recognition relying on age and weight performed more poorly than the individual model. Recognition relying on models taken from users' with similar height information was more accurate than the individual model. In fact, it performed better than all other models except for the original MP and SP non-demographic based approaches. As a guideline, if there are no labeled sensor data available for a new user, considering his/her height could be useful in selecting appropriate models.

Table 7. Comparison with demographic-based matching with the SP strategy

	Test accuracy
Age	72.1%
Weight	74.5%
Height	79.4%
Weight+Height	77.2%
All	78.2%
MP	83.4%
SP	80.7%
Individual	77.3%
Population	77.7%

Table 8 shows the average performance of the different demographic groups. Although these results would be more generalizable with a larger population, with our population, our system worked better for older adults than younger adults, males than females, light in weight users than heavy in weight users, and taller users than our shorter users.

Table 8. Average performance obtained by groups (MP strategy). The differences in most cases are quite small.

Error rate	Group 1	Group 2
Age	16.8%	16.3%
Gender	17.3%	15.6%

Weight	17.9%	15.0%
Height	16.1%	17.0%

2. Data size

To further understand the applicability of our approach, we evaluated the semi-population approach with two questions:

1) *How much calibration data have to be collected from a new user to have reasonable performance?*

2) *How many training users have to be in the pool for reasonable performance?*

Although more labeled data often improve recognition, a new user may not be able to provide a large amount of labeled calibration data. Reducing the data required for calibration would greatly increase the acceptance of our system [6]. To investigate how much we can reduce the requirement for labeled data for calibration, we varied the amount of the data (half and quarter of the 1.5h labeled data sampled by stratified selection to preserve the original class distribution) available and applied our semi-population approach to find the best matched activity models with smaller amounts of data than the original dataset.

We achieved accuracies of 83.5% and 80.7% using the MP and SP strategies, respectively, when they used all 1.5 hours of calibration data. Individual models, which used the same data for training, produced 77.3% accuracy. When using only one-half and one-quarter of the 1.5h calibration data, our approach achieved 81.1% and 79.7% with the MP strategy, and 80.3% and 78.1% with the SP strategy, all more accurate than the individual models. If a user collects 22 min of labeled data, our system is still able to outperform his individual model trained from 1.5 hours of his labeled data.

For the second question, we varied the number of training users in the pool from the original 27, to multiples of 5 users down to 5 users using random selection. As shown in Table 11, we still achieved higher accuracies than the individual models, even with only 5 training users in the pool. Using models from multiple people (MP), although the number of people may be small, is still better than picking all the models from a single person (SP).

Table 11. Average performance with different numbers of training users (results averaged across five random repeats).

Error rate	Semi-population		Individual
	MP	SP	(own)
(28-1) others	16.6%	19.3%	22.7%
25 others	16.5%	19.6%	
20 others	16.9%	20.2%	
15 others	17.0%	20.2%	
10 others	17.8%	19.8%	
5 others	18.7%	21.2%	

D Field study results

We conducted a limited, but more realistic, study in additional subjects' own homes to show that the system built from the activity data collected in the lab can be applied in a real environment. We recruited 3 males and 3 females and added an exercise bicycle to their homes. We did not move any furniture or restrict movement throughout the subjects' houses. For consistency, we asked subjects to wear the same set of wearable sensors described in our lab study. We instructed them to behave as they would normally in their homes (with the exception of the added bicycle). As such, we did not collect any data for the falling activity (although we did capture data from one subject who fell by accident). On average, the data collection lasted 30 minutes per subject.

The first half of the data collected from each volunteer was used for calibration as training data using the MP strategy, to select from the pool of models created in the lab study. The remaining half was used as test data to measure the performance of the retrieved models in the field. The evaluation was conducted twice, swapping the training and test datasets.

Table 9. Aggregated confusion matrix of hybrid activity recognition using our semi-population MP approach.

	Stand	Sit	Walk	Bike	Bend	Lie	Fall
Stand	888	133	71	3	55	0	1
Sit	59	4031	37	26	76	162	5
Walk	34	113	1113	8	153	1	22
Bike	160	238	50	1498	94	33	8
Bend	135	169	67	22	408	63	14
Lie	1	99	0	0	0	1898	10
Fall	0	0	0	0	0	0	1

The accuracy across the in-homes data collection was 82.3% (kappa: 0.76) (Table 9). The average accuracy across users was 80.6% (kappa: 0.73).

E Short time horizon results

We also validated our approach in a lab study and a more limited but realistic in-home study. The in-home evaluation was conducted with 30 minutes of data per subject, on average. This was for two reasons. First, our subjects did not permit us to install video cameras everywhere in their homes to record activities, so researchers had to do more limited manual observation and recording, *e.g.*, living room and kitchen. Second, we wanted to demonstrate the validity of our approach with a very *limited* amount of data with which to select models: *15 minutes*. With this amount of labeled data, the accuracy of our approach for the in-home and lab studies was almost the same as for the lab study.

F Limitations of the results

Given our contributions, we also recognize that the current instantiation of our approach has some limitations. First, our current system uses a number of wearable sensors, but users may only wear a few in practice [2, 3, 5, 21, 31]. Second, the features used in modeling are very naïve, where we could improve our system by incorporating common but more sophisticated features in the domain of sensor-based activity recognition. Here, features, either orientation-free or calibrated constantly in orientation, are recommended to cope with sensor displacement since end users are often not consistent or well-trained in wearing such systems [7, 16]. In our in-home study, there was a higher degree of sensor displacement and changing orientation, but the results were quite good, despite this. Third, other strategies to improve the semi-population approach could be available, *e.g.*, rather than selecting on the sub-model with the highest fitness, the *n*-highest models could be selected and combined, to support a greater diversity in activity models and subjects and to avoid over-fitting.

6. CONCLUSION

In this paper, we focused on a novel and effective approach for deploying activity systems, in which we proposed the construction of accurate activity models and sharing them across different users. We built activity models from a number of people (training users) and used these models adaptively for a new user who only needs to provide a small amount of labeled activity data. With a multimodal wearable sensing platform, we collected a variety of activity data from 28 people who varied in age, gender, height and weight, and used these data to build a number of hybrid activity models. With the models, our semi-population approach could recognize their activities with an accuracy of 83.4% while their individual models and the population model obtained 77.3% and 77.4%, respectively. In a small in-home study, our approach performed similarly well. Our approach of measuring the fitness of activity models with the F1 score outperforms an approach that relies on the similarity of people in demographics. In the future, we aim to enhance our approach by introducing new features that are more general and robust to its users and environments with further validation on other benchmark datasets. Also, an analysis of people’s indoor/outdoor activities with other information such as their fitness level in addition to their demography could be useful for understanding various aspects of human activity. Finally, we will develop and deploy an end-user application with a lighter sensing platform in a longitudinal study.

Manuscript received

REFERENCES

[1] Álvarez-García, J.A. Evaluating human activity recognition systems for AAL environments. *Communications in Computer and Information Science 362*, (2013), 131-136.

[2] Atallah, L., Lo, B., King, R., Yang, G.-Z. Sensor positioning for activity recognition using wearable accelerometers. *IEEE Trans. Biomedical Circuits and Systems 5*, 4 (2011), 320-329.

[3] Banos, O., Damas, M., Pomares, H., Rojas, I., Toth, M.A., Amft, O. A benchmark dataset to evaluate sensor displacement in activity recognition. In *Proc. Ubicomp'12*, (2012), 1026-1035.

[4] Berchtold, M., Budde, M., Gordon, D., Schmidtke, H. Actiserv: Activity recognition service for mobile phones. *ISWC*. (2010), 1-8.

[5] Bergmann, J., Graham, S., Howard, N., McGregor, A. Comparison of median frequency between traditional and functional sensor placements during activity monitoring. *Measurement 46*, 7 (2013), 2193-2200.

[6] Bhattacharya, S., Nurmia, P., Hammerla, N., Plötz, T. Using unlabeled data in a sparse-coding framework for human activity recognition. Pervasive and Mobile Computing 15, (2014), 242-262.

[7] Brooks, R. A robust layered control system for a mobile robot. *IEEE J. Robotics Automation 2*, 1 (1986), 14–23.

[8] Carletta, J. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics 22*. 2 (1996), 249-254.

[9] Chavarriaga, R., Bayati, H., Millan, J. Unsupervised adaptation for acceleration-based activity recognition: robustness to sensor displacement and rotation. *Personal and Ubiquitous Computing 17*, 3 (2013), 479-490.

[10] Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Troster, G., Millan, J.R., Roggen, D. The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters 34*, 15 (2013), 2033-2042.

[11] Chen, L., Nugent, C.D., Wang, H. A knowledge-driven approach to activity recognition in smart homes. *IEEE Trans. Knowledge and Data Engineering 24*, 6 (2012), 961-974.

[12] Chernbumroong, S., Cang, S., Atkins, A., Yu, H. Elderly activities recognition and classification for applications in assisted living. *Expert Systems with Applications 40*, 5 (2013), 1662-1674.

[13] Cvetkovic, B., Kaluza, B., Luštrek, M., Gams, M. Semi-supervised learning for adaptation of human activity recognition classifier to the user. Workshop on Space, Time and Ambient Intelligence, IJCAI. (2011).

[14] Dalton, A., OLaighin, G. Comparing supervised learning techniques on the task of physical activity recognition. *IEEE J. Biomedical and Health Informatics 17*, 1 (2013), 46-52.

[15] Fleury, A., Vacher, M., Noury, N. SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. *IEEE Trans. Information Technology in Biomedicine 14*, 2 (2010), 274-283.

[16] Gao, L., Bourke, A.K., Nelson, J. A system for activity recognition using multi-sensor fusion. In *Proc. 33rd Annual Int. Conf. of the IEEE EMBS*, (2011), 7869-7872,

[17] Geman, S., Bienenstock, E., Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation 4*, 1 (1992), 1-58.

[18] Gu, T., Wang, L., Wu, Z., Tao, X., Lu, J. A pattern mining approach to sensor-based human activity recognition. *IEEE Trans. Knowledge and Data Engineering 23*, 9 (2011), 1359-1372.

[19] Haslgrubler, M., Holzmann, C. DarSens: a framework for distributed activity recognition from body-worn sensors. In *Proc. BodyNets'10*, (2010), 240-246.

[20] Hong, J.-H., Min, J.-K., Cho, U.-K., Cho, S.-B. Fingerprint classification using one-vs-all support vector machines dynamically ordered with naive Bayes classifiers. *Pattern Recognition 41*, 2 (2008), 662-671.

[21] Keally, M., Zhou, G., Xing, G., Wu, J., Pyles, A. PBN: towards practical activity recognition using smartphone-based body sensor networks. In *Proc. Sensys'11*, (2011), 246-259.

[22] Khan, A.M., Lee, Y.-K., Lee, S.Y., Kim, T.-S. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Trans. Information Technology in Biomedicine 14*, 5 (2010), 1166-1172.

[23] Krishnan, N., Cook, D. Activity recognition on streaming sensor data. *Pervasive and Mobile Computing 10*, (2014), 138-154.

[24] Kwapisz, J.R., Weiss, G.M., Moore, S.A., Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter 12*, 2 (2010), 74-82.

[25] Lane, N., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, A.T., Zhao, F. Enabling large-scale human activity inference on smartphones using community similarity networks (CSN). In *Proc. Ubicomp'11*, (2011), 355-364.

[26] Lara, O., Labrador, M. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials 15*, 3 (2013), 1192-1209.

[27] Li, M., Rozgica V., Thatte G., Lee S., Emken A., Annavaram M., Mitra U., Spruijt-Metz D., Narayanan S. Multimodal physical activity recognition by fusing temporal and cepstral information. *IEEE Trans. Neural Systems and Rehabilitation Engineering 18*, 4 (2010), 369-380.

[28] Lu, C.-H., Ho, Y.-C., Chen, Y.-H., Fu, L.-C. Hybrid user-assisted incremental model adaptation for activity recognition in a dynamic smart-home environment. *IEEE Trans. Human-Machine Systems 43*, 5 (2013), 421-436.

[29] Maekawa, T., Kishino, Y., Sakurai, Y., Suyama, T. Activity recognition with hand-worn magnetic sensors. *Personal Ubiquitous Computing 17*, 6 (2013), 1085-1094.

[30] Magherini, T., Fantechi, A., Nugent, C.D., Vicario, E. Using temporal logic and model checking in automated recognition of human activities for ambient-assisted living. *IEEE Trans. Human-Machine Systems 43*, 6 (2013), 509-521.

[31] Martin, H., Bernardos, A.M., Iglesias, J., Casar, J.R. Activity logging using lightweight classification techniques in mobile devices. *Personal Ubiquitous Computing 17*, 4 (2013), 675-695.

[32] Parkka, J., Cluitmans, L., Ermes, M. Personalization algorithm for real-time activity recognition using PDA, wireless motion bands, and binary decision tree. *IEEE Trans. Information Technology in Biomedicine 14*, 5 (2010), 1211-1215.

[33] Powers, D. Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *J. Machine Learning Technologies 2*, 1 (2011), 37–63.

[34] Reiss, A., Stricker, D., Personalized mobile physical activity recognition. *ISWC* (2013), 25-28.

[35] Reiss, A., Stricker, D., Hendeby, G. Towards robust activity recognition for everyday life: methods and evaluation. In *Proc. 7th Int. Conf. on Pervasive Computing Technologies for Healthcare and Workshops*, (2013), 25-32.

[36] Riboni, D., Bettini, C. COSAR: hybrid reasoning for context-aware activity recognition. *Personal and Ubiquitous Computing 15*, 3 (2011), 271-289.

[37] Saguna, Zaslavsky, A., Chakraborty, D. Complex activity recognition using context-driven activity theory and activity signatures. *ACM Trans. Computer-Human Interaction 20*, 6 (2013), 32:1-32;34.

[38] Stikic, M., Larlus, D., Ebert, S., Schiele, B. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE Trans. Pattern Analysis and Machine Intelligence 33*, 12 (2011), 2521-2537.

[39] Sun, X., Kashima, H., Ueda, N. Large-scale personalized human activity recognition using online multitask learning. *IEEE Trans. Knowledge and Data Engineering 25*, 11 (2013), 2551-2563.

[40] Varkey, J.P., Pompili, D., Walls, T.A. Human motion recognition using a wireless sensor-based wearable system. *Personal Ubiquitous Computing 16*, 7 (2012), 897-910.

[41] Wang, L., Gu, T., Tao, X., Lu, J. A hierarchical approach to real-time activity recognition in body sensor networks. *Pervasive and Mobile Computing 8*, 1 (2012), 115-130.

[42] Xu, W., Zhang, M., Sawchuk, A.A., Sarrafzadeh, M. Robust human activity and sensor location corecognition via sparse signal representation. *IEEE Trans. Biomedical Engineering 59*, 11 (2012), 3169-3176.

[43] Zhang, M., Sawchuk, A. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proc. Ubicomp'12*, (2012), 1036-1045.

[44] Zhang, S., McClean S.I., Scotney, B.W., Probabilistic learning from incomplete data for recognition of activities of daily living in smart homes. *IEEE Trans. Information Technology in Biomedicine 16*, 3 (2012), 454-462.

[45] Zhang, M., Sawchuk, A.A. Human daily activity recognition with sparse representation using wearable sensors. *IEEE J. Biomedical and Health Informatics 17*, 3 (2013), 553-560.

[46] Zhao, Z., Chen, Y., Liu, J., Shen, Z., Liu, M. Cross-people mobile-phone based activity recognition. *AAAI*(2011), 2545-2550. AAAI Press.

[47] Zhu, C., Sheng, W. Motion- and location-based online human daily activity recognition. *Pervasive and Mobile Computing 7*, 2 (2011), 256-269.