The Benefits of Personalized Smartphone-Based Activity Recognition Models*

Jeffrey W. Lockhart† Gary M. Weiss†

Abstract

Activity recognition allows ubiquitous mobile devices like smartphones to be context-aware and also enables new applications, such as mobile health applications that track a user's activities over time. However, it is difficult for smartphone-based activity recognition models to perform well, since only a single body location is instrumented. Most research focuses on universal/impersonal activity recognition models, where the model is trained using data from a panel of representative users. In this paper we compare the performance of these impersonal models with those of personal models, which are trained using labeled data from the intended user, and hybrid models, which combine aspects of both types of models. Our analysis indicates that personal training data is required for high accuracybut that only a very small amount of training data is necessary. This conclusion led us to implement a self-training capability into our Actitracker smartphone-based activity recognition system[1], and we believe personal models can also benefit other activity recognition systems as well.

1 Introduction.

Activity recognition (AR) on mobile devices is a rapidly growing field. The ability of a device to recognize its user's activity is important because it enables context-aware applications and behavior. Activity recognition also makes it possible to develop mobile health applications that track a user's activities, such as our Actitracker app [1], which can help address the many health concerns that arise due to inactivity, including childhood obesity [11]. The work described in this paper relies on Android smartphones, but the tri-axial accelerometer present in these smartphones is very similar to those found in other smartphones and mobile devices.

In this paper we employ a straightforward approach for implementing AR. We collect accelerometer data from users as they walk, jog, climb stairs, sit, stand, and lie down, and then aggregate each 10 seconds of data into a single labeled example. We then induce an AR model by applying common classification algorithms to the generated training examples. The work in this paper includes data from our set of 59 test subjects [24], as well as data from 414 subjects in the HASC 2010 and 2011 data sets [9].

There has been much prior work on AR [4, 12, 13, 23], and a smaller but growing body of work on smartphone-based AR [5, 10, 15, 28]. While some work has used personal models [19, 28], which are built exclusively using labeled training data from the intended user, most work has focused on impersonal models [5, 8, 16, 20], which are built using data from a panel of users who are not the intended users of the model. Although newer work [16] has aimed at tailoring impersonal models to individuals, little work has compared personal and impersonal models on a reasonably-sized population, and no work has carefully analyzed the relative performance of these types of models. In this paper we provide a thorough analysis of the relative performance of these models and we view this as one of our main contributions. We conclude that personal AR models are extraordinarily effective and vastly superior to impersonal models even when built from a very small amount of personal training data.

Personalized AR models are quite feasible for smartphone-based AR since a smartphone typically is used by a single user and because little training data is required. Thus, we advocate for the development of personal AR models. We have in fact incorporated the ability to generate such models into our publically available AR app, Actitracker [1]. This app allows a user to quickly perform self-training and then replaces the default impersonal model with an automatically generated personal one.

2 The Activity Recognition Task.

The activity recognition task involves mapping time-series accelerometer data to a single physical user activity. We formulate this as a standard classification problem by aggregating the time-series data into examples. We consider six common activities that collectively cover much of the time in a typical user's day: walking, jogging, stair climbing (up and down), sitting, standing, and lying down. We assume the smartphone is in the user's pocket. However, we believe that an AR system could easily learn the body location either in order

*This material is based upon work supported by the National Science Foundation under Grant No. 1116124.

†Department of Computer and Information Science, Fordham University. (lockhart,gweiss)@cis.fordham.edu

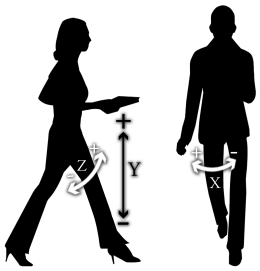

Figure 1: Orientation of accelerometer axes

to use a position-specific model, or as an implicit part of the original learning problem. The axes associated with the smartphone are aligned as indicated in Figure 1. The accelerometer measures the acceleration due to gravity, about $9.8m/s^2$, and this is incorporated into the y-axis values for activities where the phone is upright.

3 Experiment Methodology.

In this section we describe the methodology for generating AR models. We discuss data collection procedures, the method for transforming the accelerometer data into examples, and the model induction process. We also describe the methodology for generating and evaluating personal and impersonal models, as well as a hybrid model that combines elements of these two models.

3.1 Data Collection. We collected data by having 59 users carry an Android-based smartphone in their pocket while performing the six everyday activities mentioned previously. Our research team members directed the participants to perform the various activities and input the activity labels into our data collection app. The sensor data is stored on the phone and also transmitted to our server. For this study, we use a sampling rate of 20Hz. We used fifteen different Android smartphone models in our study and all of their accelerometers appeared to generate similar results.

Additionally, we analyzed the data from the HASC 2010 and 2011 data sets [9]. In order to apply the same methods to this data, we selected only the accelerometer data from pocket or waist located sensors and downsampled the readings from 100Hz to 20Hz. The resultant subset included 414 people performing 4 activities (standing, walking, running, and skipping). The results for the two data sets are presented separately, since they represent different activities and are recorded with different hardware/software.

3.2 Data Transformation. Our classification algorithms cannot directly handle time-series data, so we begin by transforming the raw accelerometer data into examples following the procedure described in our prior work [15]. Each example summarizes 10 seconds of data,

Total	Walk	Jog	Stair	Sit	Stand	Lie
$n = 9291$	3397	1948	1549	1143	689	565
100%	36.6	21.0	16.7	12.3	7.4	6.1

Table 1: Number of examples per activity over the data we collected from 59 users

an interval sufficient to capture several repetitions of periodic motions and empirically shown to perform well. Each example contains 43 features that are variations of 6 basic statistics.

Table 1 shows the number and distribution of the transformed examples, per activity, for our data. Walking is the most common activity. The time spent jogging and stair climbing was limited because these activities are strenuous, and we limited the time spent on the static activities because we found they were easy to learn. By comparison, after transformation the HASC set contains 10,718 examples from 414 users, with the following distribution: Walk - 59.9%, Run - 13.0%, Stand - 15.1%, and Skip - 13.0%. This is a different and smaller set of activities than in our data, and the class distribution is much more skewed.

3.3 Model Induction and Experiments. Our AR models are induced from the labeled examples using the following WEKA [25] classification algorithms: decision trees (J48 and Random Forest, RF), instance-based learning (IBk), neural networks (Multilayer Perceptron, NN), rule induction (J-Rip), Naïve Bayes (NB), and Logistic Regression (LR). Default settings from WEKA are used for all learning methods except NB, where kernel estimation is enabled, and IBk, where we set k=3 (IB3) so we use 3 nearest neighbors.

We induce three types of models: impersonal, personal, and hybrid. Each model addresses a slightly different learning problem and makes different assumptions about how the model is applied. The type of model impacts how we partition the data into training and test data. The different models are defined below:

Definition 3.1. *Impersonal models* use training data from a panel of users that will not subsequently use the model (thus the training and test sets have no common users). These models are applied to a new user without additional labeled training data or model regeneration.

Definition 3.2. *Personal models* use training data from only the user for whom the model is intended. These models require a training phase to collect labeled data from each user. The training and test data come from the same user, but contain distinct examples.

Definition 3.3. *Hybrid models* are a mixture of impersonal and personal models. The training set has data

from both the test subject and other users, but the test set's examples are distinct.

Impersonal models have the advantage that they can be built once for all users and can include data from many users for training purposes. These models can be viewed as universal, although technically they should only be used for users not in the training set. Personal models have the advantage that they may match the idiosyncrasies of the intended user, but require each user to provide training data, limiting the amount of data available. The hybrid model also requires training data and model generation for each user, but can potentially outperform the personal model because it utilizes additional training data from other users.

The experiments associated with each model vary in how they are set up. For impersonal models, data from 58 users is placed into the training set and data from 1 user is placed into the test set. This process is repeated 59 times, which allows us to generate reliable performance metrics and characterize the performance on a per-user basis. For personal models, 10-fold cross validation is applied to each user's data and thus 590 ($59 * 10$) personal models are evaluated. Since each user has a very limited amount of data (on average 160 examples), 10-fold cross validation is essential. The confusion matrices from both types of models are created by summing the counts in each cell over all 59 runs. The setup for the hybrid models is much simpler: we place all of the user data into a single file and then apply 10-fold cross validation. Thus, the hybrid training and test sets have overlapping sets of users.

To generate the learning curves for personal and hybrid models, we generate k folds for each user (where k is the total number of examples a user has divided by the number of examples we want in the training set). This required the generation of tens of thousands of models. The results for all folds of each user are combined, and then the results for all of the users are averaged. To generate the learning curves for impersonal models, it was impractical to generate every possible combination. Instead, we randomly selected the required number of users and then randomly selected the required number of examples from each selected user to build training sets. This process was repeated 50 times, and the results are averaged. The results we present for the HASC data set were generated using the same methodology.

4 Results.

In this section we present the results of our analysis for the two data sets (ours and HASC). We focus primarily on the results for our data because it represents a more complex set of activities and because it includes more detailed information about the participants. In key places, such as where we evaluate the effect of an increased number of users, we present more detailed findings for the HASC data.

	RF	NB	LR	IB3	NN	J48	J-Rip	Avg.
Personal	98.4	97.6	97.7	98.3	98.7	96.5	95.1	97.4
Hybrid	95.0	82.8	84.6	96.5	92.1	91.8	91.1	88.7
Impersonal	75.9	74.5	72.7	68.4	67.8	69.1	70.2	70.9
Average	89.8	85.0	85.0	87.7	86.2	85.8	85.5	85.7

Table 2: Predictive accuracy of activity recognition

The predictive accuracy associated with the personal, hybrid, and impersonal models on our data set is displayed in Table 2. These results make it quite clear that for every classification algorithm the personal models perform best, the hybrid models perform second best, and the impersonal models perform worst. Furthermore, the personal models always achieve a very high level of accuracy and perform dramatically better than the impersonal models. While this result may seem easy to justify, since people move differently from one another, the result is far from obvious since the personal models are trained from dramatically less data.

The hybrid models typically perform closer to the personal models than the impersonal models. Given how well the personal models do, this is a bit surprising. It implies that the hybrid models can make effective use of the personal data in the data set, even though only a small fraction of the data (on average 1/59) is personal data. This means that the classification algorithms can effectively identify the movement patterns of a particular user from among a host of users. In retrospect this is not so surprising, since our recent work has shown that biometric models induced from accelerometer data can identify a user from a set of users with near perfect accuracy [14]. Because the hybrid model performs more poorly than the personal model, but still requires the acquisition of labeled training data from the target user, there seems little reason to utilize the hybrid model. The only exception might be when the amount of personal data is extraordinarily small, thus increasing the importance of the relatively common impersonal data. But as we see later in this section, even when there is very little personal data the personal models outperform the impersonal models. This result surprised us, but the real surprise is just how effective personal data is, even in extremely small quantities.

The main focus of this paper is on the comparative performance of the three types of AR models, but our results also suggest which classification methods may be best suited to AR, given our formulation of the problem (see the underlined values in Table 2). For personal models, NN does best, although RF and IB3

	% of Examples Correctly Classified						
	Personal			Impersonal			Base
	IB3	RF	NN	IB3	RF	NN	Line
Walk	99.1	98.9	99.0	65.2	73.0	56.8	36.6
Jog	99.5	99.6	99.8	89.0	95.2	92.1	21.0
Stairs	96.4	96.8	98.0	65.1	61.5	68.0	16.7
Sit	98.2	98.7	98.1	67.6	81.5	66.7	12.3
Stand	86.4	97.8	97.5	75.2	91.9	88.0	7.4
Lie	95.9	95.0	97.5	34.0	45.1	45.5	6.1
Overall	98.3	98.4	98.7	68.4	75.9	67.8	36.6

Table 3: Predictive accuracy on a per-activity basis

also perform competitively; for hybrid models IB3 does best but RF performs competitively; for impersonal models, RF does best. Averaged over the three types of models, RF does best. Due to space considerations, many of our detailed results focus only on RF, IB3, and NN, the three best performers.

The personal results in Table 2 are mirrored by our results on the HASC data (results not shown). Using RF, the personal accuracy is 97.2%. However, the impersonal accuracy is higher than for our data set, at 85%. We examine the causes of this discrepancy more thoroughly throughout this section.

4.1 Accuracy by Activity. Table 3 shows the AR performance for the personal and impersonal models for each activity, using the three best-performing classification algorithms and a baseline strategy. The baseline strategy always predicts the specified activity, or, when assessing overall performance, the most common activity. The baseline allows us to consider class imbalance. Personal models outperform impersonal models for every activity, usually by a substantial amount, although impersonal models still outperform the baseline.

Table 4 provides the confusion matrices associated with the Random Forest learner for the impersonal and personal models. These results show that most of the errors, for both the impersonal and personal models, are the result of confusing walking with stairs and lying down with sitting.

The confusion between walking and stairs may be due to the similar time between footsteps and exacerbated by the differences that individual people exhibit when performing these activities (jogging probably does not exhibit this problem due to the shorter time between footsteps and the more extreme acceleration values). It is easy to see why lying down and sitting are confused, since the orientation of one's pocket will be similar for both of these stationary activities. While the results for personal models in Table 4b show that these activities are still confused the most, the frequency of such errors

(a) Impersonal	Predicted Class					
Actual Class	Walk	Jog	Stairs	Sit	Stand	Lie
Walk	**2480**	66	819	22	8	2
Jog	51	**1845**	41	1	0	1
Stairs	518	69	**593**	2	4	3
Sit	7	5	3	**931**	19	178
Stand	3	0	12	19	**633**	22
Lie	7	0	5	284	14	**255**

(b) Personal	Predicted Class					
Actual Class	Walk	Jog	Stairs	Sit	Stand	Lie
Walk	**3359**	3	30	1	3	1
Jog	5	**1940**	3	0	0	0
Stairs	40	5	**1500**	2	2	0
Sit	3	0	1	**1128**	2	9
Stand	5	0	8	2	**674**	0
Lie	3	2	4	18	1	**537**

Table 4: Per-activity confusion matrices

is reduced by more than a factor of 10. *This indicates that it is possible to learn the user-specific differences in these two sets of activities and that these differences are not the same for all people.* This is a key argument for the use of personal models and perhaps the most important conclusion of this paper.

The confusion between sitting and lying down is not a factor with the HASC data set, where standing is the only stationary activity. This strongly contributes to the increased performance of impersonal models on the HASC set; the most difficult to identify classes are missing. For this reason, we focus more on our own, more challenging, data set than on the HASC set.

4.2 Accuracy by User. The results presented thus far are averages over all users. However, it is informative to see how AR performance varies between users. Figure 2 provides this information for the personal models and shows that these models perform consistently well for almost all users. The minor outliers that do show poor performance are primarily due to the high levels of class imbalance in those users' data. For example, the user with the second highest error rate has 59 examples of walking data but only between 5 and 8 examples for each of the other activities. The user with the worst accuracy had a similar class distribution, and also had a leg injury. Thus, the few problems that do occur for the personal models appear to be due to high levels of class imbalance or due to an injury.

Figure 3 shows a much broader distribution of performance for the impersonal models. There are still some users with classification accuracies in the

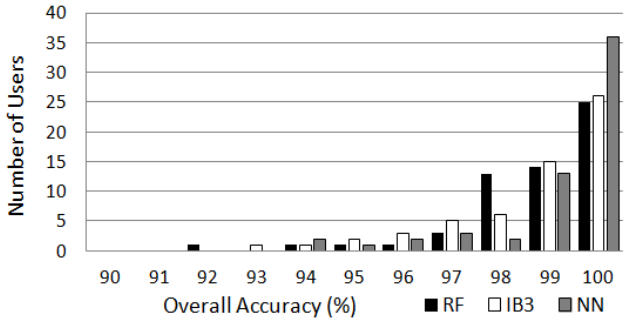

Figure 2: Personal model accuracy distribution across users

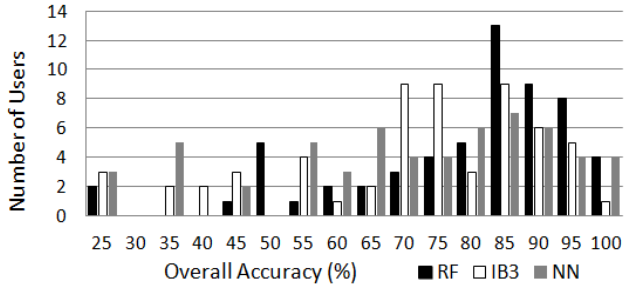

Figure 3: Impersonal model accuracy distribution across users

95-100% range, but the accuracies vary widely and there are some users with extremely low accuracies. Our detailed analysis showed that most of these very poor-performing users performed quite well when using personal models. These results further support the view that there are many users who move differently from other users, which confuses the impersonal models, while the personal models can learn these user-specific differences to achieve consistently good results.

As part of our data collection protocol, we collect information about the physical characteristics of each user (height, weight, sex, shoe size, etc.). We analyzed this information to determine if people with particular or extreme characteristics are especially hard to predict for impersonal models, but partly due to the limited number of users, we could find only suggestive patterns. For example, of the 10 users that were hardest to predict using the impersonal RF models, 3 were among the oldest users in the study. In the future we plan to collect data from substantially more users so that we can better assess the impact of such factors.

4.3 Augmented Impersonal Models. In an effort to better understand the relationship between these traits and the performance of our models, we repeated the experiments for the impersonal models several times, each time augmenting the training and testing sets with an additional attribute of personal information (viz. height, weight, shoe size, and sex). Such information requires less effort to obtain than labeled training data, and thus presents the possibility of building more targeted impersonal models, which may have some of the benefits of personal models, but without the cost. Additionally, we matched users based on the similarity of their transformed accelerometer data and constructed another set of models using data from the most similar users. These are alternate versions of the protocols used in prior research [16].

Where the prior work showed improvement by restricting impersonal models to similar users, this improvement was limited and only brought the models' accuracies per activity to rates similar to those of our impersonal models in Table 3 [16]. Because our users are already very similar (primarily college students), while other studies' subjects are less homogenous, it is not very surprising that on our data set these techniques do not improve accuracy. This suggests the poor performance of models that are not trained from labeled personal data is the result of differences that cannot be explained entirely by differences in user demographics, or compensated for by the similarity of unlabeled data. We conclude that simple demographic information, or even unlabeled data, is no substitute for labeled accelerometer data from that user, which may encode idiosyncrasies with the subject's movements.

4.4 Accuracy by Quantity of Training Data. In the context of this paper, learning curves can be particularly insightful because the varying amount of training data may impact model types differently and acquiring labeled personal data can be quite costly. We begin by analyzing the learning curves for the personal and hybrid models, presented in Figures 5 and 6. Thes figures show that personal and hybrid models improve their performance rapidly. Figure 5 shows that with only 20 seconds (2 examples) of personal data, all three classifiers clearly outperform impersonal models.

Furthermore, our results show that after 2 minutes (only 12 examples) of labeled data for each activity from a user, RF models reach 98.7% accuracy. With 3 minutes of data, this increases to 99.2% and at 5 minutes we reach 99.6%. But the key takeaway here is that we need only a miniscule amount of personal data—one example per activity—in order to outperform an impersonal model built using far more data.

To generate the same curves for hybrid models, we used the impersonal model training sets as a base and added in the data from the personal model training sets used to generate Figure 5. The resulting training sets included all available impersonal data and the amount of personal data specified on the x-axis. Figure 6 shows a similar, though less dramatic pattern for hybrid models as we saw with personal models. With 10 seconds of

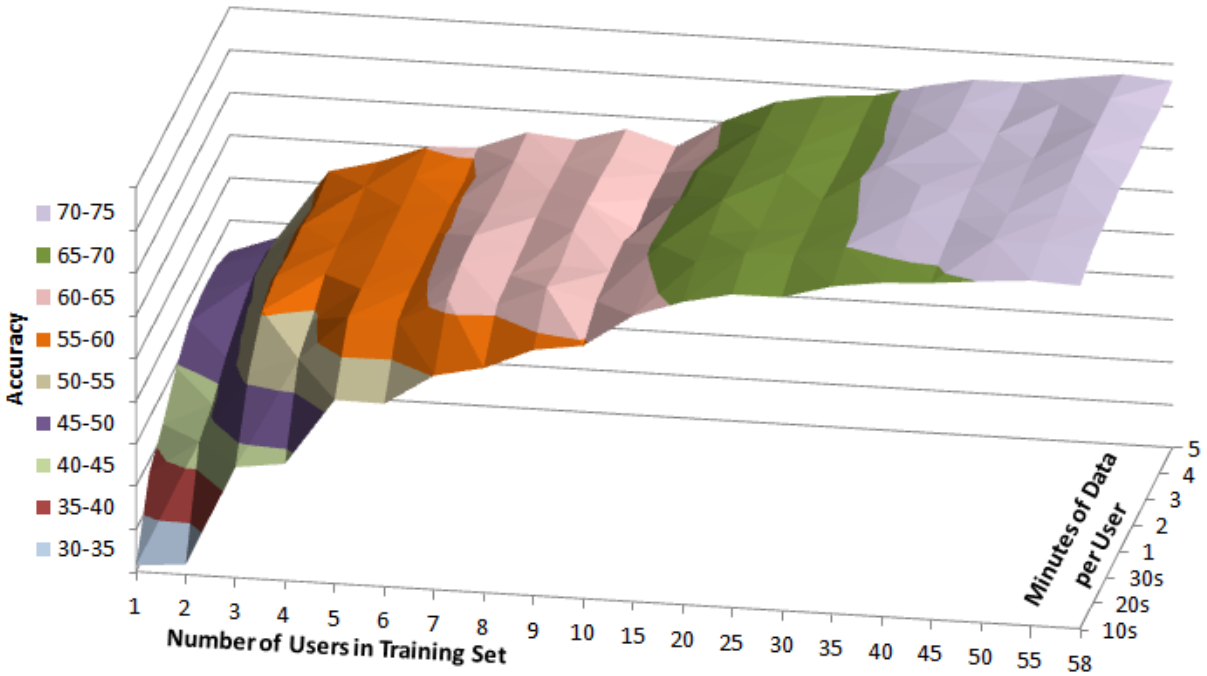

Figure 4: Learning curve for impersonal random forest models

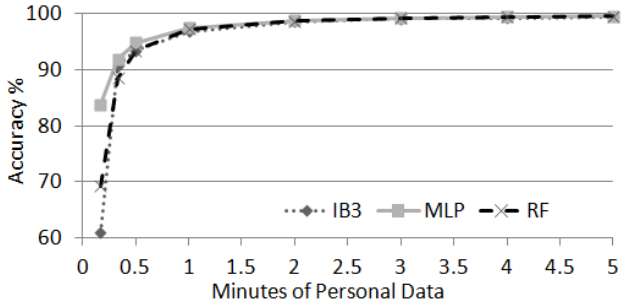

Figure 5: Learning curves for personal models

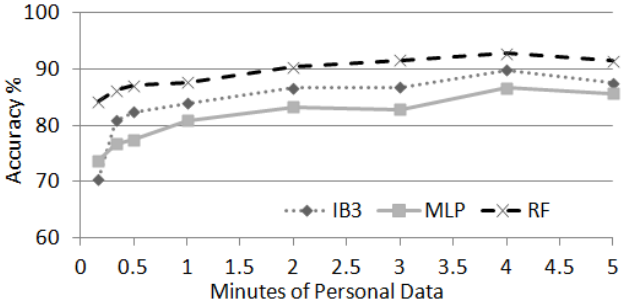

Figure 6: Learning curves for hybrid models

labeled personal data for each activity, two out of our three hybrid models outperform personal models, and all outperform impersonal ones. From 20-30 seconds and beyond, personal models dramatically outperform hybrid ones. Impersonal models have an additional factor to consider, however: we can vary both the amount of data from each user and the number of users in the training set. Figure 4 shows these two factors plotted with classifier accuracy in 3–dimensions. The surface is shaded to make its accuracy dimension clearer.

Figure 4 allows us to make several important observations, although some of these may be difficult to see at a quick glance, given that there are three dimensions involved. Like with personal and hybrid models, there is little to no improvement in accuracy from including more than two minutes of training data from each user. However, when there are few users in the training set, including more data per user (up to 2 minutes) dramatically improves performance. As the number of users in the training set increases, the value of additional data from each user decreases. With 58 users in the training set, there is little difference between using 10 seconds or 5 minutes of data per user per activity. Thus, the best way to improve the accuracy of impersonal models is to increase the number of users in the training set, rather than increasing the amount of data per user—this is another key lesson. This tradeoff is dramatic: a model generated from 5 users with 10 seconds of data per activity outperforms one generated from 2 users with 60 seconds of data per activity—even though the first data set has less than half the total data. Figure 4 also shows us that accuracy has not reached a plateau, and hence having data from more than 58 users will yield improvements in accuracy.

In order to understand this trend better, we generated the same curves for the HASC data set, which has seven times as many users. The three-dimensional chart for this set has the same shape as Figure 4, ex-

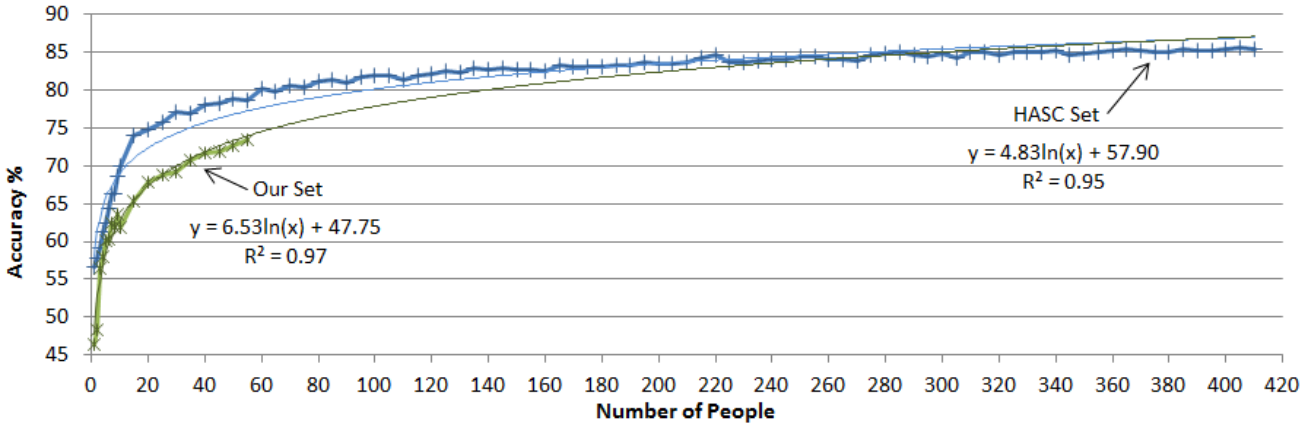

Figure 7: Learning curves for impersonal random forest models on data from our set and the HASC set

cept shifted slightly up due to the simpler activity set and greater class imballance. For ease of comparison, we show the HASC curve side by side with our own in Figure 7. Beyond 210, the increase in accuracy of impersonal models from additional users is negligible.

We also find that the performance gap between the same algorithms on the two data sets decreases as we increase the number of users in the set. With only one user, the gap is 10 points (46% to 56%), but at 55 users the gap is only 5 points (73% to 78%). Further, we fitted a logarithmic function, $f(x) = 6.53 * ln(x) + 47.75$, with $R^2 = 0.97$, to the accuracy of impersonal models on our data set so we could compare it to the HASC set for greater numbers of users. The fitted curves for each set are shown as the thinner, smooth lines in Figure 7. Based on this function, increasing the number of users from 58 to 200 would bring the average accuracy to 82%, while the accuracy on the HASC set at 200 users is only 83%. Of course, these projections may not hold.

In any case, it is clear that personal and even hybrid models will substantially outperform impersonal models, while using dramatically less data. We anticipate that no number of users would allow an impersonal model to outperform our personal models, which are based on about 24 minutes of data divided over six activities. One of our key findings is that even impersonal models using data from a homogenous group are limited by the diversity of gait patterns, an obstacle easily overcome by building personal models.

5 Related Work.

There has been prior work related to personal, impersonal, and hybrid models of AR, although in virtually all cases the work has not had these topics as their primary focus. Several AR studies only analyze impersonal models, with limited results [5, 8, 20]. Other AR systems using smartphones have achieved relatively higher accuracy, but used only personal models [5, 19, 28]. One paper described personal models that could be incrementally trained to adapt to changes in a user's gait [2]. However, their results were similar to those of our impersonal models, indicating that feature selection and choice of algorithms is still an important part of AR.

The majority of other smartphone-based systems only evaluate hybrid models [3, 7, 15, 27], and a number of other studies [6, 10] do not provide enough information in their methodology to determine the model type. We view this as problematic because of the dramatic difference in performance and applications between different model types. In one study of Parkinson's disease patients, the researchers conclude that models trained on healthy people perform more poorly on people with Parkinson's disease [27]. However, their methodology shows that they used hybrid models when evaluating their accuracy on healthy subjects and impersonal models when evaluating their accuracy on Parkinson's patients. Our work shows that the discrepancy in their results is likely due, at least in part, to their mixing of model types and not due to differences between Parkinson's patients and healthy people.

There has been relatively little comparative analysis of the different types of AR models. Two studies did compare personal and impersonal models, but both employed five accelerometers—thus any conclusions would not necessarily apply to a smartphone–based system. The first of these studies concluded that impersonal models always outperform personal ones, due to the additional training data; it further showed that when the impersonal and personal training set sizes are equalized, personal models only slightly outperform impersonal ones [4]. Our results clearly and dramatically contradict this result (but for smartphone-based systems). In the second study the personal models outperform the impersonal models but there is virtually no analysis or discussion of this, as it is not the paper's focus [23].

One paper we have already discussed [16] recognizes

the poor performance of some impersonal models and develops methods for improving their accuracy by selectively training on similar users. However, our results show that beyond a certain point (i.e., the point they [16] reached and the point we start from with our set of similar users), user similarity is not able to compensate for the differences in gait. Further, our analysis of the learning curves for various model types shows that very small amounts of personal data dramatically outperform these best-case impersonal models, and that no amount of data for impersonal models will perform competitively with personal models. Thus our paper is the most comprehensive study on the impact of model-type on AR—especially as it relates to single accelerometer, smartphone-based systems. Furthermore, we additionally evaluate hybrid models and include many more users in our study than prior studies, as well as more activities which are more challenging to distinguish, leading to more reliable, and general, results.

Many studies use very limited datasets, often with fewer than 5 users [17, 28] or 10 users [2, 7, 10]. Compounding the issue, the most widely used AR datasets, COSAR and OPPORTUNITY, have data from only 4 and 12 users, respectively [21, 22]. Larger sets, such as HASC 2010 and 2011, contain simplified sets of activities and smaller amounts of data. This motivated us to release our AR dataset [24].

6 Conclusion and Future Work.

In this paper we describe and evaluate a data mining approach for implementing activity recognition, using only smartphones. We demonstrate that nearly perfect results can be achieved if a personalized model is constructed, even using only a very small amount of user-specific training data. We further show that impersonal models perform much worse than personal models, even under the best conditions where they are trained on similar users. Analysis of the data shows that impersonal models cannot effectively distinguish between certain activities, whereas personal models can effectively learn the user-specific differences that confound impersonal models. We also show that while the poor performance of impersonal models is affected by some idiosyncratic users, whose activities cannot be accurately predicted, the problem is widespread and not restricted to a few problem users. Although there appears to be room for marginal improvement in the accuracy of impersonal models by increasing the number of people in the training set up to about 210, and related work [16] shows that selecting similar subsets of users also increases the accuracy of impersonal models to a point, it is clear that the personal models are able to substantially outperform impersonal ones while using only a very small amount of personal data. We also showed that if one does want to improve the performance of impersonal models, it is far better to obtain more users than to obtain more data per user.

In this paper we also evaluated the performance of hybrid models and showed that their performance was generally inferior to personal models but consistently better than that of even best-case impersonal models constructed with similar users. Given that hybrid models require user-specific training data, and personal models with the same amount of data generally perform better, one is better off using personal models. The one exception is the extreme case where we have only 10 seconds of data from a user, but that case is unlikely, since anyone already gathering personal data would easily be able to gather more than 10 seconds of it from healthy people. Thus we conclude that, hybrid models will never realistically outperform personal models.

This study provides the most thorough analysis of the relative performance of personal, impersonal, and hybrid AR models, and we view this as a key contribution of the paper (along with the various lessons learned that we listed above). This is also the first study to examine in great detail the effect of training set size on AR performance, as it varies by number of unique users and quantity of data per user. This work should greatly influence the design of future AR systems and the higher level activity and context representation systems which rely on them.

One of our key achievements is making this AR research available to smartphone users and researchers alike, via our downloadable app, Actitracker [1]. Our AR system tracks a user's activities and provides reports via a secure account and web interface. This mobile health application helps people ensure that they and their children are sufficiently active to maintain good health and avoid the many health conditions associated with inactivity. Because of the results of this research, we have included a self-training mode in our system, so one can quickly generate personal labeled activity data to achieve good predictive performance. This data is uploaded to our server, which automatically generates a personalized AR model that then replaces the impersonal model for that user.

In the future, we plan to extend our activity recognition work in several ways. We continue to collect data and will add many more users. We also plan to collect labeled data from users weeks and months after their personal models are built, to evaluate the impact that time and different clothes and footwear have on activity recognition. We will also cover additional activities and utilize additional sensors, especially the gyroscope, which is becoming standard in most smartphones.

References

[1] actitracker.com

[2] Z.S. Abdallah, M.M. Gaber, B. Srinivasan, and S. Krishnaswamy, *StreamAR: incremental and active learning with evolving sensory data for activity recognition*, In Proc. 24th IEEE International Conference on Tools with Artificial Intelligence, 2012.

[3] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, *Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine*, In Proc. International Workshop of Ambient Assisted Living, 2012.

[4] L. Bao, and S. Intille, *Activity recognition from user-annotated acceleration data*, Lecture Notes Computer Science 3001, pp.1–17, 2004.

[5] T. Brezmes, J.L. Gorricho, and J. Cotrina, *Activity recognition from accelerometer data on mobile phones*, In Proc. 10th International Work-Conference on Artificial Neural Networks, pp. 796–799, 2009.

[6] Y. Cho, Y. Nam, Y-J. Choi, and W-D Cho, *SmartBuckle: human activity recognition using a 3-axis accelerometer and a wearable camera*, HealthNet, 2008.

[7] S. Dernbach, B. Das, N.C. Krishnan, B.L. Thomas, D.J. Cook, *Simple and complex activity recognition through smart phones*, In Proc. 8th International Conference on Intelligent Environments, pp. 214–221, 2012.

[8] N. Gyorbiro, A. Fabian, and G. Homanyi, *An activity recognition system for mobile phones*, Mobile Networks and Applications, vol. 14, no. 1, pp. 82–91, 2008.

[9] N. Kawaguchi, et al., *HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings*, In proc. 2nd Augmented Human International Conference, pp. 27, March 2011.

[10] A.M. Khan, Y.K. Lee, S.Y. Lee, and T.S. Kim, *Human Activity Recognition via an Accelerometer-Enabled-Smartphone Using Kernel Discriminant Analysis*, In Proc. 5th International Conference on Future Information Technology, pp. 1–6, 2010.

[11] J.P. Koplan, C.T. Liverman, and V.I. Kraak, *Preventing childhood obesity: health in balance*, National Academies Press, Washington DC, 2005.

[12] N.C. Krishnan, S. Panchanathan, *Analysis of low resolution accelerometer data for human activity recognition*, In Proc. International Conference on Acoustic Speech and Signal Processing, ICASSP, 2008.

[13] N.C. Krishnan, D. Colbry, C. Juillard, S. Panchanathan, *Real time human activity recognition using tri-axial accelerometers*, In Proc. Sensors Signals and Information Processing Workshop, 2008.

[14] J.R. Kwapisz, G.M. Weiss, and S.A. Moore, *Cell phone-based biometric identification*, In Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems, 2010.

[15] J.R. Kwapisz, G.M. Weiss, and S.A. Moore, *Activity recognition using cell phone accelerometers*, SIGKDD Explorations, vol. 12, issue 2, pp. 74–82, 2010.

[16] N.D. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A.T. Campbell, and F. Zhao, *Enabling large-scale human activity inference on smartphones using community similarity networks (CSN)*, In proc. 13th International Conference on Ubiquitous Computing (UbiComp 2011), pp. 355–364, September 2011.

[17] Y. Lee, and S. Cho, *Activity recognition using hierarchical hidden markov models on a smartphone with 3D accelerometer*, In Proc. 6th international conference on Hybrid artificial intelligent systems, pp. 460–467, 2011.

[18] U. Maurer, A. Smailagic, D. Siewiorek, and M. Deisher, *Activity recognition and monitoring using multiple sensors on different body positions*, In Proc. IEEE International Workshop on Wearable and Implantable Sensor Networks, vol. 3 no. 5, 2006.

[19] E. Miluzzo, N.D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S.B. Eisenman, X. Zheng, A.T. Campbell, *Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application*, In Proc. 6th ACM Conference on Embedded Networked Sensor Systems, pp. 337–350, 2008.

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

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

[22] D. Roggen, *et al.*, *Opportunity: towards opportunistic activity and context recognition systems*, In Proc. 3rd IEEE WoWMoM Workshop on Autononomic and Opportunistic Communications, 2009.

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

[24] www.cis.fordham.edu/wisdm/dataset.php

[25] I.H. Witten, and E. Frank, *Data mining: practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, 2005.

[26] World Health Organization, *Physical inactivity a leading cause of disease and disability, warns WHO*, 2002, www.who.int/mediacentre/news/releases/release23

[27] W. Wu, S. Dasgupta, E.E. Ramirez, C. Peterson, G.J. Norman, *Classification accuracies of physical activities using smartphone motion sensors*, Jurnal of Medical Internet Research, vol. 14 no. 5, 2012

[28] J. Yang, *Tooward physical activity diary: motion recognition using simple acceleration features with mobile phones*, in Proc. 1st International Workshop on Interactive Multimedia for Consumer Electronics at ACM Multimedia, 2009.