Personalized Online Training for Physical Activity monitoring using weak labels

Federico Cruciani, Ian Cleland,
Chris Nugent and Paul McCullagh
School of Computing
Ulster University
Jordanstown, Northern Ireland, UK

Kåre Synnes and Josef Hallberg
Dept. of Computer Science, Electrical and Space Engineering
Luleå University of Technology
Luleå, Sweden

***Abstract*—The use of smartphones for activity recognition is becoming common practice. Most approaches use a single pre-trained classifier to recognize activities for all users. Research studies, however, have highlighted how a personalized trained classifier could provide better accuracy. Data labeling for ground truth generation, however, is a time-consuming process. The challenge is further exacerbated when opting for a personalized approach that requires user specific datasets to be labeled, making conventional supervised approaches unfeasible. In this work, we present early results on the investigation into a weakly supervised approach for online personalized activity recognition. This paper describes: (i) a heuristic to generate weak labels used for personalized training, (ii) a comparison of accuracy obtained using a weakly supervised classifier against a conventional ground truth trained classifier. Preliminary results show an overall accuracy of 87% of a fully supervised approach against a 74% with the proposed weakly supervised approach.**

***Index Terms*— data annotation, weakly supervised learning, smartphone activity recognition.**

I. Introduction

The use of smartphones as unobtrusive devices to perform Human Activity Recognition (HAR) is becoming increasingly prevalent in the assistive technology literature [1] [2]. These devices were once considered incapable of running HAR systems in real-time. This was either due to limited resources in terms of computational capabilities, or because of limited battery life when used to perform continuous monitoring. Within the last few years, however, advancement in technology has brought smartphones to a new level of sophistication, making them ideal candidates for HAR. This is primarily due to the availability of a wide range of on-board sensors. Furthermore, smartphones are perceived by users as unobtrusive devices despite their pervasiveness. This trend is confirmed by major mobile OSs, offering HAR functionalities within their APIs [3]. Sensor based activity recognition, in particular, the use of inertial sensors for HAR has been deeply investigated [4]. When moving proposed HAR approaches to an embedded solution, however, a number of limitations arise. For example, in most cases, studies have focused on wearable sensors (with sensor location known *a priori*) and developed solutions are not position independent [1] [2]. The assumption that users will carry the smartphone in a predefined position, however, is not valid in a free-living setting. Another major limitation resides in the lack of personalization, i.e. solutions are trained *offline* (often with data collected in controlled environments) with a single classifier for all users, although some users may have very different behaviors than others [1], and a personalized approach has shown better results [5]. Some solutions have been proposed to tackle the sensor positioning problem, however, very few studies have explored the use of smartphones for *online* training (i.e. training locally on the smartphone) and *personalized* training (i.e., training a user-specific classifier) [2]. This is mainly due to the hurdle of generating the ground truth. Unless an efficient ground truth generation method can be achieved, personalization is not viable using supervised learning approaches for HAR. Data annotation methods have been widely discussed to facilitate ground truth collection, and different approaches have been proposed. One example is tools to support manual annotation of datasets [6] [7] in order to (at least partially) automate the process. These tools can speed up the process of data annotation, however, the manual labeling is still required making this approach not viable for a personalized approach. The use of smartphones for data annotation has also been proposed. The advantage in this case is that users can directly annotate data fragments [8], yielding a more flexible approach. The fact that human interaction is still required at some stage though, makes the process time consuming for the users.

Among data-driven approaches towards HAR, it is worth mentioning that unsupervised approaches have also been proposed. These approaches do not require the availability of large labeled datasets. Nevertheless, in this study, we focus on personalization of supervised approaches, and specifically on exploring weak supervision to remove the burden of generating a user specific ground truth. In this paper, we propose a weakly supervised approach (i.e. trained using weak labels generated via a heuristic), and evaluate its performance in terms of accuracy against a conventional supervised learning method that requires manual ground truth generation.

The rest of the paper is structured as follows. Section II highlights related works, trends and limitations of the state of the art. Section III introduces the proposed approach, while the adopted methodology for evaluation is explained in section IV. Section V and VI provide details on preliminary results comparing weakly supervised approach to a fully supervised one. Finally, section VII summarizes possible directions for

978-1-5386-3227-7/18/$31.00 ©2018 IEEE

future work.

II. Related Work

The research field of sensor based HAR has been widely investigated in recent years [2] [9]. Yet, considering the computational capabilities offered by modern smartphones, there is scope for a new line of investigation towards *online* and *personalized* training using mobile devices offering the potential of improved levels of recognition performance. In this respect, personalized means a tailored approach, trained on user specific data. Similarly, by online we mean that some part of the classification workflow can be performed in real-time and locally on the smartphone. This is opposed to online machine learning, where the term online identifies techniques that can adapt to new available data points in the training set [2]. In this section, we analyze solutions proposed in the literature and their viability for an online personalized training approaches using smartphones. Sensor based HAR is an advanced and broad field of investigation. Among the wide range of proposed sensing modalities, inertial motion sensors (accelerometer, gyroscope) have been widely used [2] [9]. In smartphone based HAR, together with inertial sensors, it is more common to include other on-board information sources, as GPS [2]. A solution focusing solely on accelerometer has, however, been considered as optimal in terms of tradeoff between accuracy and battery consumption, and therefore adopted in many cases [1]. More common features used for accelerometry HAR are in the time domain (e.g. mean, variance and min-max range) [1] [2]. Frequency domain features have also been used. Although similar features may not be ideal when trying to develop a smartphone embedded solution continuously monitoring activity. This is mainly due to the complexity of the Discrete Fourier Transform (DFT) calculation, making it resource intensive [1]. When the sensor position and orientation is known *a priori*, the feature extraction on the three axes of the signal has been proven to be more informative [1] [2]. Unfortunately, this assumption is not valid when considering a smartphone based solution in free-living. To address this, a hierarchical approach has been proposed. This approach aims first to identify location and orientation of the sensor, and then to run the appropriately trained classifier [1] [2]. Alternatively, solutions based on features extracted from the magnitude of the 3D acceleration are utilized in order to have a feature set that is orientation independent [1]. Similarly, sign-invariant features based on the absolute value from the three axes have been used to provide invariance to some orientations [1]. A sampling rate between 20-30 Hz is widely adopted in monitoring physical activity, since most part of the information to monitor physical activities resides in a 10-15 Hz range [1] [2]. Feature extraction is generally realized through a sliding window approach, with a 50% overlap being the most used technique [1]. Generally, window sizes vary between 1 and 10 seconds, with 1-2 seconds being considered optimal [1]. This, however, depends on the subset of specific activities to be detected [1]. In most cases the set of target activities contains activities such as sitting, standing, walking, running, cycling or using some means of transportation [1] [2]. In terms of supervised classification methods Decision Trees (DTs), Support Vector Machines (SVMs) and k-Nearest Neighbors (kNN) are the most common in smartphone based HAR [1]. Neural Networks (NNs) and more recently Convolutional Neural Networkss (CNNs) have also been used [1] [2]. Deep learning approaches such as CNNs have been shown to improve accuracy [1]. The complexity of such approaches, however, make deep learning methods not applicable for embedded smartphone solutions, at least in the short term. Focusing on smartphone based solutions for HAR, the classification or prediction stage is usually performed locally (on the smartphone) and in real-time. Unlike classification, the training phase is, in most cases, performed offline, and one classifier is trained for all users beforehand [2]. This is confirmed by evaluation methods that often are based on leave-one-subject-out for validation [1]. Online solutions using a client-server approach have also been proposed [2]. In this case, features are extracted locally, however, classification is performed remotely on the server-side. These solutions are, however, dependent on a reliable internet connection to operate continuously.

State of the art accuracy in controlled environments for detecting activities such as sitting, standing, walking, running, cycling or use of transportation range between 85-95% depending on the target activities and classification method, however, the accuracy is lower for unknown subjects (i.e. subjects who have not been used to train the classifier) [1]. Moreover, the majority of studies only consider data gathered in controlled environments, although some studies have shown how accuracy can drop significantly (up to 17%) when moving lab-trained classifiers to uncontrolled free-living setting [10]. While much has been undertaken in exploring different techniques and approaches for sensor based HAR, more research is required in the area of personalized and intelligent solutions, able to adapt to the specific user needs [1] [2]. For instance, it has been observed that a generic classifier can work better with some users, however, have poor performance with others [1]. For supervised methods ground truth generation makes the personalized approach unviable. Some alternative approaches, such as user solicitation, have been proposed in order to facilitate the generation of labeled datasets [8]. Similarly, weakly supervised approaches have also been proposed as in [11]. In this case the weak supervision requires to label only a subset of train data points. Label propagation is then used to realize semi-supervised learning. Although this approach significantly reduces the problem of data annotation, some manual labelling is still required in order to collect the ground truth. Moreover, most of the attention so far has been focused only on the final recognition accuracy, without considering aspects as the tradeoff between accuracy and resource consumption (either in terms of battery, computational complexity or memory) [2].

III. Implementation

The proposed approach aims to perform HAR based only on accelerometer data with a sampling rate of 30 Hz. To optimize

the solution for a smartphone embedded scenario, only time domain features have been evaluated, specifically:

- rotation independent features extracted from the magnitude of the acceleration (mean, variance and min-max distance of 3D magnitude of acceleration),
- additional sign invariant features as the absolute value of mean acceleration for the three axes (i.e. absolute value of mean acceleration on X, Y and Z axes).

The proposed training method for classification is *weakly supervised*, i.e. (weak) ground truth generation is based on *weak labels* obtained through an heuristic as in [12], as opposed to *semi-supervised* methods that make use of unlabelled datapoints but still require a subset of labeled data for training [13] [11]. The generation of weak labels will be described in III-A.

As in the case of feature selection, the same approach has been followed to identify suitable classification methods, opting primarily for multi-class computationally inexpensive methods; if not in the training phase, at least in the classification/prediction stage. As such, DTs, NNs and kNN have been identified as potential candidates. Considering that our approach is based on weak labels (i.e. some data points can be mislabeled), kNN have been excluded to avoid direct use of mislabeled samples at prediction stage.

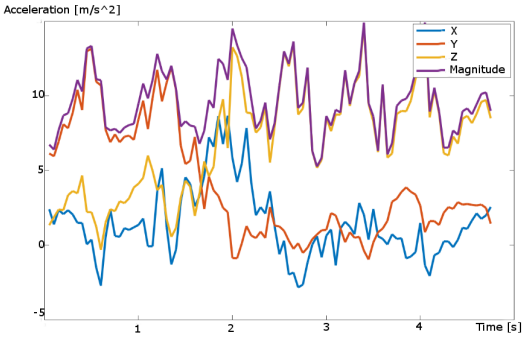
Although NNs are computationally more expensive than DTs, the approach can still be viable to run real-time prediction on the smartphone. Furthermore, NNs allow to update weights when new data points are available, performing online training. Moreover, a pretrained NNs can be used to set initial weights, thus solving the cold-start problem.

This study focused on distinguishing the following physical activities set: sitting, standing, walking, running, cycling and using any means of transportation (i.e. without distinguishing between car, train or bus). The set of target activities has been restricted to sitting, standing, walking and transportation since data gathered in free-living were not representative across all classes. Therefore, the study focused on those classes providing a more balanced dataset, since the same classes are also the more informative to monitor physical activity on our target group (older adults).

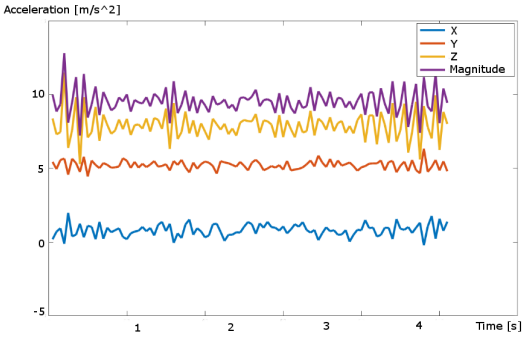
A. *Heuristic function for weak Labeling*

The heuristic used to generate weak labels is the combination of two information sources: GPS and step counter. The GPS provides valuable information that can easily discriminate between walking, running, or using transportation based on estimated speed that can be computed between consecutive timestamped GPS locations (e.g. walking $\simeq$1.4-2.0m/s, running $\simeq$3.0-6.0m/s or transportation $\geq$8m/s). A GPS only based heuristic though, would easily provide false positives (e.g. driving in traffic can result in similar speed to running or walking patterns), or missing information indoor (e.g. running on treadmill at the gym). To reduce the likelihood of having mislabeled samples, the combination with a step counter is used to refine the labeling.

Modern Android smartphones often provide an on-board step counter. A simple step detector has been implemented however, to provide compatibility with a wider range of models. The step counter calculation is performed by computing the magnitude of 3D acceleration. The resulting signal is then pre-processed through a Butterworth low-pass filter (with cut off frequency at 15 Hz as in [14]), and finally a peak detection algorithm identifies local maxima over the resulting signal as steps candidate. Potential steps are then accepted or rejected based on a simple amplitude threshold (i.e. ignoring peaks which are too small) and checking that the corresponding steps per minute (spm) rate resides in acceptable ranges (e.g. walking is typically between 60-110 spm, running 150-180 spm [15]). The final heuristic combines GPS and step counter data in a rule-based intersection of the sources. This allows for instance, to distinguish driving in traffic against a false positive of running because the step counter will not be in the acceptable range, or to detect a workout running session at the gym that would not be detectable via GPS. Similarly, false positive of the step detector can be filtered by using acceptable ranges (in terms of step per minutes rate), and GPS information (e.g. eliminating false positives detected while driving a car). Fig.1 presents an example of data fragments with the corresponding weak label Walking and Transportation assigned using the heuristic.



(a)



(b)

Fig. 1. (a) 5s of acceleromtery raw data with weak label 'walking'. (b) 5s of accelerometry raw data with weak label 'transportation'

B. Architectural view

The HAR framework consists of an Android app that collects data from on-board sensors. A periodic routine retrieves GPS location every 5 minutes, while 5 seconds of accelerometer raw data are collected every 3 minutes. This information rate has been empirically identified as being a good tradeoff between quality of resulting weak labels and the required amount of memory and data to transmit. Collected information is sent to the main server application which provides secure services for authentication, and data storage. Periodically, a Python script will retrieve data for all users and proceed to train the classifier on the server side when a sufficient amount data is available. In this prototype Python scikit-learn library [16] has been used for training the classifier. Once the classifier has been trained the parameters are saved in Predictive Model Markup Language (PMML). The parameters can be sent back to the local Android app. Based on the classifier parameters, the app can update the classification method to be used in real-time instantiating a new classifier loading the PMML model. The overall process is depicted in Fig.2.

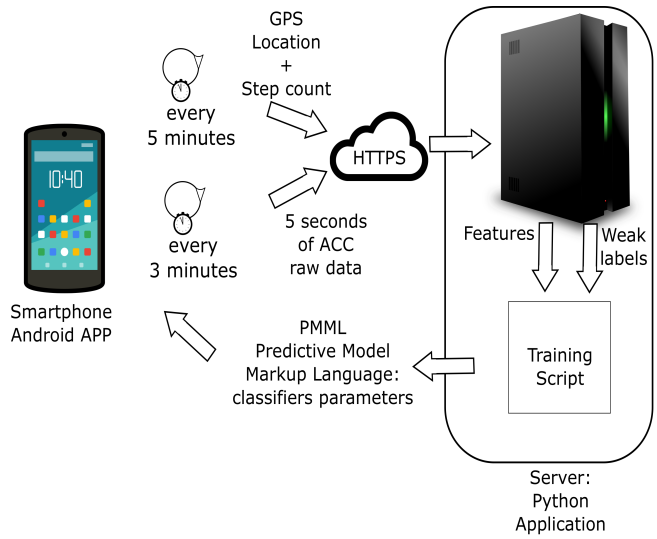


Fig. 2. Architecture for HAR personalized online training. The Android app continuously monitors accelerometry data and saves a 5s fragment every 3 minutes to be sent to the server. Weak labels obtained combining GPS and step counter info are used to train the classifier on the server-side. Parameters of the trained classifier are sent back to the app in PMML format.

C. Classification approach

As most studies focus only on accuracy of recognition methods we explored a computationally inexpensive approach, at least in the prediction stage. The computational cost of training the model is less relevant, since this step is performed on the server side. Consequently, the chosen classifiers were DTs due to their inexpensive cost at the prediction stage and NNs because of the convenience in adapting to new data points.

Feature extraction was realized with a fixed window size of 1 second and a non-overlapping sliding approach (to reduce the overhead of an overlapped method).

IV. Methodology

The experiment has been conducted in two phases. First accelerometry data have been collected for one user for 4 months using the smartphone Android app. During this initial phase, periodic tests of classifier training have been performed to identify a good tradeoff in terms of data to be sent to server and update frequency of the GPS information to reduce the impact on battery consumption. This allowed to identify the final configuration (i.e. GPS sampled every 5 minutes, and 5 seconds of raw accelerometry data sent to the server every 3 minutes).

Two smartphones have been used to evaluate battery consumption in different setup. A Sony Xperia Z3 compact with mobile network activated and other user apps running simultaneously, and a Motorola Moto G with mobile network disabled (i.e. no SIM and only WiFi turned on) and no other apps running.

The final evaluation has been performed on data collected from the smartphone with the identified setup on one subject, acquiring at the same time a manually annotated ground truth. The goal of this second phase was to evaluate the effect on accuracy of a weak supervised approach. One mobile has been used for data collection and kept at a consistent location (trouser pocket), while the second smartphone has been used for data labeling. In order to compare the accuracy obtained with the classifier trained on weak labeled dataset, a ground truth has been manually collected using an ad-hoc Android app. The user would annotate the starting of an activity by pressing the corresponding button as presented in Fig.3. The dataset comprising the manually labeled ground truth was collected in free-living conditions, recording 2-8 hours per day for 10 days for a total of $\simeq$36 hours. Data on physical activity was recorded whilst performing normal Activities of Daily Living (ADLs) such as preparing or having breakfast, commuting to work, and recording during working hours.

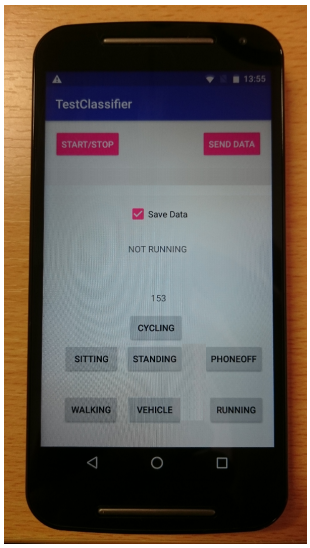


Fig. 3. The Android app use for data labeling.

The subject would annotate the start of a new activity in the transition between actions (e.g. 'sitting' to 'standing' or 'standing' to 'walking'). The final ground truth dataset has been obtained removing 1 second before and after the transition to reduce the uncertainty inherently introduced by the manual annotation.

V. Results

The collected dataset was reduced in order to balance the number of samples for each activity class. This provides a balanced dataset for validation. The final dataset consisted of 8000 samples (2000 per class) used as training (70%) and test data (30%). An additional dataset, consisting of 3000 samples, has been used as validation data. The following results refer to a NN classifier. Similar values have been obtained also both with DT and NN based classification methods with the fully supervised method. In the weakly supervised dataset the NN performed better than the DT across all classes, appearing to be more robust to mislabeled samples. The best accuracy in the weak approach has been measured with a NN with two hidden layers (14 and 10 neurons). Table I presents the resulting confusion matrix for the 4 classes (sitting, standing, walking, transportation). An overall accuracy of 87% has been obtained with a fully supervised approach compared with a 74% accuracy for the weakly supervised method.

TABLE I
CONFUSION MATRIX FOR THE FOUR CLASSES: C1 'SITTING', C2 'STANDING', C3 'WALKING' AND C4 'TRANSPORTATION'.

Fully Supervised				
	C1	***C2***	***C3***	***C4***
C1	**0.9224**	0.0531	0.0041	0.0204
C2	0.0115	**0.8269**	0.1538	0.0077
C3	0.0343	0.0882	**0.8725**	0.0049
C4	0.2308	0.0000	0.0384	**0.7308**
Weakly Supervised				
	C1	***C2***	***C3***	***C4***
C1	**0.7018**	0.1754	0.0702	0.0526
C2	0.1558	**0.7487**	0.0905	0.0050
C3	0.0292	0.1971	**0.7591**	0.0146
C4	0.1282	0.0513	0.0769	**0.7436**

The effect of the continuous monitoring, in terms of battery consumption, has been measured on the two smartphones. The test with both phones has shown how the approach is quite efficient and did not cause significant variation (between 1-2% of battery consumed by the application at the end of the day) in both smartphones. The average amount of raw data transmitted and stored on the server for training purposes was $\simeq$1MB/day, therefore easily scalable to a higher number of users in terms of the required bandwidth and storage space. Moreover, in order to maintain the possibility of experimenting different features sets, data fragments sent to the server consisted of raw-data. This can be improved in the future by sending only the features to the server, either reducing the required data to be transmitted, or increasing the number of samples to be sent.

The learning curve of the weakly supervised training showed that 5-6 days of monitoring were enough to converge to a satisfactory solution. The comparison, with the fully supervised approach also shows that the fully supervised method converges more rapidly as shown in Fig.4.

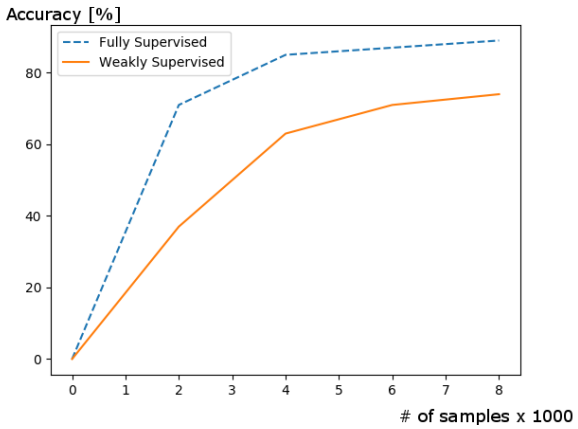


Fig. 4. Learning curve for the Fully-Supervised and Weakly Supervised methods showing Overall accuracy [%] increase rate with a growing number of samples used as training.

VI. Discussion

Gathered results show how the weakly supervised approach impacts the overall accuracy and the training process. In particular, a larger number of samples is required to train the model, however, over time the gap in accuracy compared to a fully supervised approach is reduced. The confusion matrix highlights how the uncertainty due to the heuristic (i.e. the presence of mislabeled samples) affects the accuracy. The comparison between the manually annotated ground truth and the *weak* ground truth (obtained using the heuristic), shows how the uncertainty introduced in the weak ground truth produces some mislabeling. Nevertheless, the impact of mislabeled samples is alleviated over time when a larger number of samples is collected. Both confusion matrices (weak and fully supervised) show how instances of the 'transportation' (C4) class are, in some cases, labeled as 'sitting' (C1). This is due to the situation when the user is sitting on a vehicle, but the vehicle is not moving.

In this experiment all weak labeled samples have been used for training. The quality of weak annotations generated by the heuristic could, however, be refined to eliminate some mislabeled sample artifacts. For instance, through statistical outliers elimination on samples with the same weak label; or some unsupervised clustering technique could be explored to identify isolated clusters generated by the heuristic that could be ignored. In terms of the approach adopted for the classification method, while optimal from the perspective of resource consumption, it may not be ideal purely from the accuracy perspective. In particular, CNN or other deep learning methods could be investigated to verify if they provide improved performance in terms of accuracy with a weakly labeled dataset.

VII. Conclusions

This paper describes work in progress and preliminary results that can potentially be improved by collecting a larger dataset. The goal of this experiment was to explore the viability of a weakly supervised approach to avoid the obstacle of manual ground truth generation. Although, the experiment has shown how a weak annotated dataset affects the overall accuracy, results highlight also how over time the weakly supervised approach converges towards an acceptable level of accuracy. The study provides encouraging results, however, repeating the experiment on a larger set of users is necessary to provide more representative results. In this first study we focused on labeling physical activity recognition, however, an extension will encompass more complex heuristics combining a user's diary and indoor localization to generate more refined activity labeling of ADLs. In conclusion, even though the experiment has highlighted how the presence of a (at least partially) manually labeled dataset can be beneficial, the positive results in terms of resource consumption show how there is scope for further exploration in the direction of online personalized training using weak supervision.

Acknowledgment

This work has been funded by the European Union Horizon 2020 MSCA ITN ACROSSING project (GA no. 616757). The authors would like to thank the members of the project's consortium for their valuable inputs.

References

[1] J. Morales and D. Akopian, Physical activity recognition by smartphones, a survey, Biocybernetics and Biomedical Engineering, vol. 37, no. 3, pp. 388400, 2017.

[2] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, A Survey of Online Activity Recognition Using Mobile Phones, Sensors, vol. 15, no. 1, pp. 20592085, Jan. 2015.

[3] Google Activity Recognition API. Available online: http://developer.android.com/training/ location/activity-recognition.html (accessed on July 2014).

[4] Chen, Liming, Jesse Hoey, Chris D. Nugent, Diane J. Cook, and Zhiwen Yu. "Sensor-based activity recognition." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, no. 6 (2012): 790-808.

[5] Reiss, Attila, and Didier Stricker. "Personalized mobile physical activity recognition." In Proceedings of the 2013 international symposium on wearable computers, pp. 25-28. ACM, 2013.

[6] Cruciani, F. *et al.*, "DANTE: A Video Based Annotation Tool for Smart Environments." In S-CUBE, pp. 179-188. 2010.

[7] Zimmerman, P. H. *et al.* "The Observer XT: A tool for the integration and synchronization of multimodal signals." Behavior research methods 41, no. 3 (2009): 731-735.

[8] I. Cleland *et al.* Evaluation of prompted annotation of activity data recorded from a smart phone. Sensors, vol. 14, no. 9, pp. 15 86179, Jan. 2014.

[9] Bao, Ling, and Stephen Intille. "Activity recognition from user-annotated acceleration data." Pervasive computing (2004): 1-17.

[10] Ermes, M., *et al.* "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions". IEEE Transactions on Information Technology in Biomedicine, 12(1), 2026. (2008).

[11] Stikic, M. *et al.* "Weakly Supervised Recognition of Daily Life Activities with Wearable Sensors". IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(12), 25212537.(2011).

[12] Kelly, D. and Caulfield, B. (2016). Pervasive sound sensing: A weakly supervised training approach. IEEE Transactions on Cybernetics, 46(1), 123135. [https://doi.org/10.1109/TCYB.2015.2396291].

[13] Zhou, Zhi-Hua, De-Chuan Zhan, and Qiang Yang. "Semi-supervised learning with very few labeled training examples." In AAAI, pp. 675-680. 2007.

[14] Susi, M., Renaudin M., and Lachapelle G. "Motion mode recognition and step detection algorithms for mobile phone users." Sensors 13, no. 2 (2013): 1539-1562.

[15] Choi, B. C. K., *et al.* "Daily step goal of 10,000 steps: A literature review". Clinical and Investigative Medicine. Retrieved from [http://cimonline.ca/index.php/cim/article/view/1083] (2007).

[16] Pedregosa, F. *et al.*, "Scikit-learn: Machine Learning in Python ", JMLR 12, pp. 2825-2830, 2011.