Collection of a Diverse, Realistic and Annotated Dataset for Wearable Activity Recognition

I. Cleland[*], M. P. Donnelly[*], C.D. Nugent[*], J. Hallberg[†], M. Espinilla[‡] and M. Garcia-Constantino[*].
[*]School of Computing, Ulster University, Co. Antrim, Northern Ireland, United Kingdom.
[†]Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Sweden.
[‡]Department of Computer Science, University of Jaen, Jaen, Spain
*Corresponding author i.cleland@ulster.ac.uk

Abstract—**This paper discusses the opportunities and challenges associated with the collection of a large scale, diverse dataset for Activity Recognition. The dataset was collected by 141 undergraduate students, in a controlled environment. Students collected triaxial accelerometer data from a wearable accelerometer whilst each carrying out 3 of the 18 investigated activities, categorized into 6 scenarios of daily living. This data was subsequently labelled, anonymized and uploaded to a shared repository. This paper presents an analysis of data quality, through outlier detection and assesses the suitability of the dataset for the creation and validation of Activity Recognition models. This is achieved through the application of a range of common data driven machine learning approaches. Finally, the paper describes challenges identified during the data collection process and discusses how these could be addressed. Issues surrounding data quality, in particular, identifying and addressing poor calibration of the data were identified. Results highlight the potential of harnessing these diverse data for Activity Recognition. Based on a comparison of six classification approaches, a Random Forest provided the best classification (F-measure: 0.88). In future data collection cycles, participants will be encouraged to collect a set of "common" activities, to support generation of a larger homogeneous dataset. Future work will seek to refine the methodology further and to evaluate model on new unseen data.**

Keywords—Activity Recognition; Data Collection; Data Annotaion; Crowd Sourcing; Data Sharing; Data Quality.

I. Introduction

Human Activity Recognition (AR) is now an essential component of ubiquitous and pervasive computing solutions. Indeed, as a research field, AR is reaching maturity with a growing community that has been actively working in the area for many years. Advancement of AR, however, continue to be hampered by the lack of large, diverse, high quality, accurately annotated, publicly available datasets [1]. Whilst efforts have been made towards standardizing and disseminating best practice, in terms of methodologies associated with the collection and sharing of datasets [1], shared AR datasets remain limited with regard to the number of users and the diversity of activities [1]. To address this issue, we investigated the potential of collecting a large diverse dataset with support from an undergraduate student cohort. This paper presents an analysis of a dataset collected by 141 students undertaking undergraduate studies at Ulster University. The paper first provides an analysis of publicly available datasets for AR, highlighting similarities in methodology and issues around the size and diversity. Following this, the methods used for collection of this dataset are described. Results, highlighting the quality and utility of the data, are then presented. This includes results of AR using well established data driven classification approaches. Finally, the paper highlights a number of issues identified during the data collection process and proposes a series of recommendations of how these may be addressed.

II. Background

AR is commonly achieved through the application of machine learning techniques applied to data gleaned from low level sensors, such as accelerometers and gyroscopes [3]. The training of these algorithms relies largely on the acquisition, preprocessing, segmentation and annotation of raw sensor data into distinct activity related classes. For this reason, the development of AR algorithms requires high quality, and diverse activity data to enable the desired generalization capabilities of trained models [3]. A large-scale data set is recognized as being a key step in improving and increasing the widespread adoption of AR based applications [4,5]. Such large-scale data sets must include data from a variety of sensors, be recorded during a wide range of activities and represent the subtle differences exhibited by the target occupants of the environment. Additionally, to support a supervised learning paradigm, the data must include accurate ground truth labels that are representative of each recorded activity [6].

Based upon an analysis (Table I) of publicly available AR datasets that utilized wearable sensors, it was found that it is common for datasets to represent as few as 12 participants and as little as 6 activities. There has, however, been some movement towards larger open datasets for AR.

TABLE I. An overview of publicaly available datasets for Activity Recognition using wearable sensors.

Dataset	No. Participants	No. of Activities
Opportunity [7]	12	17
WISDM [8]	29	6
DaLiAc [9]	19	13
HAR [10]	30	6
SCUT-NAA [11]	44	10
UniMiB SHAR [12]	30	17
HASC [13]	116	6
REALDISP [14]	17	33
SPHERE [21]	12	20

The Opportunity [7] and UCI Human Activity Recognition (HAR) [10] datasets, for example, have become the most

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

commonly used benchmarking datasets for developing and evaluating wearable AR solutions. The Opportunity dataset acquired data from 12 subjects whilst performing 17 activities and gestures. This dataset contains data from a vast array of sensors and modalities including 72 environmental and body worn sensors [7]. This included 5 inertial measurement units (IMU) and 12 Bluetooth accelerometers worn on the body. In addition to a wide number of sensors, the dataset also has strengths in terms of how the data was collected. Participants recorded data during two types of recording sessions; Drill sessions where the subject sequentially performed a pre-defined set of activities and "Activities of daily living " sessions which are less guided and therefore deemed more reflective of real world scenarios. Whilst there is a vast amount of information contained in the Opportunity dataset, it is limited in terms of the number of participants (12) with a subset of data (4 participants) commonly being used for benchmarking purposes.

The HAR dataset contains data from 30 participants, undertaking 6 activities (walking, siting, standing, lying, ascending and descending stairs) [10]. Data was collected from a single accelerometer and gyroscope (Samsung SII) placed at the waist. Data was annotated manually offline from video. While this dataset has a reasonable number of participants, it is limited in terms of the variety in activities represented.

One additional dataset of note, given the number of participants involved, is that collected by the Human Activity Sensing Consortium (HASC) [13]. The dataset is the result of a collaboration between 20 teams who collected data from 116 participants. Data was obtained from a single accelerometer whilst undertaking 6 activities, stay, walk, jog, skip, ascending and descending stairs. The main strength of the dataset has been noted previously as the sheer number of subjects [7]. The dataset is limited, however, by the fact that the placement of the sensor was not standardized across all participants.

For researchers who have collected data for the purposes of AR, it is often clear why resulting datasets are limited in one way or another. Collecting, cleaning and labelling of data is both time consuming and costly. It is difficult to recruit large numbers of participants coupled with challenges to obtain and standardize hardware, used to collect the data. Whilst crowed sourcing and opportunistic sensing approaches to data collection for AR are becoming increasingly feasible, they too have limitations in terms of the accuracy and availability of annotation [15].

Taking these points into consideration, there is still a need for a dataset, collected from a large number of participants (>100), that is well-documented in terms of the methodology, has a standard sensor configuration and contains data for a range of activities. To meet these requirements, this paper presents a dataset collected by 141 undergraduate students at Ulster University. The following Section provides details of the methodology undertaken.

III. Methdology for Data Collection

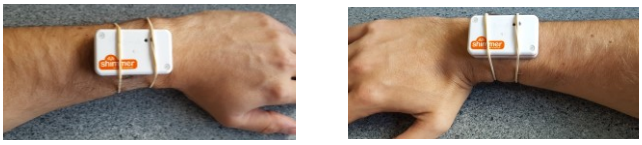
In the spring semester during 2017, 145 students enrolled onto the Pervasive Computing in Healthcare undergraduate module at Ulster University. The module provides students with relevant theory and significant practical opportunities in relation to AR and its workflow. The module culminates in a formal assessment that involves the collection, processing, modelling, analysis and supervised classification of data relating to various AR scenarios. Prior to undertaking data collection, students were equally assigned to one of six Scenarios, each represented by three activities (Table II). Students where provided with video instruction detailing each step of the data collection methodology including, calibration, placement and what each activity should look like, however, students where not supervised during this process. Activity data was recorded for two minutes per activity resulting in the collection of approximately six minutes of data per student.

TABLE II. Detals of Scenarios and Activities which students where asigned to and performed. Including the number of participants across the six scenarIos.

No.	Scenario	Activities	No. of participants
1	Self-Care	Hair Grooming, Washing Hands, Brushing Teeth	24 (72 files)
2	Exercise (Cardio)	Walking, Jogging, Stepping-Up	23 (69 files)
3	House Cleaning	Ironing Clothes, Washing Windows, Washing Dishes	25 (75 files)
4	Exercise (Weights)	Arm Curls, Dead Lift, Lateral Arm Raise	21 (63 files)
5	Sport	Bounce Ball, Catch Ball, Pass Ball	25 (75 files)
6	Food Preparation	Mixing Food, Chopping Vegetables. Sieving Flour	23 (69 files)
Total			**141 (423 files)**

A. Data aquisition and annotation

The Shimmer wireless sensors platform (Shimmer 2R, Shimmer Research, Dublin, Ireland) was used for data collection. Prior to collecting the data, the device was calibrated in line with manufacturer guidelines. Accelerometer data was recorded from the device placed on the dorsal aspect of the dominant Wrist (i.e. data could come from either arm as shown in Figure 1). The wrist was chosen as a suitable location for a single sensor as it is convenient to wear, is a common location for wearable devices, such as smart watches and provides reasonable accuracy in AR as previous research has shown [16] Participants placed the sensor themselves in the orientation depicted in Fig. 1.



(a) (b)

Fig. 1. Image depicting the location of the sensor on the (a) left or (b) right arm. Consistant orientation of the sensor was also maintained.

Data was recorded at a sample rate of 51.2Hz with a sensitivity range of ±1.5G. Shimmer Connect (Shimmer Connect V0.7, Shimmer Ressearch, D|ublin, Ireland) was used to stream data via Bluetooth. Data from each activity was saved to a seperate file. These files where then named with a label, representing the secenario and activity e.g. 1-3 for the activity brushing teeth. This data was then uploaded to a closed group repository.

IV. Dataset Description and Cleaning

From the original 145 students participating in the assignment, 141 uploaded data pertaining to the 6 scenarios. Table II presents the breakdown of the number of participants across each of the scenarios. With data collected, the next step was to clean and validate the dataset to ensure participants had correctly adhered to the protocol. All cleaning of the data was automated within Matlab (Mathworks, 2017a) using a suite of developed in-house scripts.

For labelling of the dataset, all participants labelled their data with a scenario and activity id as detailed in the provided protocol documentation. Some small discrepancies in file name format (i.e. scenario id and activity id in the wrong format) where subsequently corrected automatically within Matlab. Each file was checked to ensure it exceeded 2 minutes (6144 samples). Approximately 50% of the files were found to be less than 2 minutes in duration, however, only 2% of files collected had a duration outside 90 and 150 seconds (Fig. 2). Therefore, all files where included in the data analysis.

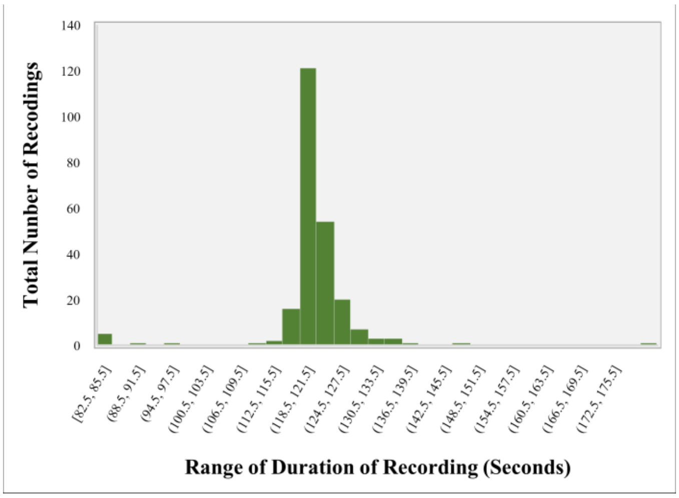


Fig. 2. Histogram of file duration. 98% of files uploaded were between 90 and 150 seconds in length.

All participants used a sample rate of 51.2Hz, as specified in the data collection protocol. When checking device sensitivity settings relating to each recording, 25 files were removed for having a maximum or minimum value of acceleration above the measurable range of the sensor (6g or 58.8m/s^2). These files were labelled as non-representative recordings and discarded. The remaining files were considered for further analysis.

V. Benchmark Classification Process

To investigate the potential of using data collected from a student population for developing AR models, the performance of six well known classification methods was assessed. Fig. 3 presents a common activity recognition pipeline that was adopted to process the data [18]. The following Sections provide further details of the process within each of these steps.

A. Data Preprocessing

Data was processed using each axis of accelerometry independently (x, y and z) and by combining the three axis to extract the signal magnitude vector (SMV), equation (1). The SMV is viewed as independent of orientation of the sensor node and was therefore its calculation was a crucial step in the AR pipeline, particularly as the sensor was permitted be placed on either the right or left wrist during the data recordings.

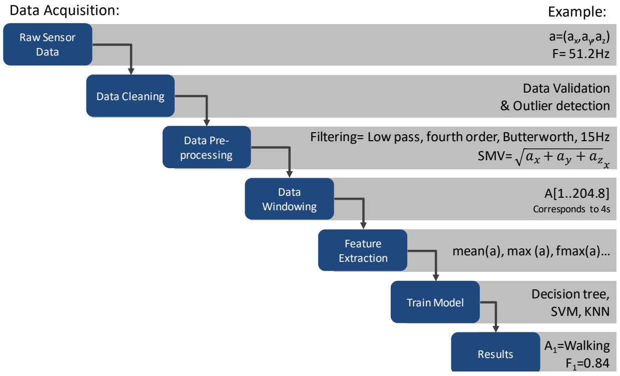


Fig. 3. Activity recognition pipeline showing the main processing steps.

$$\text{SMV}= \sqrt{a_x + a_y + a_z}_{x} \qquad (1)$$

The accelerometer signal was low pass filtered using a fourth order Butterworth filter, with a cut-off frequency set to 15Hz. This limited the bandwidth of the signal to frequencies commonly observed in human motion and eliminated high frequency noise. Approximately 99% of bodily acceleration from human activities is represented in frequencies below 15Hz.

B. Data Windowing

The accelerometer signal, including the SMV, was partitioned into 4 second (204 samples) non-overlapping windows. This window size has previously been reported in AR literature and is found to provide a balance between providing sufficient data within the window to capture a data sample representative of the entire activity, and the delay associated with larger window sizes [16]. Table III. presents the number of instances generated per windowed class. In total, 9612 instances were approximately equally represented across the 18 investigated activities.

C. Feature Extraction and Selection

A set of common features, defined in previous work [16, 18] were extracted from the x, y, z axis and SMV, as presented in Table IV. These features were chosen to represent both temporal and frequency domain information. Features 1-24 are common statistical metrics, computed from the time domain and extracted from the SMV. Feature 25 (Signal Magnitude Area (SMA)), represented in Equation 2, has been found to be a suitable measure for distinguishing between static and dynamic activities when employing triaxial accelerometer signals [18].

$$\text{SMA} = \sum_{i=1}^{N}(|x(i)|) + (|y(i)|) + (|z(i)|) \qquad (2)$$

where $x(i)$, $y(i)$ and $z(i)$, represent the acceleration signal along the x-axis, y-axis, and z-axis, respectively.

TABLE III. NUMBER OF FEATURES FOR EACH ACTIVITY WITHIN THE FINAL CLEANED DATASET.

Scenario Name	Activity	Activity id	No. Of instances
Self-Care	hair grooming	1	577
	washing hands	2	551
	teeth brushing	3	527
Exercise (Cardio)	Walking	4	491
	Jogging	5	510
	Stepping	6	500
House cleaning	Ironing	7	579
	window washing	8	555
	dish washing	9	577
Exercise (Weights)	arm curls	10	516
	Deadlift	11	469
	lateral arm raises	12	511
Sport	Pass	13	627
	Bounce	14	563
	Catch	15	598
Food Preparation	mixing food in a bowl	16	498
	chopping vegetables	17	475
	sieving flour	18	488
Total			**9612**

Feature 26 (Spectral Entropy) is the sum of the squared magnitude of the discrete fast Fourier transform (FFT) components of a signal. Feature 27 (Total Energy), is the sum of the squared discrete FFT component magnitudes of the SMV. The sum is divided by the window length for the purposes of normalization (1). This feature has been reported to result in accurate detection of specific postures and activities [19]. For instance, the energy of a subject's acceleration can discriminate low intensity activities such as lying from moderate intensity activities such as walking or high intensity activities such as jogging [25]. If ×1, ×2, ... are the FFT components of the window, then the energy can be represented by Equation (3):

$$\mathrm{Energy}_x = \frac{\sum_{i=1}^{|w|} |\mathrm{SMV}_i|^2}{|w|} \tag{3}$$

where *SMVi* are the *FFT* components of the window for the SMV axis and *w* is the length of the window.

Following feature extraction, the data was examined to highlight the existence of redundant and irrelevant information. Specifically, features were ranked based on the Information Gain from which the top 27, 25, 20, 15, 10 and 5 features were selected. A filter method, utilizing information gain was chosen over wrapper based methods as it is independent of learning methods. To identify the most appropriate number of features, we compared the performance of a C4.5 pruned decision tree (DT) in Weka (J48) on these 6 feature subsets. The decision tree was used here as it is a common classifier and has shown reasonable results in previous work [3].

The performance obtained from all 27 features was used as baseline for comparison. Table V presents the average F-measure following 10-fold cross validation of the five feature subsets. A paired t-test with significance level of *p*= 0.05 was used to compare the statistical significance of the results from that of the base line.

TABLE IV. FEATURES CONSIDERED WITHIN THIS WORK, INCULDING BOTH TEMPORAL AND FREQUENCY INFORMATION.

Feature No.	Feature Name	Feature Description	Selected Y/N [a]
1-4	Mean value	Mean value of the x, y, z and SMV in the window.	Y/Y/Y/N
5-8	Maximum	Maximum value of the x, y, z and SMV in the window.	Y/Y/Y/Y
9-12	Minimum	Minimum value of the x, y, z and SMV in the window.	Y/Y/Y/Y
13-16	Standard Deviation	Standard deviation of the samples x, y, z and SMV in the window.	Y/Y/Y/Y
17-20	Range	Range of the samples of SMV in the window.	Y/Y/Y/Y
21-24	Root Mean Square	Root Mean Square of the values of x, y, z and SMV in the window.	Y/Y/Y/Y
25	Signal Magnitude area	Signal Magnitude Area (SMA) across the acceleration signal in x, y and z axis.	Y
26	Spectral Entropy	The normalized information entropy magnitudes of the discrete FFT components of the signal.	N
27	Total Energy	Sum of the squared magnitudes of the discrete FFT components of the signal	Y

[a] The selected column shows which features where included following feature selection.

TABLE V. F MEASUE SCORE OF PERFORMANCE OF DECISION TREE USING SUBSETS OF 25, 20, 15, 10 AND 5 FEATURES RANKED BY INOFRMATION GAIN.

No. of Features	F measure of Decision Tree (DT)[a]
27 (All)	0.78
25	0.74^{-}
20	0.72*
15	0.72*
10	0.69*
5	0.56*

a. Markers denote – no statistical or * statistically worse than base line. Significance 0.05.

This analysis highlighted no significant decrease in F-measure of the DT when employing 25 features. For this reason, 25 features where used in subsequent tests.

D. Data Cleaning

In addition to the data cleaning that took place, post data collection, additional data cleaning was undertaken to identify potentially outlying features / instances that could have resulted from participants not adhering correctly to the protocol. This was conducted at a feature level by identifying outliers in the represented feature space. Allowing for the inter-subject variability among participants, it was decided to identify outliers within subjects for each activity, independently. Consequently, outlier values were deemed to be those more than three scaled absolute deviations away from the median (Median Absolute Deviation (MAD)). MAD has been shown to be a more robust method of detecting outliers compared to standard deviation around the mean [17]. If an instance contained two or more outlying features, that instance was removed from the dataset.

In total, 1575 instances were removed as these contained an outlier in one or more feature. Fig.4. presents a boxplot of the cleaned feature space for each feature. Given that data was collected in a controlled lab setting, it was anticipated that those data existing at the start and end of each recording may not be reflective of the target activity. This was observed, for example, when there was a short delay between the time that a participant commenced data acquisition and the time at which they simulated the target activity. Consequently, a single window at both the start and end of each file was discarded from the dataset. This aspect of pre-processing has previously been reported in the literature [16].

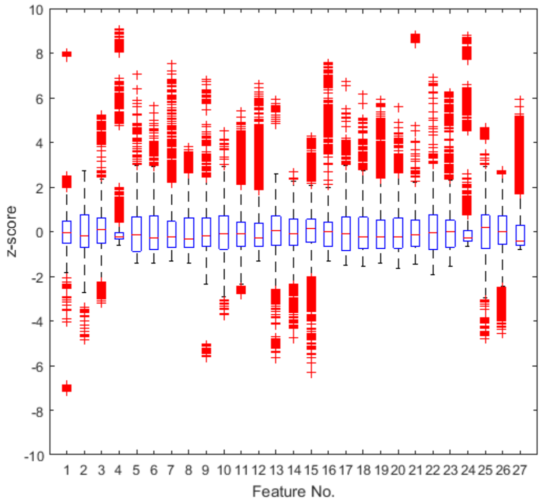


Fig. 4. Stadardised z-score showing the spread of features around the mean. Data shown is the cleaned dataset post pre-procsessing and removal of outliers.

E. Classification Models

Six classifiers were benchmarked against the available data; namely, C4.5 Decision Tree (DT), Naïve Bayes (NB), Neural Network (NN) (Multilayer Perceptron), K- Nearest Neighbor (KNN), Random Forest (RF) and Support Vector Machine (SVM). These approaches are well cited in the literature and an overview of each, including their respective benefits and limitations can be found in Preece *et al.* [2].

F. Validation

To benchmark each selected machine learning algorithm, a 10-fold cross validation was performed within Weka Experimenter (University of Waikato, Version 3.8.1). F-measure was used as a non-bias performance, providing a weighted harmonic mean of the precision and recall of each classifier, represented as in Equation (4):

$$\text{F−measure} = 2 \times \frac{precision \times recall}{presicion + recall} \quad (4)$$

where *Precision (positive predictive value)* is the fraction of retrieved instances that are relevant and *Recall (sensitivity)* is the fraction of relevant instances that are retrieved. A higher F-measure value is indicative of better performance. A paired t-test was applied to the results to identify if the F-measure was significantly different using the RF when compared to the DT, NB, KNN, NN or SVM. The RF was used as the baseline scheme, with the other five algorithms being compared to it. A value of less than p = 0.05 was considered statistically significant.

VI. Results

Results summarizing the classification are presented in Fig. 5. Of the six classifiers investigated, RF and KNN resulted in the highest F-measures of 0.88 and 0.86, respectively. These results were a statistically significant improvement compared to the DT (0.74), NN (0.76), NB (0.67) and SVM (0.67).

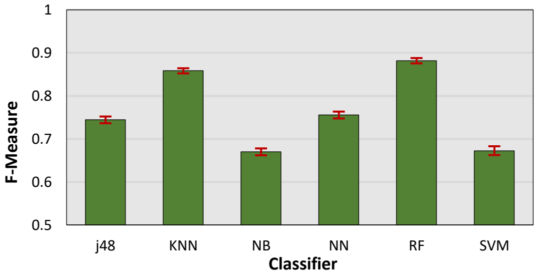
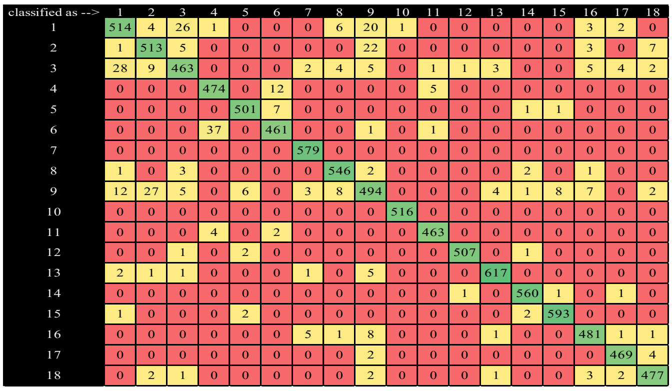


Fig. 5. Average F-measure, over 10 folds, for the DT, KNN, NB, NN, RF and SVM classifiers. Error bars show 95% confidence intervals.

Examining performance at a class level, the confusion matrix of the RF is presented in Fig. 6. From this, it is noted that the activities of hair grooming (1), washing hands (2) and dishwashing (9) were regarded as the most challenging activities to discriminate between. This is postulated to be due to the natural randomness inherent in performing these types of activities, where hand movements vary from person to person, however, is also indicative of the similarity of movements within these three activities. This is the case for other activities which represent similar movements such as walking (4) and stepping (6). Other activities, contained less “randomness” in terms of movement and therefore had better classification accuracy. For example, arm raise (12), bounce (14) and catch (15) are all activities which contained more rhythmic repetitive movements and therefore achieved high rates of accuracy classification (>90%).

VII. Discussion & Conclusion

This paper has presented and attempted to appraise a methodology to acquire and annotate population diverse data sets for AR, within an academic teaching environment. Analysis of the data set has highlighted several benefits of using a student population to create a rich large-scale data sets for creating robust AR models. Specifically, data collection with such a large group of participants has produced a high level of diversity and therefore generalization within the data. As data was collected by students in an unsupervised setting, variations in the data collection methodology has led to a highly diverse dataset. This is arguably more representative of a dataset collected “in the wild” than data collected under more controlled circumstances. This, however, also highlighted several challenges, not least the requirement for assuring data quality and to identify outliers in the labelled training data. There is a fine balance to afford data that is representative of the activities without detracting from natural inter-subject variability, capturing the subtle differences in how individuals perform activities. The preliminary results presented in this paper demonstrate the potential to use these data to achieve recognition rates that are comparable to those reported in the literature [7-14]. Undertaking this analysis has also highlighted a number of methodological limitations that will be improved upon during the next data collection cycle, planned for early 2018. Specifically, the data set described within this paper is limited by the fact that participants collected data for a small number of activities (3), within their allocated scenario



classified as -->	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	514	4	26	1	0	0	0	6	20	1	0	0	0	0	0	3	2	0
2	1	513	5	0	0	0	0	0	22	0	0	0	0	0	0	3	0	7
3	28	9	463	0	0	0	2	4	5	0	1	1	3	0	0	5	4	2
4	0	0	0	474	0	12	0	0	0	0	5	0	0	0	0	0	0	0
5	0	0	0	0	501	7	0	0	0	0	0	0	0	1	1	0	0	0
6	0	0	0	37	0	461	0	0	1	0	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	579	0	0	0	0	0	0	0	0	0	0	0
8	1	0	3	0	0	0	0	546	2	0	0	0	0	2	0	1	0	0
9	12	27	5	0	6	0	3	8	494	0	0	0	4	1	8	7	0	2
10	0	0	0	0	0	0	0	0	0	516	0	0	0	0	0	0	0	0
11	0	0	0	4	0	2	0	0	0	0	463	0	0	0	0	0	0	0
12	0	0	1	0	2	0	0	0	0	0	0	507	0	1	0	0	0	0
13	2	1	1	0	0	0	1	0	5	0	0	0	617	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	0	560	1	0	1	0
15	1	0	0	0	2	0	0	0	0	0	0	0	0	2	593	0	0	0
16	0	0	0	0	0	0	5	1	8	0	0	0	1	0	0	481	1	1
17	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	469	4
18	0	2	1	0	0	0	0	0	2	0	0	0	1	0	0	3	2	477

Fig. 6. Confusion matrix for RF classifier. Numbers represent activity ID from table V.

In future data collection cycles, participants will be encouraged to collect a set of "common" activities, to support the generation of the number of common instances. Within this study, participants were also permitted to calibrate their own device prior to data collection, which raises open questions surrounding the quality of these calibrations. While a number of instances were removed during the data cleaning process, not all instances of poor calibration may have been identified. Future data collection cycles will therefore include the capture of stationary data for each axis as a validation measure and facilitate automatic recalibration of data, if required [20]. Furthermore, the accelerometer dynamic range was set to ± 1.5g. In practice, this range resulted in clipping of the signal for the more vigorous activities, for example, bouncing a ball. A larger range of at least ± 4g will be employed to ensure this is not the case within future work. Students were provided with video instructions of how to perform activities during data collection. This may have in some way impacted upon how the individual would naturally carry out that activity. Given the high diversity of the data, however, the investigators believe that the impact of this on the dataset was limited. In summary, this paper has shown the potential of utilizing a student population to collect a highly diverse dataset for AR. Challenges have presented in terms of assuring that the data collected is in line with the collection methodology. Work is, therefore, required to create automated techniques for data cleaning and validation.

Future work will additionally seek to evaluate the models generated on this data set with completely unseen test data, collected by a new intake of student participants, during the next academic year. This will include a more in-depth evaluation of appropriate classification techniques including a leave-one-subject-out validation methodology.

Acknowledgments

This research has received funding under the REMIND project Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020, under Grant Agreement No. 734355. Invest Northern Ireland is acknowledged for supporting this project under the Competence Centre Programme Grant RD0513853 - Connected Health Innovation Centre.

References

[1] Lockhart, Jeffrey W., and Gary M. Weiss. "Limitations with activity recognition methodology & data sets." In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 747-756. ACM, 2014.

[2] Preece, S.J.; Goulermas, J.Y.; Kenney, L.P.J.; Howard, D.; Meijer, K.; Crompton, R. Activity Identification using Body-Mounted Sensors—A Review of Classification Techniques. Physiol. Meas. 2009, 30, R1.

[3] Avci, A.; Bosch, S.; Marin-Perianu, M.; Marin-Perianu, R.; Havinga, P. Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey. In Proceedings of the 23rd International Conference on Architecture of Computing Systems (ARCS), Hannover, Germany, 22–23 February 2010; pp. 1–10.

[4] Intille, S.S.; Lester, J.; Sallis, J.F.; Duncan, G. New Horizons in Sensor Development. Med. Sci. Sports Exerc. 2012, 44, S24–S31.

[5] Lane, N.D.; Miluzzo, E.; Lu, H.; Peebles, D.; Choudhury, T.. A Survey of Mobile Phone Sensing. IEEE Commun. Mag. 2010, 48, 140–150.

[6] Hossmann, T.; Efstratiou, C.; Mascolo, C. Collecting Big Datasets of Human Activity One Check in at a Time. In Proceedings of the 4th ACM International Workshop on Hot Topics in Planet-Scale Measurement, Ambleside, UK, 25–29 June 2012; pp. 15–20.

[7] Chavarriaga, Ricardo, Hesam Sagha, Alberto Calatroni, *et. al.* "The Opportunity challenge: A benchmark database for sensor-based activity recognition." Pattern Recognition Letters 34, no. 15 (2013): 2033-2042.

[8] Kwapisz, Jennifer R., Gary M. Weiss, and Samuel A. Moore. "Activity recognition using cell phone accelerometers." ACM SigKDD Explorations Newsletter 12, no. 2 (2011): 74-82.

[9] Leutheuser, Heike, Dominik Schuldhaus, and Bjoern M. Eskofier. "Hierarchical, multi-sensor based classification of daily life activities." PloS one 8, no. 10 (2013): e75196.

[10] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Spain. Dec 2012.

[11] Xue, Yang, and Lianwen Jin. "A naturalistic 3D acceleration-based activity dataset & benchmark evaluations." In Systems Man and Cybernetics (SMC), 2010 IEEE, pp. 4081-4085. IEEE, 2010.

[12] Micucci, Daniela, Marco Mobilio, and Paolo Napoletano. "UniMiB SHAR: a new dataset for human activity recognition using acceleration data from smartphones." arXiv preprint arXiv:1611.07688 (2016).

[13] Kawaguchi, N., et al. HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings. In Proc. 2nd Augmented Human International Conference (2011).

[14] Banos, O., Toth M. A., Damas, M., Pomares, H., Rojas, I. Dealing with the effects of sensor displacement in wearable activity recognition. Sensors vol. 14, no. 6, pp. 9995-10023 (2014).

[15] Cleland, Ian, Chris Nugent, and Sungyoung Lee. "The ground truth is out there: challenges with using pervasive technologies for behavior change." In Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 322-325, 2016.

[16] Mannini, Andrea, Mary Rosenberger, William L. Haskell, Angelo M. Sabatini, and Stephen S. Intille. "Activity Recognition in Youth Using Single Accelerometer Placed at Wrist or Ankle." Medicine and science in sports and exercise 49, no. 4 (2017): 801-812.

[17] Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median." J Exper. Social Psycho. 49, no. 4 (2013): 764-766.

[18] Bulling, Andreas, Ulf Blanke, and Bernt Schiele. "A tutorial on human activity recognition using body-worn inertial sensors." ACM Computing Surveys (CSUR) 46, no. 3 (2014): 33.

[19] Sugimoto, A.; Hara, Y.; Findley, T.; Yoncmoto, K. A useful method for measuring daily physical activity by a three-direction monitor. Scand. J. Rehabil. Med. 1997, 29, 37–42.

[20] Van Hees, Vincent T., et al. "Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents." Journal of Applied Physiology 117.7 (2014): 738-7.

[21] Twomey, N., Diethe, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P. and Craddock, I., 2016. The SPHERE challenge: Activity recognition with multimodal sensor data. arXiv preprint arXiv:1603.00797.