# Improving the Quality of User Generated Data Sets for Activity Recognition

Chris Nugent[1,4(✉)], Jonathan Synnott[1], Celeste Gabrielli[2], Shuai Zhang[1], Macarena Espinilla[3], Alberto Calzada[1], Jens Lundstrom[4], Ian Cleland[1], Kare Synnes[5], Josef Hallberg[5], Susanna Spinsante[2], and Miguel Angel Ortiz Barrios[6]

[1] School of Computing and Mathematics, Ulster University, Jordanstown, Northern Ireland, UK
{cd.nugent, j.synnott, s.zhang, a.calzada, i.cleland}@ulster.ac.uk
[2] Dipartimento dell'ingegneria dell'informazione, Universita Politecnica Delle Marche, Ancona, Italy
sl052886@univpm.it, s.spinsante@staff.univpm.it
[3] Department of Computer Sciences, University of Jaen, Jaen, Spain
mestevez@ujaen.es
[4] School of Information Technology, Halmstad University, Halmstad, Sweden
Jens.lundstrom@hh.se
[5] Department of Computer Science, Electrical and Space Engineering, Lulea Technical University, Luleå, Sweden
{unicorn, Josef.hallberg}@ltu.se
[6] Industrial Engineering Department, Universidad de La Costa CUC, Barranquilla, Colombia
mortizl@cuc.edu.co

**Abstract.** It is fully appreciated that progress in the development of data driven approaches to activity recognition are being hampered due to the lack of large scale, high quality, annotated data sets. In an effort to address this the Open Data Initiative (ODI) was conceived as a potential solution for the creation of shared resources for the collection and sharing of open data sets. As part of this process, an analysis was undertaken of datasets collected using a smart environment simulation tool. A noticeable difference was found in the first 1–2 cycles of users generating data. Further analysis demonstrated the effects that this had on the development of activity recognition models with a decrease of performance for both support vector machine and decision tree based classifiers. The outcome of the study has led to the production of a strategy to ensure an initial training phase is considered prior to full scale collection of the data.

**Keywords:** Activity recognition · Open data sets · Data validation · Data driven classification


C.R. García et al. (Eds.): UCAmI 2016, Part II, LNCS 10070, pp. 104–110, 2016.
DOI: 10.1007/978-3-319-48799-1_13

## 1 Introduction

It is fully appreciated that data driven approaches developed for the purposes of activity recognition are being constrained due to the lack of the availability of large annotated data sets which provide a high level of quality in terms of both ground truth and the underlying data. A number of efforts have been made towards the creation of such datasets, however, this has led to a number of individual studies being undertaken and the generation of datasets, which although have aimed to address the same problem, have resulted in non-common protocols being used with differing technologies providing data of different formats. In an effort to address these problems a number of organisations have worked together to define both common protocols and technology platforms to support the collection and storage of data. Through this collaboration the Open Data Initiative was conceived [1]. By adopting such an approach it becomes possible to collect data at different sites and aggregate into one common dataset.

In our previous work the aspirations of the ODI were followed in being able to generate data and make it available to independent researchers. This provides the ability for researchers to compare and contrast innovative approaches on exactly the same data and therefore being able to make true comparisons between approaches [2]. To facilitate this approach required the generation of a simulated dataset. Four researchers were asked to complete a series of activities within a smart environment by following a predefined protocol. Analysis of this data demonstrated that there was a large variance in the initial iterations of data collection [3]. The work presented within this paper proposes strategies to both assess and improve the quality of user generated data sets and highlights the importance for appropriate user training.

The remainder of the paper is structured as follows. Section 2 presents background to generation of activity related data sets and introduces the ODI. Section 3 explains the process by which the data was collected using IESim and Sect. 4 presents the results from various activity recognition models. Section 5 summarises the paper with an overview of the findings and recommendations.

## 2 Related Work

There is evidence that the research community have not only recognised the importance of having large, high quality datasets, however, there are now efforts being made to create data sets which can be shared. This will have the desired knock-on effect of improving the efficiency with which data is stored and aggregated in addition to the improved development of data driven algorithms themselves.

A number of crowd sourced approaches have emerged as potential solutions to this problem. Crowdsignals.io and the UbiHealth Sensing Campaign [4] have both launched strategies to support the definition of protocols for the collection of data in addition to the collection and annotation of data sets which can then be made publically available.

The European Union funded Project OPPORTUNITY in their work created a common platform whereby researchers working in different organisations could have access to a common data set and therefore were able to compare their results with

others [5]. A limited number of online repositories have supported the notion of shared datasets. Two notable examples are the UC Irvine Machine Learning repository [6] and Physionet [7]. The former has recently extended its datasets to include a small number of activity recognition related resources.

The ODI, an initiative established by the authors of this paper aimed to address both the definition of a protocol for the collection of data at different sites in addition to a framework for evaluation [1]. To date efforts from the ODI have led to the development of a common dataset which has-been collected in different organisations using a mix of both real and simulated sensing environments. The work has been subsequently extended to provide a platform whereby a comparison of a range data driven approaches, independently developed by researchers from different organisations, has been undertaken [2].

Although all of these studies have embraced the notion of creating easily accessible resources for the sharing of protocols and data, little effort has been reported whereby the quality of the data and guidance provided to those who are collecting the data has been considered.

## 3 Generation of Data Sets in Simulated Environments

Initial efforts of those involved with the ODI have led to the development of a number of simulated datasets using the IESim platform [8]. IESim is a simulation tool which supports the replication of real environments through the creation of a simulated environment. Sensors can be added within the simulated environment to replicate their placement within the real environment. Users can then engage with the simulated environment through use of an avatar and generate data relating to interaction with sensorised objects [8]. Figure 1 presents an overview of a real test bed environment (from Ulster University) and its subsequent realisation within IESim. The format of the data produced by IESim can be tailored to meet the requirements of any future processing modules. At present IESim can support up to 4 differing types of data formats.

In the current study IESim was used to create an environment with 5 rooms. Throughout the environment 21 sensors were included to record the activities being

(a) (b)

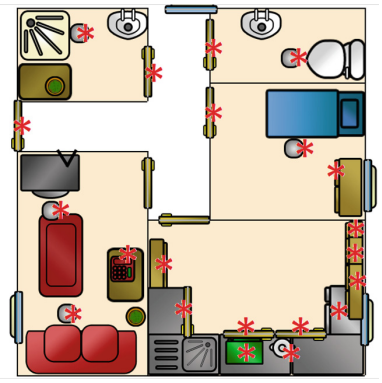
**Fig. 1.** (a) Overview of IESim replicating smart kitchen environment from the (b) Smart Environments Research Group at Ulster University's smart labs.

undertaken. Figure 2 provides the layout for the environment and an overview of the activities each participant was asked to undertake. Data was collected from 8 participants who used the simulation tool to complete the predefined set of activities. Each participant repeated each of the 11 activities 7 times, producing a data set with a total of 616 instances. The time taken for each participant to complete the set of activities (n = 77) was less than 30 min.

Upon analysis of the data collected it was found that the initial recordings made by participants using IESim varied (Fig. 3). Based on visual inspection it can be viewed that the 1st and 2nd replication tend to be different than the others for both activities presented in Fig. 3 (*WatchTV* and *LeaveHouse*). An outlier may be used an indication of bad-quality data. If detected, the outlying value must be deleted to avoid significant statistical changes in the distribution of the data. In the current work a Grubb' s test was performed with a 95 % level of confidence as an outlier identification test [10, 11].

| Activity 1 – Go to Bed |
|---|
| Activity 2 – Use Toilet |
| Activity 3 – Watch TV |
| Activity 4 – Prepare Breakfast |
| Activity 5 – Take Shower |
| Activity 6 – Leave House |
| Activity 7 – Get Cold Drink |
| Activity 8 – Get Hot Drink |
| Activity 9 – Prepare Dinner |
| Activity 10 – Get Dressed |
| Activity 11 – Use Telephone |

(a)



(b)

**Fig. 2.** (a) List of activities undertaken by participants. (b) Environment used for data collection. Each sequence of 11 activities was repeated 7 times by each participant. Stars indicate where the sensors (n = 21) were placed.

It was found that the 1st replication of User 2 was identified as an outlier with a p-value = 0.00307 ($\alpha = 0.05$) and 1.92069 standard deviations from the sample mean. This is an indication that this point does not follow the statistical behaviour of the sample and must be removed since it generates a significant variance change. To further elaborate on this occurrence the same analysis was carried out for the *Leave-House* activity. By applying the outlier identification technique it was found that the 1st replication of User 3 was identified as an outlier with a p-value = 0.0114 ($\alpha = 0.05$) and 2.13201 standard deviations from the sample mean. This is an indication that this point does not follow the statistical behaviour of the sample and must be removed as it causes a significant variance change.

Although participants found IESim intuitive to use a number of usability errors were recorded. The main error was one of incorrect self-annotation. The effects of this process were sensor events being assigned to either the previous or following activity.

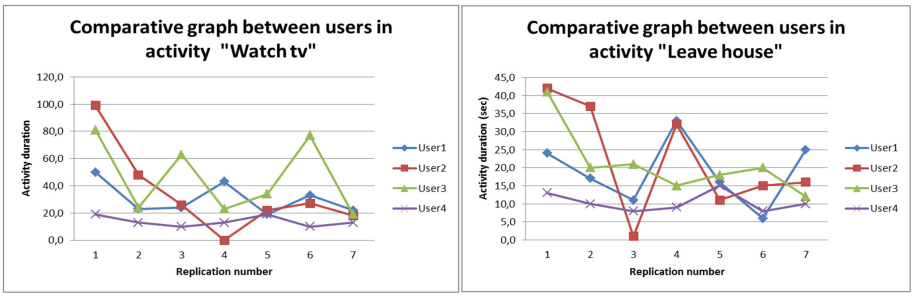



**Fig. 3.** Examples of assessment of time of 4 participants completing activities *WatchTV* and *LeaveHouse* completed using IESim. Each replication represents the completion of the set of 11 activities.

In addition, on a small number of occasions, participants skipped sub-tasks within the activities they were completing. These errors had a minor effect on the quality of the data and were considered to have provided a more realistic distribution of sensor events over a range of different activities.

## 4 Evaluation

In order to evaluate the quality of the data on the development of data driven activity recognition approaches three versions of the dataset were considered:

1. The original data set with all 7 iterations from all participants.
2. The original data set with the first iteration of experiments performed by all participants removed.
3. The original data with the first and second iteration of experiments performed by all participants removed.

Removal of replications 1 and 2 was intended to avoid the effects of non-representative completion of activities during a period of time of first usage with IESim when participants made themselves familiar with the simulator.

Two classification models have been considered; support vector machines (SVM) and decision trees. SVM algorithms use a non-linear kernel to discriminate the feature space into various classes. This approach offers the advantage in our current problem where the activity classes may have a non-linear relationship with the captured data. Additionally, SVMs handle a high dimensionality of the feature space, in our case represented by the large number of sensors deployed to monitor the designed activities. The decision tree approach, in particular, offers the advantage of intuition, where both the methodology, the derived model and the results are coherent. This has the potential to provide additional valuable information on the discriminative ability of the sensors in the environment. The decision tree approach may, however, suffer from differentiating classes which are not linearly separable. Decision tree approaches have demonstrated their superior performance over other popular machine learning approaches,

i.e., decision table, instance-based learning or nearest neighbour, and Naive Bayes classifiers in activity recognition tasks [9]. Table 1 presents the results attained with both classifiers using 10 fold cross validation on the 3 datasets.

**Table 1.** Summary of results from usage of 2 classifiers with 3 different datasets.

| | Accuracy with 7 repetitions (n = 616) | Accuracy with 6 repetitions (n = 528) | Accuracy with 5 repetitions (n = 440) |
|---|---|---|---|
| Support vector machine (10 fold cross validation) | 93,83 % | 94,69 % | 94,32 % |
| Decision tree (10 fold cross validation) | 93,83 % | 94,51 % | 94,77 % |

## 5 Summary

In this paper we have demonstrated the effects that poor quality data can have when developing data driven approaches to activity recognition. Upon closer examination of a dataset which was previously collected by the ODI it was found that the initial 1–2 replications through the data, when those generating the data where learning the approach to adopt for data collection, where largely different to data collected in later cycles. As such it was found that this data was not representative of the target activities and should be removed. This was further evidenced through the application of outlier detection testing. To mitigate the impact of such data in the development of the classification process it is recommended that participants are provided with an opportunity to be trained with the simulation tool and that the initial replications are not recorded in the final dataset until the point in time when the participant can use the system confidently. This approach can be extended into the more general domain of pervasive healthcare where data is both generated and collected in the wild. There is the potential to improve the quality of such data through periodic training sessions. It may also be the case that algorithms developed will never reach 100 % accuracy due to the complexity of the problem, however, efforts should be made to provide the training process with as high a quality data set as possible. Future work will involve testing of this concept further through the collection of new datasets and generation and analysis of a range of classification models. In addition, further efforts will be made to analyze the effects on an activity per activity basis.

**Acknowledgments.** Invest Northern Ireland partially supported this project under the Competence Centre Program Grant RD0513853 – Connected Health Innovation Centre.

## References


1. Nugent, C., Cleland, I., Epsinilla, M., Santanna, A., Synnott, J., Banos, O., Lundstrom, J., Hallberg, J., Calzada, A.: An initiative for the creation of open datasets within pervasive healthcare. In: Future of Pervasive Health Workshop. ACM (2016). doi:10.4108/eai.16-5-2016.2263830
2. Synnott, J., Nugent, C., Zhang, S., et al.: Environment simulation for the promotion of the open data initiative. In: SmartSys Workshop (2016)
3. Ortiz Barrios, M., Nugent, C., Synnott, J.: A methodology for assessing the quality of datasets in support of data driven activity recognition. In: EMBC 2106 (2016, in press)
4. Ubihealth project. http://www.ubihealth-project.eu/index.php. Accessed 8 March 2016
5. Sagha, H., et al.: Benchmarking classification techniques using the opportunity human activity dataset, In: 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Anchorage, AK, pp. 36–40 (2011)
6. UCI machine learning repository. http://archive.ics.uci.edu/ml/. Accessed 08 March 2016
7. PhysioNet: the research resource for complex physiologic signals. https://www.physionet.org/. Accessed 08 March 16
8. Synnott, J., Chen, L., Nugent, C.D., Moore, G.: The creation of simulated activity datasets using a graphical intelligent environment simulation tool. In: Engineering in Medicine and Biology Society (EMBC), pp. 4143–4146 (2014)
9. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
10. Barrios, M.O., Jiménez, H.F.: Reduction of average lead time in outpatient service of obstetrics through six sigma methodology. In: Bravo, J., Hervás, R., Villarreal, V. (eds.) AmIHEALTH 2015. LNCS, vol. 9456, pp. 293–302. Springer, Heidelberg (2015)
11. Herazo-Padilla, N., Montoya-Torres, J.R., Muñoz-Villamizar, A., Isaza, S.N., Polo, L.R.: Coupling ant colony optimization and discrete-event simulation to solve a stochastic location-routing problem. In 2013 Winter Simulations Conference (WSC), pp. 3352–3362, December 2013. IEEE (2013)