An initiative for the creation of open datasets within pervasive healthcare

Chris Nugent
Ulster University, UK
Halmstad University, Sweden
+44 2890 368330
cd.nugent@ulster.ac.uk

Ian Cleland
Ulster University
Northern Ireland, UK
+44 2890 368840
i.cleland@ulster.ac.uk

Anita Santanna
Halmstad University
Sweden
+46 (0) 35 16 78 49
anisan@hh.se

Macarena Espinilla
University of Jaen,
Spain
+34 953212897
mestevez@ujaen.es

Jonathan Synnott
Ulster University
Northern Ireland, UK
+44 2890 368840
j.synnott@ulster.ac.uk

Oresti Banos
University of Twente
The Netherlands
+31 534895329
O.banoslegran@utwente.nl

Jens Lundström
Halmstad University
Sweden
+46 (0) 35 16 78 65
Jens.lundstrom@hh.se

Josef Hallberg
Lulea Technical University
Sweden
+46 920493177
Josef.hallberg@ltu.se

Alberto Calzada
Ulster University
Northern Ireland, UK
albertocalsa@gmail.com

ABSTRACT

In this paper issues surrounding the collection, annotation, management and sharing of data gathered from pervasive health systems are presented. The overarching motivation for this work has been to provide an approach whereby annotated data sets can be made readily accessible to the research community in an effort to assist the advancement of the state-of-the-art in activity recognition and behavioural analysis using pervasive health systems. Recommendations of how this can be made a reality are presented in addition to the initial steps which have been taken to facilitate such an initiative involving the definition of common formats for data storage and a common set of tools for data processing and visualization.

CCS Concepts

Applied computing→Life and medical sciences→Health care information systems**.**

Keywords

Activity recognition; Pervasive Computing; Data Sets; Repository.

1. INTRODUCTION

Recent advances in information technology have revolutionised science and how it is applied to society in all aspects of life. Specifically, within the realms of pervasive computing, this has provided new opportunities for researchers to generate and share data and build on each other's work. Mining large datasets and combining them with information from other heterogeneous sources offers a vast potential to advance developments in public health, applications in pervasive and connected health and development of new strategic policies.

The importance of data sharing in advancing the state of the art is central to all of these initiatives. The need for making health related data openly available is becoming increasingly recognised and has been strongly endorsed by the H8 group of global health organisations. Additionally, when developing health related solutions, it is imperative that the solutions are tested with diverse populations with a range of ethnic, gender, social and cultural backgrounds. From a technical perspective consideration should also be given to different sensor types being used, their locations within the environment, different physical environments and different users. This will assist to ensure that developed solutions are extensible and scalable on a global level and will be more generalizable to unknown situations.

Much of the data collection that could improve health research is expensive and time-consuming. In many cases, the processes to both generate and collect the data are duplicated which further adds to the expense of the process. Making research data sets available beyond the original research team where they have been generated, in a timely and responsible manner, subject to appropriate safeguards and standards, will offer a number of benefits.

At present, the interest in progressing activity recognition and behavioural analysis research based on data gleaned from pervasive health systems is ever increasing [1]. This is leading to large amounts of similar efforts within the domain designing similar experiments, collecting similar data and trying to design similar approaches to behavioural analysis. The research is, however, being hindered by the limited size of the datasets, especially in instances of data driven behavioural model design.

Coupled with this is the lack of generalisation of solutions given limited diversity of the sample size considered in data collection.

This paper presents efforts, referred to as the Open Data Initiative (ODI), which are currently being undertaken in an attempt to address these challenges and offers a structured approach to provide annotated data sets in an accessible format for the research community. It is hoped that with these efforts an advancement of the state-of-the-art in activity recognition and behavioural analysis will be achieved.

2. Related Works

The research domain of activity recognition and behavioural analysis is starting to show evidence of researchers working together in an attempt to be more efficient in both the collection and sharing of data sets and the development of intelligent data analysis techniques.

An initiative managed by CrowdSignals.io has recently provided a crowdfunding campaign in an effort to address the challenges related to the lack of publically available datasets. The aspirations of the campaign are to provide the largest set of fully annotated data from sensors and mobile phones through the provision of a uniquely developed data collection platform. At the time of writing this article the platform was in alpha version and had not been fully released to the general public.

The UbiHealth Sensing Campaign aims to define a common protocol whereby researchers in different regions can be involved in the collection of research data using mobile solutions [2]. This initiative is taking a step forward in the development of shared approaches to avoid the duplication of efforts in the collection of data sets to later be used as the basis for activity recognition and behavioural analysis. The research is based around a common framework referred to as the mk-sense framework which involves an application for a mobile phone and a data management application deployed on a server. At the time of writing this article the campaign was collecting its first data set across 9 institutions (and 7 countries) to study how friendly an area was for walking in.

Research by Nugent *et al*. aimed to develop an open format for the representation of data collected within smart environments [3]. This work led to the production of an xml based schema which successfully proved to have the ability to retrospectively repurpose data previously collected within research studies [4]. The challenge with the approach was faced when trying to raise awareness within the research community and engaging with other researchers to encourage them to adopt a similar approach when collecting and formatting their own data.

Efforts within the European Union funded Project OPPORTUNITY recognised the challenges behind progressing research in activity recognition when multiple teams of researchers at different sites were both generating and using their own data [5]. A major limitation of this approach was recognised as the inability to make valid comparisons between approaches given the underlying data used for development purposes was different. The Project aimed to offer a common platform whereby researchers could avail of exactly the same data for training purposes and therefore could directly compare their independently developed activity recognition approaches. Although the efforts in this research were directed more towards the robustness of the activity recognition process itself, it further endorses the need to have publicly available datasets that can be used for developmental purposes.

Further rationale for the need to consolidate efforts in the production and sharing of datasets is evidenced by researchers sharing their datasets through online repositories such as the UC Irvine Machine Learning Repository [6] and PhysioNet [7]. At the time of writing this article the UCI Repository had 8 datasets (from over 350) which could be considered as supportive of the notion to share data within the activity recognition research community.

What is evident from all of these related works is that there is a clear appreciation that further efforts should be made towards the streamlining of the process of collecting and sharing data within the activity recognition and behavioural analysis research domains. Efforts have been focused on data formats, common data collection protocols, common sensing and aggregation platforms for the collection of data and finally the use of common data sets for the comparison of independently developed approaches for data analysis. What is now required is a further consolidated effort to bring all of these approaches together under one common initiative which is openly available within the research community.

3. Development of an Open Data initiative

In this Section an outline of the required and proposed elements of the ODI are presented in addition to outlining the next steps required in the realization of an operational solution.

3.1 Background to initiative

The ODI is being driven a consortium of researchers active within the field of Pervasive Computing from Ulster University (UK), Lulea Technical University (Sweden), Halmstad University (Sweden), University of Jaen (Spain) and the University of Twente (The Netherlands). It is a consolidation of skill sets from domains ranging from activity recognition, behavioural analysis, signal processing, machine learning, wearable computing and pervasive and mobile computing. In 2007 Nugent *et al*. proposed the homeML format as an open format for the storage and exchange of data within smart environments [3]. This work was further extended through collaborations with Lulea University in the development of open formats for the exchange of rules within smart environments and the development of visual editors to encapsulate domain knowledge [8]. At the time, however, the work fell short of wide spread penetration within the research community. During November 2015 the ODI was fully revived at a Workshop on Intelligent Environments Supporting Health and Wellbeing held at Halmstad University. The outcome of the workshop was the definition of the components required to deliver a functional end-to-end Open Data Initiative in pervasive healthcare. Figure 1 presents an overview of the components and activities required and planned by the ODI.

By embracing this initiative a number or benefits are expected as follows:

- faster progress in improving state of the art through usage of readily available datasets and avoidance of time in collecting and annotating new datasets.
- new insights gained from development of innovative data analysis techniques as a result of larger representative datasets being available.
- better value for money for funders with both data and results being made truly openly accessible.
- easier to initiate and benefit from national/international collaborations.

- increased likelihood of moving from the research domain to a scalable and extensible solution.
- increased replicability of experiments and reproducibility of results.

The following Sections outline the requirements and plans for each of these activities within the realms of the ODI.

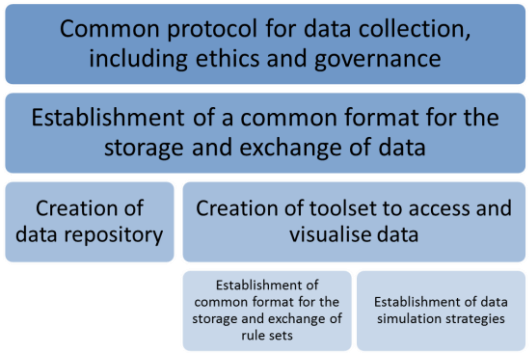

Figure 1. Overview of the components comprising the foreseen Open Data Initiative.

3.2 Common Protocol for data collection

To assist in avoiding duplication of efforts in data collection and to maximize diversity in the data sets from a cultural, societal and geographical perspective the ODI will strive to create a suite of openly available protocols specifying the technology platforms and sensing technologies to be used, the types of activities to be undertaken and exactly how these should be performed in addition to detailing the ethical implications to be considered. Researchers will therefore be able to collect data following exactly the same protocols as others. By following such an approach a central data set will be established and will incrementally grow with each new set of recordings being added once they have been validated.

This approach has been tested through the sharing of a common protocol between Ulster University and the University of Jaen. Using a simple installation of binary sensors a range of activities within a smart kitchen have been collected by two different sets of researchers following the same protocol, with all data being stored in the same repository. Figure 2 presents the smart kitchen environment where the data was recorded at Ulster University. This has resulted in a total of 396 instances across 9 activities being recorded with 32 instances being recorded by researchers at Ulster and 364 instances being recorded by researchers from Jaen [9, 10]. Plans are now underway to replicate the same experiments in smart labs in the University of Jaen and Halmstad University.

Figure 2. Smart kitchen used during data collection from 2 sets of researchers using the same data collection protocol.

3.3 Common format for storage and exchange of data

To assist in the storage and exchange of data within the ODI a common format is required. Given the previous work of the Authors in this domain and the ability of the homeML approach to successfully repurpose a number of datasets this approach will be reused. Figure 3 presents the xml schema to be used for the storage of the data. As a working group, one of the key items for consideration will be appropriateness of this format for data exchange and consideration of other potentially more suitable approaches.

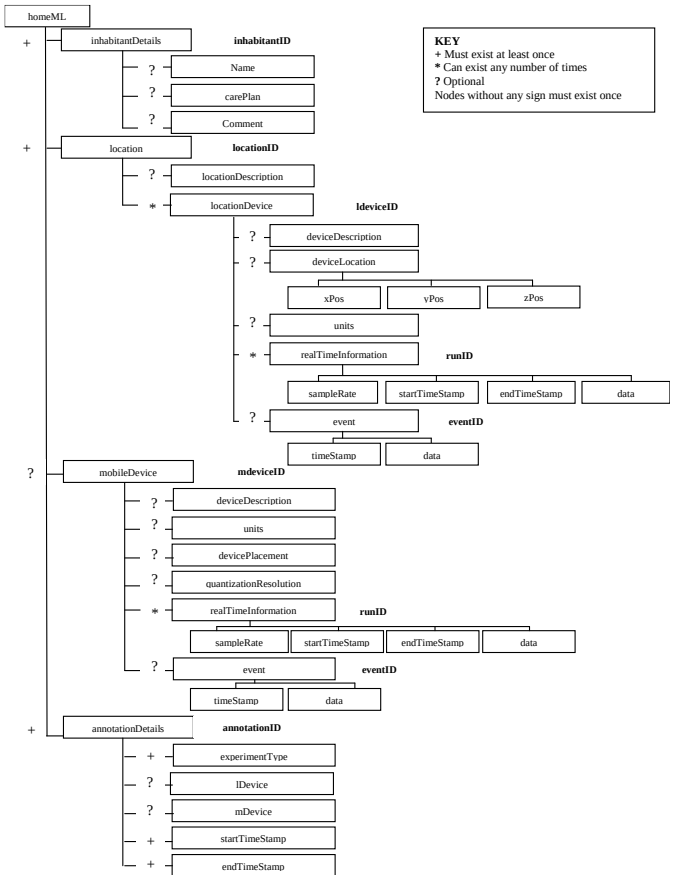

Figure 3. homeML schema to be used within the ODI as the common format for the storage of data.

3.4 Creation of data repository

The aim is to create an online repository where data sets in the domain of activity recognition and behavioural analysis can be both stored and added to. For the former a vetting process will be established whereby any new datasets will be validated for accuracy and reliability prior to making their content openly available. All datasets to be uploaded must be accompanied with a full description of the data and its annotations, the protocol used for data collection and specification of the technology platform used. This is to ensure that the experimentation can be easily replicated by others within the field and that the data generated and its annotations will be compliant with what has already been recorded. Once a datatset has been created and the protocol has been shared additional instances collected by other researchers can be added provided they show evidence of following the protocol and usage of identical technology platforms and data annotation strategies.

3.4.1 Creation of toolset for the access and visualization of data

To aid in the analysis of the datasets a suite of tools will be made available whereby datasets stored on the repository can be accessed, queried and visualized. This will assist researchers in investigating the distribution of datasets prior to accessing them.

3.4.2 Establishment of large scale simulated datasets

In addition to providing researchers with access to datasets collected within real environments the ODI will also create tools to produce simulated datasets based on the information stored within the repository. This provides large datasets representative of specific activities and behaviours to be made available to the research community [11, 12].

3.4.3 Testing with new approaches to Activity Recognition

The end point in the ODI is to provide Researchers with a range of datasets that they can use to support the development of activity recognition and behavioral modeling approaches and to have the ability to compare these within a unified evaluation framework. As presented in Figure 4, the ODI provides an initiative whereby multiple data sets can be collected using the same protocol and technology platforms can be aggregated and used by multiple independent researchers to develop activity recognition algorithms. With such an approach a single evaluation framework can then be used for comparative approaches. This permits performance metrics to be openly stored and referenced for all algorithms tested using the same data.

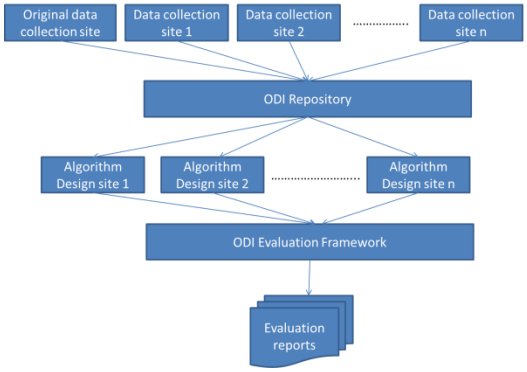

Figure 4. ODI framework for evaluation of activity recognition algorithms using a unified evaluation framework.

4. Conclusion

This paper has outlined the efforts of an international collaboration to progress an initiative to make data more openly available within the research community of activity recognition and behavioral modeling. Details of work undertaken to date and strategies for future work have been presented. The following challenges will underpin the further development of the ODI:

- how can we ensure that standards of data management are developed, promoted and entrenched so that research data can be shared routinely, and re-used effectively?
- what should be considered to ensure that data is collected in an ethical manner, in keeping with best practice?
- what frameworks should be implemented for data sharing to support the creation of an online repository?
- how can challenges around data labeling and ground truth annotation be overcome to ensure the high quality validation of data being considered?
- how can best practices be transferred both to and from other domains such as affective computing, persuasive computing, mobility analysis, cognitive computing, to name but a few.

The next immediate steps in the ODI will revolve around the creation of the repository and the establishment of a number of initial datasets to validate the overall concepts presented in this paper.

5. ACKNOWLEDGMENTS

Invest Northern Ireland is acknowledged for supporting this project under the Competence Centre Programme Grant RD0513853 - Connected Health Innovation Centre.

REFERENCES

[1] Chen, L. Hoey, L., Nugent, CD, Cook, DJ and Yu, Z. Sensor-Based Activity Recognition. 2012. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790-808.

[2] Ubihealth Project, http://www.ubihealth-project.eu/index.php, last accessed 8th March 2016.

[3] Nugent, CD et al., 2007. homeML – An Open Standard for the Exchange of Data Within Smart Environments, Volume 4541 of the series LNCS, pp 121-129

[4] McDonald, H.A., Nugent, C.D., Moore, G., Finlay, D.D.: An XML Based Format for the Storage of Data Generated within Smart Home Environments. In: *The 10th IEEE International Conference on Information Technology and Applications in Biomedicine*. pp. 1-4. (2010).

[5] H. Sagha *et al*., Benchmarking classification techniques using the Opportunity human activity dataset, *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, Anchorage, AK, 2011, pp. 36-40.

[6] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/, last accessed 08/03/16.

[7] PhysioNet: the research resource for complex physiologic signals, https://www.physionet.org/, last accessed 08/03/16.

[8] Nugent. CD *et al*., 2007. HomeCI - A visual editor for healthcare professionals in the design of home based care. *Engineering in Medicine and Biology Society, EMBS 2007. 29th Annual International Conference of the IEEE*, Lyon, 2007, pp. 2787-2790.

[9] Quesada, F, Moya, F, Medina, J, Martınez, L., Nugent, CD and Espinilla, M. 2015, Generation of a Partitioned Dataset with Single, Interleave and Multioccupancy Daily Living Activities, Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information Volume 9454 LNCS, pp 60-71.

[10] Shewell, C. Nugent, M. Donnelly, and H. Wang, 2016, "Indoor Localisation Through Object Detection on Real-Time Video Implementing a Single Wearable Camera," in *Mediterranean Conference on Medical and Biological Engineering and Computing*, pp. 1231-1236.

[11] Synnott, J. Chen, L., Nugent, CD and Moore, G. 2014, The creation of simulated activity datasets using a graphical intelligent environment simulation tool. *Engineering in Medicine and Biology Society* (EMBC), pp. 4143–4146.

[12] Lundstrom, J, Synnott, J, Jarpe, E., Nugent, CD. 2015, Smart Home Simulation using Avatar Control and Probabilistic Sampling, *The 2nd International Workshop on Smart Environments: Closing the Loop*, pp. 336-341.