Environment Simulation for the Promotion of the Open Data Initiative

Jonathan Synnott, Chris Nugent, Shuai Zhang, Alberto Calzada, Ian Cleland
School of Computing and Mathematics
University of Ulster
Jordanstown, Northern Ireland
{j.synnott, cd.nugent, s.zhang, i.cleland}@ulster.ac.uk, albertocalsa@gmail.com

Macarena Espinilla, Javier Medina Quero
Department of Computer Science
University of Jaén
Jaén, Spain
{mestevez, jmquero}@ujaen.es

Jens Lundström
Department of Intelligent Systems
Halmstad University
Halmstad, Sweden
jens.lundstrom@hh.se

***Abstract*— The development, testing and evaluation of novel approaches to Intelligent Environment data processing require access to datasets which are of high quality, validated and annotated. Access to such datasets is limited due to issues including cost, flexibility, practicality, and a lack of a globally standardized data format. These limitations are detrimental to the progress of research. This paper provides an overview of the Open Data Initiative and the use of simulation software (IE Sim) to provide a platform for the objective assessment and comparison of activity recognition solutions. To demonstrate the approach, a dataset was generated and distributed to 3 international research organizations. Results from this study demonstrate that the approach is capable of providing a platform for benchmarking and comparison of novel approaches.**

Keywords*— simulation; intelligent environments; data sharing; activity recognition

I. Introduction

The development and testing of novel approaches involving the processing of intelligent environment (IE) sensor data requires access to high quality, annotated and validated sensor datasets generated within IEs. One example is the development and evaluation of activity recognition approaches. This relies on test data for the assessment of the performance of new algorithms [1], models [2] and classification mechanisms [3]. Despite the demand for these datasets, the acquisition of high quality datasets is subject to several limitations [1]. The implementation of IEs is costly in terms of financial resource, time, and space. Considerable planning is required, and the optimum configuration may not be known prior to construction [1]. Additionally, these environments may lack flexibility as there are practical and ethical limitations dictating the reasonable modifications that can be made to an already existing environment. Comprehensive testing of novel approaches involves the collection of data describing all possible scenarios which may be encountered within an IE. This may not be possible due to recruitment, ethical and regulatory limitations [1], [4]. These issues with the collection of IE sensor data are detrimental to research progress and are slowing down advances in the development of new approaches and also increasing the time required to make real solutions available which can be deployed in real life scenarios [5].

This work aims to reduce the time currently being taken to make available, on a wide scale, activity recognition solutions that can be deployed within real life scenarios. The remainder of this paper is structured as follows. Section 2 provides an overview of the background of the research area, including existing work in the area of simulation and what is being referred to as the Open Data Initiative (ODI). Section 3 describes IE Sim, a software solution for IE data simulation. Section 4 describes an approach for the validation of the ODI through the collection and distribution of data generated within IE Sim. Section 5 presents the results and discussion, and Section 6 provides concluding marks with recommendations for future work.

II. Background & Related Work

This Section describes two key approaches which have the potential to address the data collection limitations identified in Section 1. These are: Environment Simulation, and the ODI.

A. Environment Simulation

The simulation of IEs for the generation of synthetic sensor datasets is one popular area of research which may be capable of addressing the aforementioned limitations [6]. Such approaches have the potential to accelerate research in related areas through the generation of vast sensor datasets [1]. These approaches offer increased control over the environment and the resulting data. Additionally, the layout of environments can be modified rapidly and without cost to adapt to the researcher's requirements. This allows researchers to quickly assess the impact of various sensor layouts and configurations without trial and error investment in expensive hardware. Experiments can be re-run repeatedly with small adjustments to the experiment protocol or environment. Simulations also facilitate adjustment of parameters such as time, catering for the generation of rapid generation of datasets spanning extended periods of time.

Such approaches ultimately facilitate rapid, robust and cost effective testing and evaluation of novel solutions, and studies that rely on simulation during the design phase are often more likely to include more robust designs [1]. Additionally, the digital nature of these simulation approaches promotes collaboration and open problem solving to a wider research community [1], particularly when combined with initiatives such as the ODI.

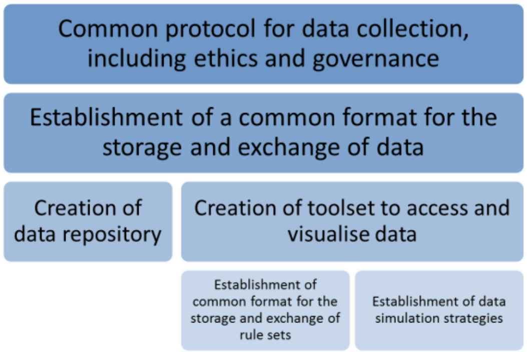
B. *The Open Data Initiative*

There is significant interest within the activity recognition community to improve upon the performance of solutions which are deployed for the purposes of both detecting and profiling activities of daily living within the home environment. The outputs of such approaches provide objective health assessments and dramatically reduce the amount of effort required by healthcare professionals in the care management process. The negative effect of this work has been that large amounts of efforts, all of which are largely similar, have been invested into designing experiments, collecting data and finally analysing the data. This has limited the overall size and diversity of datasets which are available to support data driven approaches for activity recognition.

As an effort to address this challenge a group of researchers have been proactive to establish the ODI. The ODI has as its main aim the ability to provide a structured approach to provide annotated data sets in an accessible format for the research community. By making such resources available a secondary aim is to reduce the gap between research efforts and real solutions for activity recognition which can be deployed in real life scenarios.

To date a range of efforts within the research community have attempted to define common data formats, common data collection protocols, common data aggregation platforms and approaches for comparison of analysis techniques. There is a general appreciation that further efforts should be made to streamline these efforts. 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. The ODI is being driven by a consortium of researchers active within the field of Pervasive Computing from Ulster University (UK), Luleå Technical University (Sweden), Halmstad University (Sweden), University of Jaén (Spain) and the University of Twente (The Netherlands). An overview of the main components identified for the ODI are presented in Fig. 1. Through embracing this approach it is expected that a number of benefits will be achieved:

Fig. 1. Overview of the components comprising the Open Data Initiative.


Common protocol for data collection, including ethics and governance
Establishment of a common format for the storage and exchange of data
Creation of data repository
Creation of toolset to access and visualise data
Establishment of common format for the storage and exchange of rule sets
Establishment of data simulation strategies

- 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.

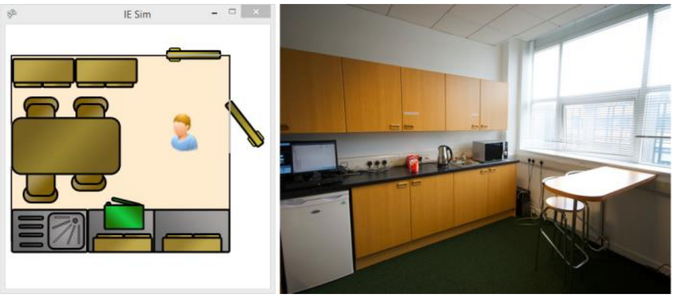
The approach demonstrated in this paper aims to combine environment simulation and the ODI approach in order to illustrate the potential to facilitate data generation, sharing, and objective comparisons between approaches developed by independent organizations.

III. Intelligent Environment Simulation

IE Sim has been designed to facilitate the rapid creation of simulated environments populated with objects and sensors [7]. It incorporates a visual, interactive approach designed for use by both technical and non-technical users to rapidly prototype novel environments and perform initial testing on novel activity recognition or assisted living approaches. The software provides a platform for the sharing of environments and the performance of repeatable experiments. This may facilitate collaboration, objective evaluation and comparison of data driven approaches by independent researchers. Fig. 2 provides an example of an environment created within IE Sim (Fig. 2 (a)). This environment was created to represent the smart kitchen (Fig. 2 (b)) located within Ulster University's Smart Environments Research Group [8].

Environments created within IE Sim are saved to an online repository in an XML format. Version control facilitates the creation of a variety of environment configurations associated with unique IDs. Any created dataset can be traced to the exact

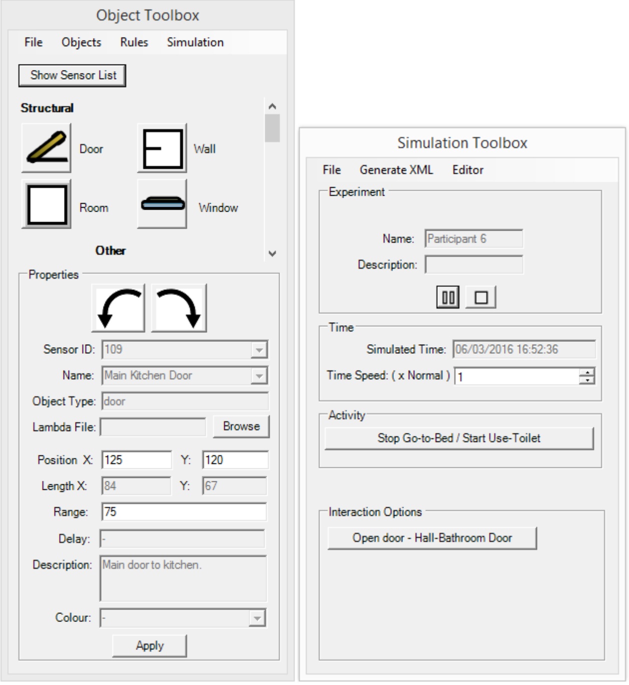
Fig. 2. An example of a simulated (a) and real (b) smart kitchen.



(a) (b)

configuration within which it was created, allowing experiments to be repeated and compared. This also facilitates assessment of the impact of small changes in the environment or activity performance on the success of data processing approaches. Environments are created using the object toolbox (Fig. 3 (a)). This toolbox provides a variety of objects such as rooms, walls, furniture and decorations, allowing users to sculpt a layout representative of a real environment. The toolbox also provides a variety of sensors including PIR sensors, pressure sensors, and contact sensors for use in objects such as doors, ovens, and refrigerators. Users can also create custom objects for inclusion within environments. Sensor parameters including detection range and firing frequency can be adjusted. The simulation toolbox (Fig. 3 (b)) facilitates experiment setup and execution. Users are able to assign an experiment name and description, start, pause, and stop an experiment, set the simulated time and the progression speed, interact with environment objects and annotate activity performance.

Fig. 3. The two main IE Sim interaction menus: (a) The object toolbox, (b) The simulation toolbox.



Sensor ID: 109
Name: Main Kitchen Door
Object Type: door
Lambda File:
Position X: 125 Y: 120
Length X: 84 Y: 67
Range: 75
Delay: -
Description: Main door to kitchen.
Colour: -

Name: Participant 6
Description:
Simulated Time: 06/03/2016 16:52:36
Time Speed: (x Normal) 1

(a) (b)

The arrow keys on a keyboard are used to navigate an avatar throughout a simulated environment. The avatar can passively or actively interact with the sensors within the environment. Passive interaction includes moving throughout an environment and entering the detection range of sensors such as passive infrared (PIR) sensors or pressure sensors. Active interaction includes the user explicitly interacting with objects such as doors, microwaves, or kettles through the use of a context menu when the avatar is within interaction range. An offline version of IE Sim has previously been evaluated by 21 international researchers, receiving positive feedback [9]. A collaboration with Halmstad University, Sweden, aimed to improve the realism of the data generated and to facilitate the rapid generation of datasets spanning extended periods of time [10]. In particular, this collaboration focussed on improving the realism of generated PIR sensor data. A joint avatar and probabilistic model was proposed to simulate both PIR sensor triggering. The number of PIR events per room and time interval was modelled by the Poisson distribution and implemented in the simulator by random sampling of the exponential distribution. Parameters of the adopted models were modelled from data acquired from a real-life setting. The results suggest that the proposed approach successfuly increased the realism of simulated data. IE Sim currently supports data output in multiple formats: Labelled activity vector, MySQL database records, and homeML, which is an open standard for exchange of data within smart environments.

IV. Approach

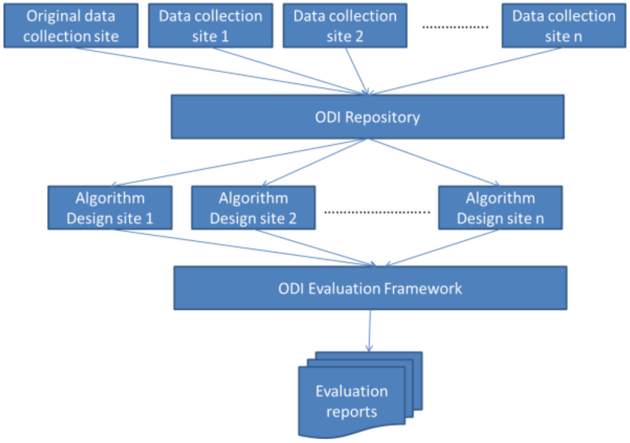
This study adopted the ODI methodology, which involves firstly creating a dataset, then independently providing data and finally using a common platform for objective assessment. This was completed in two phases: Simulated dataset generation using IE Sim, and analysis of the simulated dataset by independent researchers using a variety of activity recognition solutions.

As presented in Fig. 4, the ODI provides an initiative whereby multiple datasets can be collected using the same protocol and technology platforms can be aggregated and used by multiple independent researchers to develop independent activity recognition algorithms. The advantage of adopting such a methodology is that a single evaluation framework can then be used for comparative approaches. This was the methodology followed within the current work. In the first instance data was generated by a variety of researchers to produce an aggregated data set. Secondly, independent researchers from 3 organizations developed approaches for activity recognition which were then evaluated under the one common framework.

A. Phase 1 – Simulated Data Collection

The data collection phase of the study involved the recruitment of 8 participants who were staff, students, or visiting scholars of Ulster University's School of Computing and Mathematics. These participants used IE Sim for the first time after being shown a demonstration of an activity being completed. Prior to the study beginning, a simulated environment (Fig. 5) was created. This environment was used by the participants for the recording of activity completion.

Fig.4. ODI framework for generation of datasets and objective evaluation of activity recognition algorithms using a unified evaluation framework.



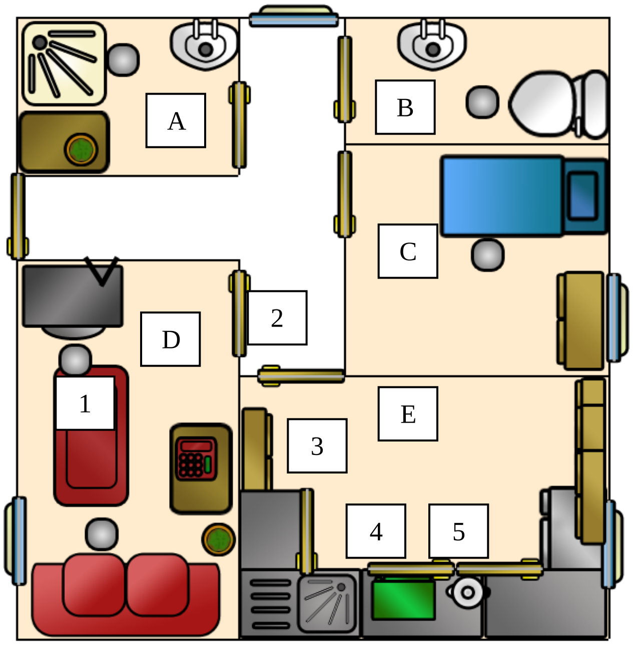
This environment was modelled upon a single floor residential environment, consisting of: a living room, kitchen, bathroom, toilet room, bedroom and hallway. The environment was populated with a range of objects including contact sensors, pressure sensors, and typical household objects such as a bed, shower and sinks.

Participants were given a list of 11 activities to complete, including the key tasks within each activity. The series of activities was completed 7 times by each participant. The activities were: Go to bed, Use toilet, Watch Television, Prepare Breakfast, Take Shower, Leave House, Get Cold Drink, Get Hot Drink, Prepare Dinner, Get Dressed and Use Telephone. These activities were selected in order to promote interaction with a wide range of sensors located throughout each room within the simulated environment. The following is an example of the instructions for the "Prepare Dinner" activity:

1. Walk into the kitchen
2. Open/Close the freezer
3. Open/Close the groceries cupboard
4. Open/Close the plates cupboard
5. Open/Close the cups cupboard
6. Open/Close the Microwave
7. Last step - Press the STOP "Prepare Dinner" / START "Use Telephone" button

Participants were instructed that the steps did not need to be completed in the order provided, however, should be completed in a logical order. Participants were required to manually annotate activity performance by pressing a "Stop [Activity *n*] / Start [Activity *n* + 1]" button once each activity was complete. This resulted in the corresponding activity name being inserted at the correct position within the generated dataset. The researcher supervising the trial directly monitored the completion of the first two activities to ensure correct completion. Participants were then instructed to complete the remainder of the activities without direct supervision. The researcher remained in the same room as the participants in order to answer any queries that arose.

Fig. 5. The simulated environment used for data collection. This environment was created using the IE Sim software. Areas: (A) The shower room, (B) The toilet room, (C) The bedroom, (D) The living room, (E) The kitchen. Sensors include pressure sensors such as (1), and contact sensors in doors such as (2), cupboards (3), a microwave (4) and kettle (5).



B. *Phase 2 - Data Analysis*

The data analysis phase involved distribution of the simulated dataset to independent researchers from three organisations. These were: two researchers from the University of Jaén, a Visiting Scholar with Ulster University who also holds the position of senior data science engineer in UK industry, and a researcher from Ulster University who has not previously been involved in the development of IE Sim. Each researcher is actively involved in the creation of novel data driven approaches to activity recognition. Each researcher was provided with a training dataset which consisted of the first 5 performances of each activity by each user (71.42% of the total data n=220), and a test dataset which consisted of the last 2 performances of each activity by each user (28.57% of the total data n=88). The researchers were asked to train and test their activity recognition approaches, and provide an overview of the resulting classification accuracy.

The researcher from Ulster University analysed the dataset through the use of neural networks and deep neural networks. Neural networks are non-parametric approaches that can implicitly detect complex nonlinear relationships between data and their classifications. Here, a Multilayer Perceptron with one input layer, one output layer and one hidden layer with ten hidden neurons was created. The network performance was measured using the cross-entropy cost function. The network was trained on the training dataset using scaled conjugate gradient backpropagation [11] for the update of the weights and bias values. Network performance on the validation dataset was used as one of the stopping criteria for training to reduce over-fitting.

Deep Learning refers to models that are composed of multiple layers of non-linear information processing, for data representations with multiple levels of abstraction, where each layer processes the outputs of the previous layer for pattern analysis and classification. Deep neural networks have drawn much attention in recent years due to their outstanding advancement to the state-of-the-art for tasks in a range of application areas including computer vision, speech recognition and natural language processing. Relatively speaking, there is less effort on the exploration of using deep learning approaches for activity recognition in IEs. Deep learning approaches may have the potential to uncover the complexity and dynamics of human activities of daily living captured from the environment, from sub-activities encoded in lower layers to more complex activities in upper layers [12]. Here, a deep neural network was constructed with two stacked autoencoders and a third fully connected hidden layer. An autoencoder is a neural network learned in an unsupervised manner to derive an abstract representation of the data revealing interesting patterns. The training of the deep neural network was carried out using the cross-entropy cost function and scaled conjugate gradient backpropagation. L2 weight regularization for autoencoders was used to prevent over-fitting.

The senior data science engineer from UK industry analysed the dataset through the use of the Dynamic Instance Activation (DIA) approach. This approach is a generalized version of the Dynamic Rule Activation approach [13] which uses partitioned datasets of activities of daily living. The DIA approach was designed to maximize activity recognition accuracy by optimizing a set of similar activities when compared to a sequence of activated sensors.

Finally, the methodology for activity recognition used by the researchers from the University of Jaén consisted of a knowledge driven, fuzzy rule-based inference engine in real time. This methodology used the fuzzy linguistic approach [14] as a solution to improve the management of temporal information in knowledge-driven approaches to activity recognition in real-time. The methodology analyses the data streams generated from sensor devices based on a rule-based inference engine, exploiting a fuzzy linguistic approach to model the temporal information. The Linguistic Rule-Based Inference is based on a collection of fuzzy temporal logic rules in the form of IF-THEN statements, where each antecedent includes linguistic terms from sensor streams and the temporal linguistic term in the form of an adverbial -WHEN- that determines when the measures of sensor streams have been collected. The inference engine determines the degree of matching of antecedents based on the states of sensor streams, assigning them to the activity consequence in real time. The main advantage of this approach is the flexibility of fuzzy temporal rules for modelling the sensor information, providing, furthermore, interpretation for defining rules by the knowledge experts. One such rule used by the approach in this study is shown in (1).

```
IF (Cups Cupboard IS activated WHEN recently)
AND (Refrigerator IS activated WHEN recently)          (1)
THEN activity IS GetColdDrink
```

V. Results & Discussion

During the simulated data collection phase, participants, although previously unfamiliar with the IE Sim software, universally found that they were quickly able to learn to use the software after the completion of several activities. They quickly became familiar with the environment layout, and were able to find and identify the majority of objects and sensors within the environment. Participants typically required approximately 1 hour to complete 7 iterations of the series of 11 activities. Whilst the participants found IE Sim intuitive to use, there were some unforeseen usability issues. The main issue was with regards to self-annotation of activities. 80% of participants incorrectly annotated at least one out of the 77 activities they completed. These errors included forgetting to press the annotation button once an activity was complete, pressing the annotation button instead of an object interaction button, and accidentally pressing the annotation button twice. This resulted in sensor activations being assigned to either the previous or next activity. It should be noted that the majority of participants who made annotation errors noticed they had made an error soon after it had occurred. In the data collected from 4 participants, these annotation issues were so significant that the data were not included within the dataset used in Phase 2. The data from another 2 participants only had minor issues and the data from the remaining 2 participants were annotated correctly without issue. All participants occasionally skipped steps in activities, or completed extra steps. Such errors were not manually corrected prior to providing the data to the participating institutions for the data analysis phase. This was to ensure that the data being analysed accurately reflected the quality of typical data output by IE Sim in its current state.

The neural network approach used by the researcher from Ulster University achieved the highest classification accuracy of 97.72%. The results demonstrate that even with only one hidden layer, the network has performed exceptionally well, with two misclassifications to the activity "Take Shower" from the activities “Prepare Breakfast” and "Leave House". One common drawback of using the neural network approach is the local minima problem from the weight adjusting with a gradient descent. Therefore, repeated training with random starting weights has been attempted to address this problem. The deep neural network approach achieved an accuracy of 96.59%.

Unfortunately, the potential of the deep neural network has not been shown by its performance in this study. Two of the misclassifications are the same as the performance of the neural network discussed previously. One of the fundamental factors to the success of deep learning approaches in general is the availability of large amounts of data for training. There are a large number of parameters to be learned in the given network as a result of the complexity of the network structure. The size of the training data could be insufficient for this deep neural network to achieve its optimal performance given the size of parameters required for training. This will be addressed in the future work following this success of our testbed trial.

The DIA approach used by the senior data science engineer from UK industry also achieved a classification accuracy of 96.59%, providing similar results to those obtained from the deep neural network approach. Misclassifications by this approach included the classification of “Get Cold Drink” as “Leave House” (1 instance), “Leave House” as “Take Shower” (1 instance) and “Prepare Breakfast” as “Take Shower” (1 instance).

The fuzzy rule-based approach used by the researchers from the University of Jaén also achieved an accuracy of 96.59%. Misclassifications by this approach included “Get Dressed” as “Use Telephone” (1 instance), “Prepare Dinner” as “Get Hot Drink” (1 instance) and “Prepare Breakfast” as “Take Shower” (1 instance).

Some of the classification errors produced by these approaches were as a result of the aforementioned annotation issues, in which some sensor activations were incorrectly assigned to the wrong activity. Additionally, the annotation button implemented in IE Sim for this study simply ended the current activity and immediately began the next activity. The result of this was that occasionally the last sensor activation of the previous activity would be repeated as the first sensor activation of the following activity. This was particularly prevalent in activities which included the use of a pressure sensor. For example, the last step in one activity may have been to sit on a sofa. When beginning the next activity, the user may still have been sitting on the sofa.

VI. Conclusions & Future Work

This paper has introduced an initiative to address the existing limitations with access to high quality annotated and validated IE datasets. IE Sim was used to generate a dataset describing the completion of 616 activities within a simulated environment. Following the ODI approach, this dataset was provided to 3 independent organizations. Researchers from these organizations used this dataset to successfully evaluate and compare the performance of their activity recognition algorithms. As such, this has validated the combination of data simulation and the ODI as a platform for objective benchmarking.

Future work will aim to make IE Sim a fully online platform to further promote collaboration and sharing of datasets in line with the ODI approach. Additionally, the annotation mechanism within IE Sim will be further refined in order to reduce the likelihood of annotation errors occurring.

Acknowledgment

Invest Northern Ireland is acknowledged for partially supporting this project under the Competence Centre Programme Grant RD0513853 - Connected Health Innovation Centre. Additional support was provided by the Department of Employment and Learning, Northern Ireland.

References

[1] S. Helal, J. W. Lee, S. Hossain, E. Kim, H. Hagras, and D. Cook, “Persim - Simulator for Human Activities in Pervasive Spaces,” in *2011 Seventh International Conference on Intelligent Environments*, 2011, pp. 192–199.

[2] M. Youngblood, D. J. Cook, and L. B. Holder, “Seamlessly Engineering a Smart Environment,” in *2005 IEEE International Conference on Systems, Man and Cybernetics*, 2005, vol. 1, pp. 548–553.

[3] M. Buchmayr, W. Kurschl, and J. Küng, “A Simulator for Generating and Visualizing Sensor Data for Ambient Intelligence Environments,” *Procedia Comput. Sci.*, vol. 5, pp. 90–97, 2011.

[4] M. P. Poland, C. D. Nugent, H. Wang, and L. Chen, “Development of a smart home simulator for use as a heuristic tool for management of sensor distribution,” *Technol. Heal. Care*, vol. 17, no. 3, pp. 171–182, Aug. 2009.

[5] S. Helal, E. Kim, and S. Hossain, “Scalable Approaches to Activity Recognition Research,” in *Proceedings of the 8th International Conference Pervasive Workshop*, 2010, pp. 450–453.

[6] J. Synnott, C. Nugent, and P. Jeffers, “Simulation of Smart Home Activity Datasets,” *Sensors*, vol. 15, no. 6, pp. 14162–14179, Jun. 2015.

[7] J. Synnott, L. Chen, C. D. Nugent, and G. Moore, “The creation of simulated activity datasets using a graphical intelligent environment simulation tool,” in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 4143–4146.

[8] C. Nugent, M. Mulvenna, X. Hong, and S. Devlin, “Experiences in the development of a Smart Lab,” *International Journal of Biomedical Engineering and Technology*. Inderscience, 28-Jan-2009.

[9] J. Synnott, L. Chen, C. Nugent, and G. Moore, “IE Sim – A Flexible Tool for the Simulation of Data Generated within Intelligent Environments,” *Lecture Notes in Computer Science*. Springer, 01-Nov-2012.

[10] J. Lundström, J. Synnott, E. Jarpe, and C. D. Nugent, “Smart Home Simulation using Avatar Control and Probabilistic Sampling,” in *The 2nd IEEE PerCom Workshop on Smart Environments: Closing the Loop in conjunction with PerCom 2015*, 2015.

[11] M. F. Møller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[12] F. Ordóñez and D. Roggen, “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition,” *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[13] A. Calzada, J. Liu, H. Wang, and A. Kashyap, “A New Dynamic Rule Activation Method for Extended Belief Rule-Based Systems,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 880–894, Apr. 2015.

[14] R. Mikut, J. Jäkel, and L. Gröll, “Interpretability issues in data-based learning of fuzzy systems,” *Fuzzy Sets Syst.*, vol. 150, no. 2, pp. 179–197, 2005.