Understanding the Quality of Calibrations for Indoor Localisation

Ryan McConville, Dallan Byrne, Ian Craddock, Robert Piechocki, James Pope and Raúl Santos-Rodríguez
University of Bristol
United Kingdom
{ryan.mcconville,dallan.byrne,ian.craddock,r.j.piechocki,james.pope,enrsr}@bristol.ac.uk

***Abstract*—The efficient and effective deployment of Internet of Things (IoT) systems in real world scenarios remains a challenge, particularly in applications such as indoor localisation. Various methods have been proposed recently to calibrate localisation systems, ranging from precise but time consuming processes to those involving little explicit calibration based on a crowdsourced collection of data over time. However it is not clear how to estimate and compare the quality of a specific instance of a calibration. In this paper we present a simple yet effective method of calibrating a Smart Home in a Box (SHiB) together with a framework to combine calibrations while assessing their quality. Our empirical results demonstrate that our calibration method can be performed by untrained users in a short period of time yet is capable of up to 92% accuracy in room level localisation on free living experimental data.**

I. Introduction

Due to the the proliferation of the Internet of Things (IoT), connected devices within home environments are becoming increasingly common. These devices can both interact with and collect vast amounts of data on the surrounding environment and thus naturally have many useful applications ranging from transport to healthcare. The use of wireless sensor networks facilitates many tasks, including behaviour and activity monitoring [1], environmental monitoring [2] and indoor localisation [3]. The effective deployment of such systems remains a challenge, with many requiring installation and calibration by experts.

Recent work in the so called Smart Home in a Box (SHiB) area aims to provide smart home capabilities that can be deployed and used quickly without extensive training. One such system is the EurValve SHiB [4] consisting of four gateways which receive accelerometer data from a wrist worn wearable device over Bluetooth Low Energy (BLE). The primary objective of the system is to perform activity monitoring and localisation; it is currently under deployment in the homes of 60 patients undergoing heart valve surgery. In order to do this the wrist worn wearable samples accelerometer readings at 25Hz and sends the data to each gateway. We refer the reader to Appendix. A for a detailed description. Importantly, the Received Signal Strength Indication (RSSI) from the wearable is recorded at each gateway and thus provides a potential means of determining the senders location. RSSI is attractive as it is typically provided by the radio hardware and thus inexpensive and readily available to use. However RSSI is often considered unreliable for localisation purposes [5] because the surrounding environment has a significant influence on the calculated RSSI values. Radio frequency noise, interference of objects and a constantly changing environment, e.g. due to the movement of people, cause RSSI values to exhibit a strong variability. Nonetheless there have been a number of approaches in the literature attempting to solve these issues and utilise this commonly available signal [6], [7].

However, the initial calibration of the system is one of the key remaining challenges. There exist two main strands of work in this area; the first involves a detailed calibration, usually by an expert and may require additional information such as floorplans. The second approach attempts to perform the calibration in a (partially) unsupervised manner. The downside of the former is that it is often a time consuming and convoluted process requiring an expert, while the latter usually relies on crowdsourcing data gathered from multiple people over time [8]. Both contradict our initial aim of a SHiB deployed and setup in their home by individual users.

The main contributions of this work are as follows.

- We propose an quick and easy method of calibrating a SHiB for indoor localisation.
- We analyse the influence on localisation performance of the skill of the individual carrying out the calibration.
- We show that the experience of the calibrator does not necessarily improve the performance as evidence that our calibration process is simple.
- We present a framework to combine (when available) several calibrations, based on the concept of learning from crowds, to evaluate the quality of each calibration while achieving a better performance.
- We make available for the community a new dataset for indoor localisation including calibrations by ten individuals as well as a free living dataset[1].

II. Related Work

Indoor localisation in Smart Homes has been thoroughly studied in the past decade. An interesting overview of numerous techniques were proposed in the Microsoft indoor localisation competition in 2014 [3]. They conclude the indoor localisation problem is not solved with one of the major issues being the deployment overhead; with an average setup time of 5 hours.

[1]The dataset can be downloaded at https://seis.bris.ac.uk/ rm17770/calibration/

Although common approaches propose the use of radio propagation methods, these tend to require expert knowledge and many precise measurements taken for calibration[9], [10]. These types of methods are not compatible with our ease of deployment requirement.

Alternatively, fingerprint based techniques such as [11] introduce methods for room level localisation although require floorplan. Similarly, HIWL [7] requires knowledge of the rooms and a time consuming calibration process to train a Hidden Markov Model based approach on RSSI data. The authors use a programmed node that collects RSSI values throughout the room, taking over 5 hours. Along the same lines UMLI [6] uses hierarchical clustering to first achieve floor level localisation, followed by finer grained room level localisation. However, they rely on the availability of a large number of access points. Finally, these approaches can be extended with the use of additional data sources (e.g., GPS [12]). These algorithms do not meet our goal of not relying on extensive prior layout knowledge.

Very recently, Zhen et al. have presented BigLoc [13] which specifically attempts to solve the localisation problem in large indoor spaces using hundreds of access points for floor level localisation. Chen et al. propose a method Graph Loc [8] which requires no upfront calibration. Their graph based method does however require a physical floor plan and crowdsourcing in order to collect a large amount of RSSI data. Similarly Shimosaka et al. [14] propose an approach which requires a floorplan. We refer the reader to a recent review [15] which discusses existing localisation methods as well as practical deployment concerns, mainly reducing the fingerprinting step (site survey), the calibration of heterogeneous devices, and energy efficiency. For our approach, the first and last concerns are of the up-most importance. In terms of energy efficiency our system utilises a wearable with approximately 3 weeks of battery life. Our fingerprinting step is also designed to be as simple as possible such that untrained users can deploy and calibrate it, yet also provide enough information that we can accurately perform room level localisation.

III. A SIMPLE CALIBRATION

The key to the effective use of our SHiB for localisation is the calibration process that will provide us with labels we can use to build a model for each deployment. We have a number of constraints that must be taken into account by any calibration process. Firstly, we have no knowledge of the home environment the kit will be deployed in. Secondly, the kit will be deployed by the user. Our typical user profile is expected to be elderly and of potentially poor health. The users are chosen due to medical necessity and thus may not be technically capable. In order to gather a fingerprint for the entire room, ideally the user could walk around the room to cover the whole area. In order to reduce the complexity of this task, we just ask the user to perform the activity that they most commonly undertake in each room. This might not provide us with the best coverage but it is simpler while also characterising the habitual use of the room. Finally, the user will have the system for a period of time and there is no guarantee that they will perform the calibration immediately upon installation. Thus the method must also result in a clear signal that a calibration is taking place so the calibration labels can be extracted. The entire calibration process should take around around ten minutes.

A. Installation

The user deploys five pieces of hardware, a 4G router for data transfer and four gateways that are responsible for receiving data from the wearable. Three of the gateways are labelled with common rooms of interest in a residential setting; the living room, kitchen, and bedroom. A fourth gateway is provided for the user to place at a location of their own choosing.

B. Calibration

After each gateway is plugged into a power socket and turned on, the user must complete the following steps. For each gateway, the user places the wearable very close to the gateway for ten seconds. This causes a spike in the strength of the RSSI. The purpose of this is to provide a landmark in the data such that each calibration can easily be automatically extracted. This landmark can be used for automatic extraction and segmentation of the calibration. Following this, the user then performs a specific activity in the room for two minutes. The living room consists of sitting down, the kitchen involves walking around, the bedroom requires lying down and the final gateway, standing. The reason for this is twofold. The first is that these activities loosely correspond to the expected usual activities carried out in the room. The second reason is that, but not within scope of this paper, we plan to perform activity recognition and this provides us with labels for four different types of activity.

IV. MEASURING THE QUALITY OF A CALIBRATION

1) Data Collection: In order to provide a better understanding of our system and move towards a measure of the quality of a calibration, we collected data in the smart home environment of the SPHERE house [16] with the EurValve SHiB deployed. There, we instructed ten users to complete a full calibration. In all our our experiments the mean RSSI value from each gateway is computed over a one second window.

2) Evaluation of Individual Calibrators: Our first task analyses each calibration individually in order to estimate the localisation performance possible for each user. We will do this via stratified 10-Fold cross-validation on each individual calibration.

As we can see in Fig. 1 that the performance of each calibrator varies. Nonetheless with many of the classifiers the performance of all calibrators is similarly high. One source of error is related to the unreliability of RSSI. As we can see in Fig. 3, it is not always the case that the largest RSSI corresponds to the closest (in the same room) gateway. For instance, the living room and bedroom RSSI values in the first quarter of the chart in Fig. 3((a)) are somewhat interlaced, even though the participant is mostly static sitting within the living room. In comparison, the other RSSI plot in Fig. 3((b)), produced just thirty minutes later, the signals have much more separation. The floor plans of the building can be found in [17], where the living room and bedroom correspond to the 'lounge' and 'bedroom 2' respectively.

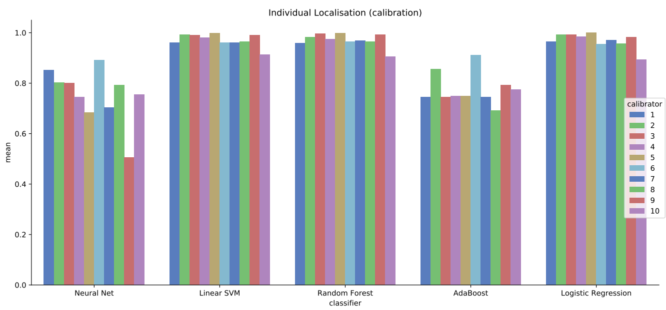


Fig. 1: Calibration performance over a number of classifiers for all calibrators.

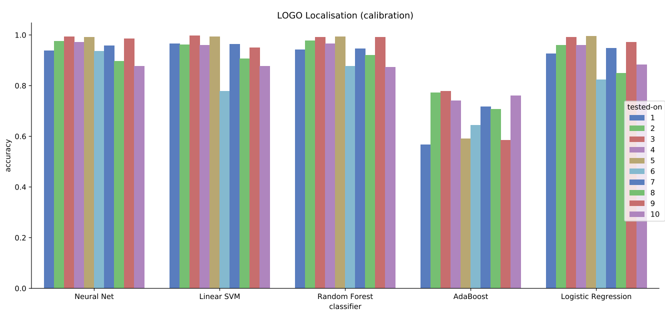


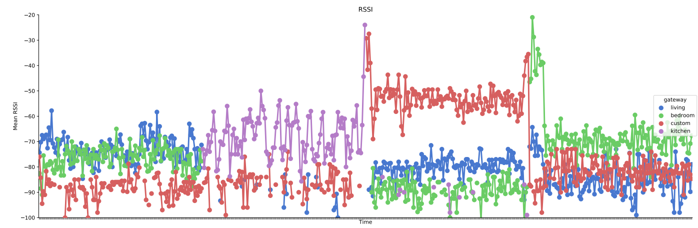
Fig. 2: Calibration performance when using a Leave One Group Out approach. A group here corresponds to a calibrator.

We also conduct a leave-one-out evaluation by training on the calibrations of 9 calibrators and testing on the final withheld calibrator. In Fig. 2 we can see that there is a higher degree of variability in the predictive performance across calibrators. For example, testing on some calibrators achieves almost perfect accuracy, while others achieve under 80%.

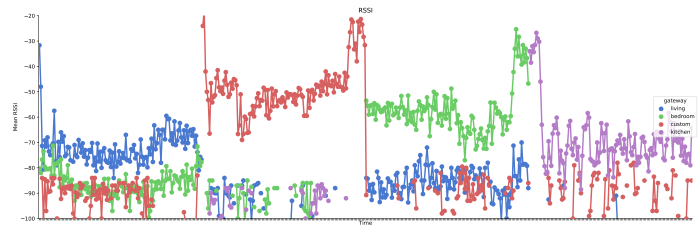
In fact, this intuition can be seen in the overlap present in the Principal Component Analysis (PCA) plots taken from the same two calibrations. Fig. 4 shows the difficulty in separating the living room and bedroom RSSI vectors even though they are on separate floors. This is most obvious in the worst performing PCA plot.

3) Evaluation of Calibrator Experience: In this section we would like to investigate if the experience of the calibrator has any influence on localisation performance. Intuitively we would expect that carrying out the same calibration steps again would result in as stronger adherence to the process, and thus perhaps, increased performance. This is of particular interest as our system is designed to be easy to set up, and as part of the EurValve protocol the participants will be setting up the system several times, over a period of approximately nine months. In order to study the effect of calibration experience we asked three calibrators to perform a second calibration on a different day than their first. Each time they followed the same calibration steps.

From Fig. 5 it is clear that calibration experience does not significantly improve the predictive power for localisation. In fact, only one of the three calibrators saw an improvement in performance with a subsequent calibration. This same calibrator then performed a third calibration, again on a different day than the previous two. The results of which can be seen in Fig. 5(b). In this third calibration the performance decreased. From these empirical experiments we see that there is no clear correlation between calibration experience and the resulting localisation performance. Further we see this as initial evidence that our IoT system and calibration process meets our objective of being straightforward and simple to calibrate.

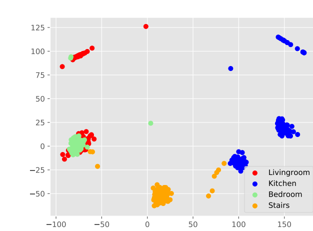


((a)) In the first quarter the livingroom and bedroom RSSI values are often intertwined despite being stationary in the livingroom.

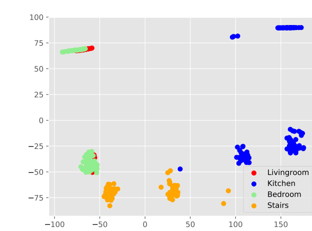


((b)) A much better seperation of RSSI values from each room is clear in this calibration.

Fig. 3: The RSSI values received during the calibration of two calibrators.



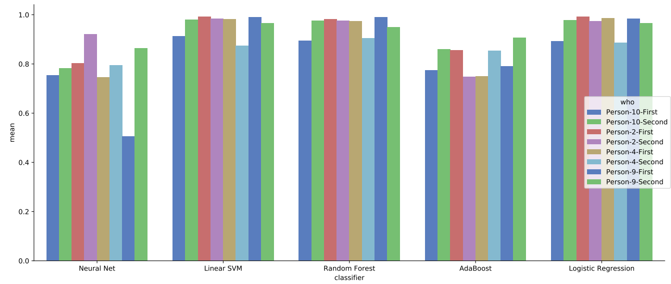
((a)) The calibration with the highest accuracy.



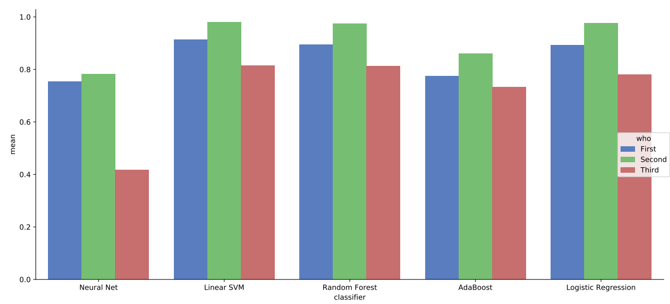
((b)) The calibration with the lowest accuracy.

Fig. 4: RSSI PCA plot from the best performing calibration (a) and worst performing (b).

4) Evaluation on Free Living Data: So far our evaluations have only been performed on data gathered as part of the calibration process. To properly evaluate our calibration process we collected data where one participant was carrying out unscripted 'free living' within the house for a number of hours. In our test data this included common activities across rooms such as making and eating lunch and sitting on the sofa. This makes a suitable test dataset to establish each of the calibrators predictive performance on, as the test participant was carrying out natural activities in different parts of the rooms and indeed house. The ground truth for this dataset was gathered with the



((a)) Three calibrators each with two separate calibrations.



((b)) One calibrator over three separate calibrations.

Fig. 5: The calibration performance of calibrators over 2 (a) and 3 (b) separate calibrations performed on different days.

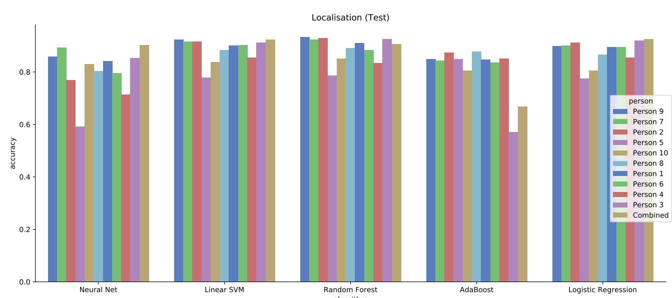


Fig. 6: The performance of each calibrator where a linear SVM was trained on their calibration and tested on a free living experiment.

use of a 4K video camera and numerical tags placed on the floor of the house. There were typically five tags per room including one at the entrance to the room.

For each of the ten calibrations we trained a number of classifiers and then tested the models on the free living test dataset. From Fig. 6 we can see that for nearly all classifiers we achieve good predictive performance for most of the calibrators. In two of the confusion matrices from the free living experiment it is clear that a number of the errors come again from the difficulty in distinguishing the living room and bedroom gateways. Again this corresponds with what we discovered previously in Sec. IV-2.

Looking closer at the results, and using the linear SVM as an example, the majority of calibrators achieve accuracy scores of over 90%. However, one calibrator achieves a score of 78%, a drop of around 12%. Thus, it is clear that not all

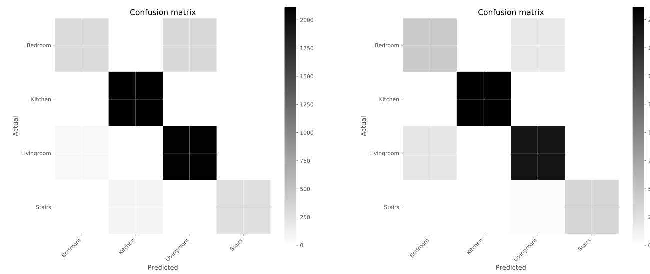


Fig. 7: Confusion matrices from two of the calibrators. Here it clear there is varying levels of misclassification between the livingroom and bedroom in both and the kitchen and stairs in one.

calibrations are equally good on the test set. This has been a consistent trend in our investigations thus far. In the following section we will explore a framework to determine the quality of each calibration while improving performance by learning from the combination of calibrations.

V. Combining Multiple Calibrations: Learning from Crowds

The idea of utilising the knowledge of crowds [18] has became increasingly attractive with online crowd sourcing services that allow many people to carry out requested tasks. Researchers in Machine Learning have used such services to crowdsource the annotation of labels for their datasets. The result is instances from the dataset are labelled by a number of annotators. One major problem however is that the annotators may have various levels of expertise and thus the labels are of varying qualities.

We propose that our problem can be framed in a similar way; specifically each calibrator can be considered an annotator. In our protocol we receive a number of calibrations per user, and while we have shown that expertise does not necessarily increase with experience there is a clear variability in the quality of each calibration. Thus an approach such as this may help improve the overall performance by weighting each calibration by its effectiveness. However in our proposed approach the calibrator is not *directly* annotating the test subjects location in the test set. Rather their annotations on the test set are inferred from their calibration. In order to infer these annotations we use the predictions of a classifier trained on their calibration which is then then used to predict their annotations for the test set. In other words, their annotations are the predictions of a model built on their calibration. The result is ten annotations for each second of the test set.

We will have two types of error in the annotations due to the, (1) varying ability of the calibrator and, (2) the model itself. Importantly the quality of the annotations vary in a similar way to that of traditional crowdsourced annotators. Furthermore, as we have previously shown that there is often misclassification between certain rooms, we propose that given different annotations the 'crowd' may be able to rectify this.

In order to evaluate the learning from crowds approaches we test each methods performance at estimating the single ground truth and build the model on part of the test set, as well as the final model performance on an unseen portion of the test dataset. As baselines two majority voting based methods were used. In hard instance (MVH) the majority decision by all of the annotators are chosen as the label. In the soft instance (MVS) the decisions of the annotators for each instance are probabilistic. Two less naive methods are also employed to estimate the quality of the annotators, specifically that of Raykar et al. (Raykar) [18] and Rodrigues et al. (MA) [19]. The Raykar method learns the ground truth labels and the classifier jointly. They treat the unobserved ground truth labels as latent variables and use Maximum Likelihood Estimation to find the parameters for their model by estimating the ground the posterior distribution in order to determine the quality of annotators and their logistic regression parameters. The MA method primarily differs in that the reliability of the annotators is treated as a latent variable.

The average performance for each approach on the test set can be seen in Fig. 8 along with the average accuracy of each individual. In general the best performance, up to 6 percentage points higher, is found by the probabilistic and learning from crowd approaches. As we have the annotations for the test set (note that these are not used during training) we can calculate the average accuracy of the annotators on the test set as 87%. This is also reflected in the poorer performance of the hard majority voting approach which discards any probabilistic information about an annotation; the annotator was either correct or not. The probabilistic majority voting (MVS) scored 92.5%. The MA method achieved the highest accuracy with 93.1%, with Raykar achieving 92.8%. The better performance of the MA and Raykar methods can be explained due to the fact that they took into account the reliability and quality of the annotators and thus achieved better performance on the unseen test set. Note that while the first two bars in Figure 8 show the mean performance, the performance of the individual calibrators ranged from 75% to 92%. The mean of 87% reflects the fact that most scores were indeed much closer to the higher end than the lower. The best learning from crowds approach, MA, was able to improve upon the true ground truth performance of all of the annotators. The highest performing annotator was 92% in the ground truth of the test set while the MA model achieved 93%.

A full evaluation of the performance of the methods are out of scope of this paper, given space limitations, however it is clear that by utilising the multiple annotations we can achieve higher levels of accuracy for localisation. As we have seen in Sect. IV-3 while an annotator may not improve with experience, methods involving utilising all of the annotations leads to improved performance in all cases.

VI. Conclusion

In this work we proposed a simple and efficient calibration method for the deployment of a SHiB kit that is capable of accurate room level localisation. We performed the calibration process with ten different participants and thoroughly evaluated machine learning methods on the individual and combined calibrations in order to evaluate their quality for localisation. Given the simplicity of the calibration, we also discovered that experience does not necessarily improve the calibration process and thus it appears that there is nothing to be gained from the use of an experienced calibrator with our system and calibration process. In addition we evaluated the calibration process on an unscripted free living experiment and found the best calibration could achieve up to 92% accuracy. Finally, we explored the use of learning from crowd based techniques that could exploit the fact that each calibration will be carried out numerous times by the same participant over a period of time. The results of these techniques were promising in that the localisation performance could be improved to 93% from the mean performance across individuals on the same test set of 87%.

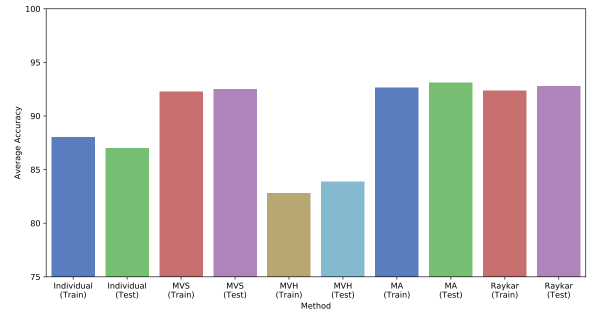


Fig. 8: A comparison of the average accuracy of each method on the test set where each method had no access to the test set annotations. The mean accuracy of the true annotations by the individual annotators is also included.

Appendix A
EurValve Smart Home in a Box

The EurValve system [4] was designed to be a scalable smart home in a box (SHiB). The system is to be deployed in the homes of 60 patients who will be undergoing heart valve surgery and will collect data before and after the surgery. It is low cost and relatively easy to install and use. The system was designed such that the patients would be able to deploy the kit themselves with as little instruction as possible. Due to design constraints the system only provides accelerometer data and RSSI for each gateway. The system is comprised as follows.

1) Wifi/Cellular router: The router provides Internet access (over the cellular link) for accessing network time protocol servers, submitting gateway system monitoring information, and transferring gateway daily log files to the central database. The gateways communicate with the router over WiFi.

2) Wrist worn wearable: The wearable is comprised of an integrated microprocessor and Bluetooth Low Power radio chip, flash module, and a three dimensional accelerometer. The flash module can store up to 512 MBs of data and is used to store all accelerometer samples.

The wearable takes 25 three dimensional accelerometer samples per second. Each sample provides the x, y, and z gravitational measure with 8 bits of accuracy between ± 4g. Every 200 milliseconds, 5 accelerometer samples are put into a BLE advertisement packet and transmitted.

The wearable constantly transmits 5 BLE advertisements each with 5 accelerometer samples until the battery is discharged. Under this workload, the wearable has been shown to last for up to three weeks without recharging.

3) Gateways: The four gateways act as the localisation anchor points for the system. Each gateway is a Raspberry Pi 3, Model B, which come with an integrated WiFi and Bluetooth Low Power radio. When a BLE advertisement packet is received, the contained accelerometer samples are logged to the daily log file stored on the gateway's SD card. The log for each packet includes the RSSI and a network time protocol based time stamp. The gateways report system monitoring information hourly and each night send the daily log file to the central database.

4) Operation: The users wear the wearable device on their wrist and only take it off when bathing (during this time the wearable can be charged). When in the home, it is expected that one or more of the gateways will be within range and receive the constantly transmitted BLE advertisements from the wearable. Accelerometer data produced while out of the home can be retrieved later from the flash memory of the wearable. Although when out of range no gateways can collect the packet and thus no RSSI information will be available. Nonetheless, this is acceptable for the problem of indoor localisation.

ACKNOWLEDGEMENT

This work was supported in part by EurValve (Personalised Decision Support for Heart Valve Disease), Project Number: H2020 PHC-30-2015, 689617 and the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

REFERENCES

[1] H. D. Mehr, H. Polat, and A. Cetin, "Resident activity recognition in smart homes by using artificial neural networks," in *2016 4th International Istanbul Smart Grid Congress and Fair (ICSG)*, apr 2016, pp. 1–5.

[2] F. Montori, L. Bedogni, and L. Bononi, "A Collaborative Internet of Things Architecture for Smart Cities and Environmental Monitoring," *IEEE Internet of Things Journal*, vol. PP, no. 99, p. 1, 2017.

[3] F. Lemic, J. Buesch, Z. Jiang, H. Zou, H. Jiang, C. Zhang, A. Ashok, C. Xu, P. Lazik, N. Rajagopal, A. Rowe, A. Ghose, N. Ahmed, Z. Xiao, H. Wen, T. E. Abrudan, A. Markham, T. Schmid, D. Lee, M. Klepal, C. Beder, M. Nikodem, S. Szymczak, P. Hoffmann, L. Selavo, D. Giustiniano, V. Lenders, M. Rea, A. Marcaletti, C. Laoudias, D. Zeinalipour-Yazti, Y.-K. Tsai, A. Bestmann, R. Reimann, L. Li, C. Zhao, S. Adler, S. Schmitt, V. Dentamaro, D. Colucci, P. Ambrosini, A. Ferraz, L. Martins, P. Bello, A. Alvino, V. Sark, G. Pirkl, P. Hevesi, X. Yang, R. R. Choudhury, V. Handziski, S. Sen, D. Lymberopoulos, and J. Liu, "A Realistic Evaluation and Comparison of Indoor Location Technologies: Experiences and Lessons Learned," in *The 14th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN '15)*. ACM Association for Computing Machinery, apr 2015.

[4] J. Pope, R. McConville, M. Kozlowski, X. Fafoutis, R. Santos-Rodriguez, R. Piechocki, and I. Craddock, *SPHERE in a Box: Practical and Scalable EurValve Activity Monitoring Smart Home Kit*. Institute of Electrical and Electronics Engineers (IEEE), 2017.

[5] A. T. Parameswaran, M. I. Husain, S. Upadhyaya *et al.*, "Is rssi a reliable parameter in sensor localization algorithms: An experimental study," in *Field Failure Data Analysis Workshop (F2DA09)*, vol. 5. IEEE, 2009.

[6] N. T. Nguyen, R. Zheng, and Z. Han, "UMLI: An unsupervised mobile locations extraction approach with incomplete data," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, apr 2013, pp. 2119–2124.

[7] L. Li, W. Yang, and G. Wang, "HIWL: An Unsupervised Learning Algorithm for Indoor Wireless Localization," *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 1747–1753, 2013.

[8] Y. Chen, M. Guo, J. Shen, and J. Cao, "GraphLoc: A Graph-based Method for Indoor Subarea Localization with Zero-configuration," *Personal Ubiquitous Comput.*, vol. 21, no. 3, pp. 489–505, jun 2017.

[9] M. K. Mwila, K. Djouani, and A. Kurien, "Approach to Sensor Node Calibration for Efficient Localisation in Wireless Sensor Networks in Realistic Scenarios," *Procedia Computer Science*, vol. 32, pp. 166–173, 2014.

[10] M. Ficco, C. Esposito, and A. Napolitano, "Calibrating Indoor Positioning Systems with Low Efforts," *IEEE Transactions on Mobile Computing*, vol. 13, no. 4, pp. 737–751, apr 2014.

[11] C. Wu, Z. Yang, Y. Liu, and W. Xi, "WILL: Wireless Indoor Localization Without Site Survey," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 4, pp. 839–848, apr 2013.

[12] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan, "Indoor Localization Without the Pain," in *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '10. New York, NY, USA: ACM, 2010, pp. 173–184.

[13] Z. Zheng, Y. Chen, S. Chen, L. Sun, and D. Chen, *International Journal of Distributed Sensor Networks*, no. 6.

[14] M. Shimosaka and O. Saisho, "Efficient Calibration for Rssi-based Indoor Localization by Bayesian Experimental Design on Multi-task Classification," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 244–249.

[15] S. He and S. H. G. Chan, "Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 466–490, 2016.

[16] A. Elsts, G. Oikonomou, X. Fafoutis, and R. Piechocki, *Internet of things for smart homes: lessons learned from the SPHERE case study*. Institute of Electrical and Electronics Engineers (IEEE), 2017.

[17] N. Twomey, T. Diethe, M. Kull, H. Song, M. Camplani, S. Hannuna, X. Fafoutis, N. Zhu, P. Woznowski, P. Flach, and I. Craddock, "The SPHERE Challenge: Activity Recognition with Multimodal Sensor Data," *arXiv*, 2016.

[18] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning From Crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, aug 2010.

[19] F. Rodrigues, F. Pereira, and B. Ribeiro, "Learning from multiple annotators: Distinguishing good from random labelers," *Pattern Recognition Letters*, vol. 34, no. 12, pp. 1428–1436, 2013.