RFID Indoor Localization Using Statistical Features

Abstract

In this paper, we present a method that uses the signal strength indication of RFID antennas with statistical features to perform relative positioning in a smart home. The goal of the proposed method is to enable the tracking of most objects inside a smart home in real-time, allowing activity recognition based on this tracking. This paper also introduce a new dataset of 4 100 000 RFID data collected in a real full-scale smart home setting. The dataset is available for the community. The method has an accuracy of 95.5% which is similar to previous work but require a fifth of the time to compute.

Keywords

Indoor localization, passive RFID, data mining, feature extraction

Introduction

World population ageing is a situation most government are now fully aware they must face in the following years (World Health Organization 2015). Related challenges are diverse, and so are the solutions proposed by researchers around the world. The challenges our teams are working on link to the increasing difficulty to provide adequate healthcare services to the ageing population. As life expectancy has increased in recent years, so have the number of persons suffering from age related diseases like dementia and Alzheimer's disease. This put healthcare systems under great pressure, both financially and in human resources.

Recent progress in miniaturization, especially for the Internet of Things (IoT), and the evolution of artificial intelligence algorithms have opened the way to the realization of the long-time dream of a real assistive

smart home. From a healthcare point of view, a smart home is a regular housing that offers assistance in the realization of activities of daily living (ADLs). The goal of this assistance is to enable ageing in place while preserving the quality of life of all inhabitants of the smart home, thus delaying hospitalization and reducing the pressure on healthcare facilities and informal caregivers (Al-Shaqi 2016). Indeed, the loss of autonomy of the elders often leads to long-term care in senior housing for non-autonomous or recurring hospitalization. The assistance can take many forms, from enhancing security of certain task like cooking to reducing the number of interventions needed from a relative living with an impaired person.

To replace effectively a long-term facility and to prevent accidents leading to hospitalization, a smart home must offers a large array of services and dispose of extensive data on its occupant (Cook 2013). However, collecting data on people is always a delicate subject as privacy concerns rapidly arises. To mitigate this problem, smart homes tend to avoid usage of video camera and instead rely on low level information collected using ubiquitous devices and technologies like passive infrared sensors, electromagnetic contacts, thermometers, wearable sensors and so on (Cook 2013) (Hsu 2017) (Bouchard 2012). The upcoming field of IoT is regularly adding more low-level sensors a smart home can use. All those sensors provide useful data used to monitor ADLs in real-time (Krishnan 2014). Smart homes are also safer if they work on a private network, preventing access to cloud computing and forcing efficient algorithms that can execute on regular computers.

Still, data from unobtrusive low-level sensors are often not enough to provide precise information about what is going on, at least in their raw form. While they might be enough to determine that someone is cooking, they generally cannot tell what meal is the person preparing nor at what step a person is in a recipe. Several methods have been proposed over the past decade to recognize ongoing ADLs, but this endeavour remains problematic due to the low granularity of the current solutions. The granularity, in activity recognition, refers to the level of abstraction provided by the method. For instance, from the lowest to the highest granularity, the same ongoing ADL could be defined as: *Cooking, Preparing pasta,*

Preparing shrimp fettucine Alfredo, or even as the atomic step *Putting fettucine in the boiling water*. While our teams at the LIARA laboratory and the DOMUS laboratory are fairly sensors agnostic (Belley 2014) (Pigot 2015), we believe that one of the solutions with the highest potential to solve this granularity problem is the passive Radio-Frequency Identification (RFID) technology. The main advantage of passive RFID is that several tags can be installed on daily usage objects in the smart home to enable their tracking in real-time. Therefore, such system could provide highly reliable spatial information to feed an activity recognition algorithm for better granularity. As passive RFID tags are small and cheap, they can be placed on most objects. Nevertheless, RFID tags are not suitable to place on perishables or and cannot survive microwave oven nor high temperature.

In this paper, a localization system based on techniques for machine learning/data mining is proposed. The method build upon the work of Bouchard (Bouchard 2017) and Bergeron et al. (Bergeron 2018) which is, in our knowledge, the only example of localization of several objects based on supervised data mining. Indeed, very few authors have worked on the problem of localizing daily usage object, and unfortunately, the best methods for humans/robots tracking often cannot be used straightforwardly because the technology used is too big (require batteries, antennas on the objects, etc.), is too costly, or requires several references points (disposing those in a smart home is not always feasible). As it will be argued further in the paper, daily objects localization is more challenging than human or robot tracking, and the accuracy and precision of the state-of-the-art is still arbitrary. To address this challenge, in this paper, the RFID Received Signal Strength Indication (RSSI) is viewed as a time-series, as they were in (Bouchard 2017). In the aforementioned paper, Bouchard identified the low sampling of the dataset from (Bergeron 2018) as a limit for a good evaluation of the proposed method. A new larger dataset is introduced to remove this limit. This paper pursue with the same research question Bouchard expressed: ”How useful at improving RFID localization methods would be the statistical features commonly used in machine learning?”. Then, we follow with a new research question: ”What statistical features are significant in

improving our RFID localization method". Accordingly, the contributions to the field of this paper, as an extension to (Bouchard 2017), are three-fold:

- A new RFID dataset
- An improved pipeline for passive RFID localization
- A feature selection to improve computing efficiency

The dataset used in this paper is a new dataset of 4 100 000 RFID readings generated from real data collected in full-scale kitchen infrastructures and is available to the scientific community at www.Kevin-Bouchard.com and www.usherbrooke.ca/domus. The previous dataset only contained 673 000 data dispersed in six rooms.

The remainder of this paper is as follow. We first begin with a quick review of indoor localization methods, then we present of experimental settings along with the new dataset. After, we present the three series of experiment we conducted with their conclusions.

Related work

Research on indoor localization has been active for more than two decades (Gutmann 1996). Over the years, many approaches where created and tested for many different intended usages. A review of some usages can be found in (Pahlavan 2015), along with some challenges localization still poses before the emergence of a smart world.

Methods using wireless technologies can be regrouped in three main categories. A first category concerns the proximity-based methods. In proximity-based methods, we use the known position of fixed objects or tags to infer the position of the target object. The fixed objects can be wireless antennas, like with NFC localization (Meng 2014), or tags, like in the LANDMARC system (Ni 2004). Considering the low range of NFC readers, the global idea is the say the position of the NFC tag is the same as the reader. For reference

system like LANDMARC, the strongest signal among the references determines the position. Statistical features can help improve proximity-based method, as in this work (Bouchard 2016) where the standard deviation of Bluetooth RSSI is used.

The second category are the lateration techniques that uses geometric properties to localize the target. Wireless antennas usually provide two types of information we can use for localization: the signal strength (the RSSI) and the angle of arrival (the AoA). Trilateration is the most used lateration technique using RSSI. The idea is to map the RSSI to a distance measure from the antennas and draw virtual ellipsoid to pinpoint the location at the intersection of few reference points (Fortin-Simard 2015). On the other hand, triangulation is the most popular method using the AoA. Instead of drawing virtual ellipsoid, triangulation demands to draw virtual straight lines and places the target at the intersection of at least two of them (Tekdas 2010).

The last family and the one of interest for this paper is the fingerprinting. The fingerprinting technique is usually used in conjunction with a better, more precise, localization system to build a radio map of the environment. The technique is, then, to use the learned map and compare, in real-time, the RSSI to associate the tracked entity to the closest location in a similar fashion than with landmarks. The main drawback is, however, the requirement for the high performance localization system (usually based on ultrasonic sensors) (Hightower 2001). The more precise system can be replaced by a manual collection of the fingerprints (Bergeron 2018). Nevertheless, manual collection is a long and tedious process. Still, it allows doing relative positioning and varying the precision at will. Fingerprints also allow using less antennas than lateration technique as the target only need to be in the range of a single antenna for the method to work.

Fingerprinting are also used with non-wireless technologies. Those includes sounds, magnetic fields and, to some extent, light. SurroundSense (Azizyan 2009), for instance, is a user-centred fingerprinting

localization application that uses the user's cellphone to capture the photo-acoustic signature of a place to later provide localization. SurroundSense also fingerprints the motion using accelerometers, the colour using the camera and the Wi-Fi, when available, to provide better localization. In their paper, the authors tried to localize in what shop the users were between 51 shops. LocateMe (Subbu 2013) is another user-centred localization method. It uses the magnetometer of a cellphone to record the ambient magnetic field (a combination of the magnetic field of the Earth and the local distortions provoked by metallic structures) in hallways. Then it uses dynamic time warping to match ongoing a user's location to a previously recorded fingerprint. By using dynamic time warping, LocateMe can adjust to various users with various walking speed or disabilities like blindness and paralysis requiring a wheelchair. However, those systems mainly use a cellphone that as to be carried by a user for the localization to occur. In their current form, the technologies they use cannot be placed on daily living objects. Moreover, LocateMe computes the position directly on the cellphone, in about 5 seconds. To our sense, this is too slow to use for real-time activity recognition, our final goal.

Methodology

In this section, the goal is to explain the methodology used to validate the research questions formulated in the introduction. While the emphasis of this paper is on the localization of one object in one smart home, the reader should keep in mind the bigger picture, which is about tracking several objects in real-time for ADLs recognition in smart homes to foster aging at home. Our team has already used the spatial data from passive RFID localization in activity recognition system in the past (Bouchard 2014) and improvements in the localization tend to translate directly in better activity recognition.

This paper is an update on the work of Bergeron et al. (Bergeron 2018) conducted with the LIARA and the DOMUS teams from which the author are members. The method exploited in the aforementioned paper relied, similarly to the literature, on using the raw RSSI signal from the passive tags to perform the

localization. This paper is also a follow up to the work of Bouchard (Bouchard 2017). In contrast, Bouchard, in his project, sees the RSSI as a time-series. Therefore, despite the low sampling of the dataset he used, the localization is performed over a data window, which is an aggregation of many consecutive readings.

The importance of this work relies on the premise that daily objects localization is more difficult than human/robot localization. The arguments are that daily objects can be very small (e.g.: a spoon, or a fork), numerous (in the kitchen there are several plates, containers, glasses, etc.) and that occlusion will often occur. Small objects implies that the tracking technology also have to be small and light. Numerous objects in turn implies that the localization process needs to be as efficient as possible in order to resolve in real-time and to scale graciously as more objects are added. Frequent occlusion disallows the use of line-of-sight technologies. It also requires the radio signal to operate at a frequency where interference with the environment are minimal. A smart home is hardly considered an open environment and thus the challenges of localization are different. The method we propose with this work tackles those three challenges by using passive RFID tags with an efficient localization method described below.

Smart home

In the introduction, we mentioned that a major contribution from this paper is a large dataset of RFID readings collected in a realistic smart home setting (Bergeron 2018). This subsection will delve into the smart home and the next will present in more details the new features of this dataset. The smart home is a full-scale apartment including a bedroom, a kitchen, a dining room, a living room and a bathroom. It is equipped with 20 polarized directional antennas distributed to cover the entire surface, as shown in Figure 1. These antennas are connected to five RFID readers and work on the 928Mhz band as specified by the Canadian Radio-Television and Telecommunications Commission (CRTC). Therefore, they have to be strategically installed to minimize collisions and maximize coverage. Collisions cannot occur among the antennas connected to the same reader since they work on a round robin. A derogation can often be

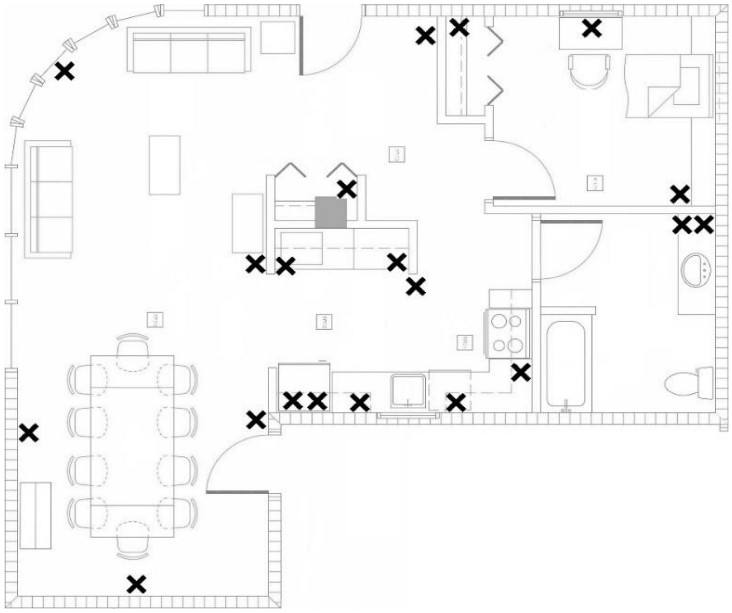

Figure 1 Map of the smart home and position of the RFID rantennas

obtained through the CRTC to change the band, but since our goal is to use the smart homes for aging in place, this would not be practical. In theory, the RFID system can collect the tags ID up to every 20ms. However, RFID is not a real-time system, thus the results are often very different. In practice, it has been observed to be reliably able to collect data under 100ms. They were set to collect at every 200ms for this experiment. The smart home features many more sensors, like pressure plates, infrared detectors and thermometer. For this paper, only data from RFID antennas are used. However, all other data sources were active during that time and might have caused some ambient noise. Still, this noise is expected to be present in a real-life smart home and is accordingly considered beneficial to this work.

Dataset

The smart home presented previously offers a surface of about 100 square meters. As explained in previous papers (Bergeron 2018) (Bouchard 2017), we consider the localization problem as a classification problem where positions are given in zones rather than in coordinates from an arbitrary origin. As for most classification problems, we first need to collect a dataset to train a classifier. The smart home being

quite large, we choose to only collect RFID readings in the kitchen for the current dataset, as opposed to the six rooms of the previous dataset. To better compare with the previous articles, we used the same 205 zones from the kitchen as before. We also used the same plastic bottle with the same four RFID tags facing different directions. What changed for this experiment was the number of collected readings and the state of the smart home. Indeed, this time we collected 1000 readings for each zone (it was only 50 per zone in the previous dataset). At 200ms between each reading, it took 3 minutes and 20 seconds per zone. During that time, we tried to mimic usual activities performed in any kitchen, like cleaning the dishes or cooking. Whereas in the first dataset readings where collected under perfect conditions with no human caused interference, this dataset contains normal interference we expect to find in an inhabited smart home. Therefore, 1000 readings per class for each antennas were recorded resulting in 205 000 vectors of twenty RSSI + one class or 4 100 000 data. The variation of the RSSI values is bounded between -38 to -69 decibel. Due to the high number of readings per zone, some vectors appear more than once in the dataset. In total, there are 184057 distinct vectors. This dataset will be exploited as is in the first experiment to present the new baseline for future comparisons.

Statistical Features

Bouchard showed in (Bouchard 2017) that statistical features extracted from the dataset collected in (Bergeron 2018) can improve the classification accuracy. Since this new dataset is essentially an expended version of the previous one, the hypothesis of the second experiment is that the features should also improve the classification accuracy on the new dataset.

The vectors contained in the dataset each represent the values at each antennas at a 200ms interval. They are regrouped to form a time-series. We refer to the size of the grouping as the data window. Then, on those time-series, it is possible to extract statistical features to take advantage of this group of readings. The features we used are presented in Table 1. The notation used considers **M** to be the matrix of the

data window made of **k** lines and **n** columns (the antennas). There are nine statistical features applied to each time-series and eight applied globally (to all 20 time-series). For instance, the Mean RSSI is the sum of all RSSI in a window for an antenna divided by the window size. The Global Mean RSSI is the sum of all RSSI in that window divided by the total number of elements in that window (**n*k**). Consequently, the size of each features vector is 189 (20 time-series * 9 statistical features + 8 global features + 1 class = 189).

Tableau 1 List of features applied on the RFID readings

Mean RSSI	Global Mean RSSI	Min RSSI
$\bar{x}_j = \frac{\sum_{i=1}^{k} x_{i,j}}{k}$	$GAvg = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{i,j}}{n * k}$	$\min(j) = \min_k\{x_k, j\}$
Variance of RSSI	Standard Dev of RSSI	Global Min RSSI
$Var(X_j) = \frac{1}{k}\sum_{i=1}^{k}(x_{i,j} - \bar{x}_j)^2$	$\sigma_j = \sqrt{Var(X_j)}$	$GMin = \min_n\{\min(j)\}$
Count Non-Zero	Global Mean Standard Dev	Max RSSI
$NZ_j = \sum_{i=1}^{k} 1 - \delta_{x_{i,j}}$	$GStDev = \frac{1}{n}\sum_{j=1}^{n}\sigma_j$	$\max(j) = \min_k\{x_{k,j}\}$
Global Non-Zero	Absolute Energy	Global Max RSSI
$GNZ = \sum_{j=1}^{n} NZ_j$	$E_j = \sum_{i=1}^{k} x_{i,j}^2$	$GMax = \max_n\{\max(j)\}$
Absolute Sum of Changes	Global Absolute Energy	Mean RSSI Change
$SC_j = \sum_{i=1}^{n}\lvert x_{i,j} - x_{i-1,j}\rvert$	$GE = \sum_{j=1}^{n} E_j$	$MC_j = \frac{1}{k}\sum_{i=1}^{k} x_{i,j} - x_{i-1,j}$
Global Sum of Changes	Global Total Power	
$GSC = \sum_{j=1}^{n} SC_j$	$TP = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{i,j}$	

Most of these features are common knowledge, but few of them may need a proper introduction. The Count Non-Zero, and by extend the Global NZ, count the number of occurrences where a signal was read, or to simply put where the RSSI was different from zero. It is expressed using the Kronecker delta. The Absolute Energy is the sum over the squared RSSI values. The Mean RSSI Change is the average fluctuation in RSSI that can be expected on the time-series. The Absolute Sum of Changes (and Global SC) is the sum over the absolute difference between each consecutive RSSI values. Finally, the Global Total Power is the sum of all RSSI values over each time-series of the window.

There are many more features in the literature that we could have used. Features based on Fourier transform, for instance, are popular to work on time series. However, they require expensive mathematical computation that might not scale well for a real time usage on many objects. In fact, this real-time constraint prevent us to use more complex features than those presented above, even thought some of them might improve the accuracy of our classifiers.

Experiments and Results

Three experiments are presented in this section. The first reprises the classification work from (Bergeron 2018) using the new dataset. The second reprises the work from (Bouchard 2017) also using this new dataset. The last one goes further on the statistical features by examining which are the most used. The third experiment, along with the new dataset, is considered the main contribution of this article.

Raw readings classification

The goal of the first of the three experiments is to establish a new baseline for classification on the new dataset. To do so, we used the same protocol we used before. We use classification algorithms from the popular Waikato Environment for Knowledge Analysis (Witten 2016), Weka, on the dataset without any pre-treatment using only default parameter. For this part, the goal is not to find what parameters

combination gives the highest accuracy. Instead, the goal is to provide a rough preview of what can be achieved by common classifier on this dataset for the task of indoor localization. The results presented in this section were all obtained using default parameters and 10-fold cross validation as provided in Weka. Table 2 presents the accuracy for a selection of algorithms and compare them to the results presented in (Bergeron 2016) for the kitchen.

Tableau 2 Accuracy of some classification algorithms on the raw dataset.

Algorithm	Previous Dataset	New Dataset	Diff +/-
CART	73.403%	80.601%	7.198%
J48 (C4.5)	74.966%	80.269%	5.303%
Random tree	67.815%	75.696%	7.881%
Random forest	88.916%	86.590%	-2.236%
Naïve Bayes	83.227%	46.054%	-37.173%
1-NN	78.824%	82.299%	3.475%

Results show an increase of the accuracy for most algorithms. The random forest performs a little worst, but is still the best algorithm among those we tried. The weighted F-Measure for all algorithms is the same as the accuracy. For J48, it is 0.8077 and for CART it is 0.8060. In all cases, there is still a need for improvement, as being wrong 15% of the time is unpractical in order to build a reliable ADLs recognition system using this method. Given there are twenty times more training examples in the new dataset, an increase in accuracy was expected. With more example, we are closer to have all possible values seen while training. This also means that the goal is not to have a classifier good at generalizing but a classifier that can remember well. Trees are good in this context and it reflects in the accuracy.

Statistical features classification

The first experiment showed that using a larger dataset allows an increase in the accuracy. However, this amelioration is not enough to build a robust ADLs recognition system on top of it. As Bouchard showed in (Bouchard 2017), statistical features extracted from a time series formed by the RFID vectors contained

in the dataset can improve the accuracy of the indoor localization while retaining real-time capacities. The second experiment of this paper simply pick up the work of Bouchard and applies it to the new larger and noisier dataset.

Tableau 3 Accuracy of some classification algorithms on the feature dataset with a window of 5.

Algorithm	Raw dataset	Feature dataset	Diff to raw +/-	Diff to Bouchard +/-
CART	80.601%	94.437%	13.836%	0.437%
J48 (C4.5)	80.269%	95.378%	15.109%	-1.022%
Random tree	75.696%	83.649%	7.953%	-5.951%
Random forest	86.590%	98.232%	11.642%	-1.568%
Naïve Bayes	46.054%	70.670%	70.670%	24.616%
1-NN	82.299%	83.995%	1.696%	15.495%

Table 3 presents results with a data window of five. The accuracy of the various algorithms are similar to those obtained by Bouchard. Considering that the new dataset contains noise, a small decrease in the accuracy was to be expected. The difference in accuracy between the feature dataset and the raw dataset is less marked than in (Bouchard 2017), but since the baseline accuracy is higher than before, this too was to be expected. The weighted F-Measure of those algorithms reflects the accuracy. For instances, the weighted F-Measure of J48 is 0.9537 and is 0.8404 for the nearest-neighbour algorithm (1-NN).

The impact of Windowing

In his paper, Bouchard also explored the impact of windowing. His conclusion were that accuracy improves with the window size with a saturation around 99.53%. We also reproduced this experiment with a J48 classifier and the results are given in Figure 2. With a window size of one, the accuracy is 80,714%, which is slightly higher than the accuracy on the raw dataset. The accuracy never ceases to improve within the selected boundaries. At size 25, the accuracy is 99,813%, meaning that only 375 examples are misclassified. If accuracy was the only criterion, this window size would be a good choice. However, a window size that big is not practical for real-time positioning, as most moving objects would cross multiple

zones during the five seconds needed to record 25 readings. This is unpractical for online real-time tracking. A window of five readings, at 95,370% accuracy, seems a good trade-of between accuracy and practicability when the delay between two consecutive RFID readings is 200ms. The choice of the window should reflect both the RFID sampling rate and the speed at which objects are expected to move. If object are expected to be fixed, like in a warehouse, a large window will offer a better accuracy. For activity recognition based on object movement, a smaller window allows a more precise tracking, especially while a tracked object is being used. With the size of our zones (20 cm), we expect objects to be moving at a speed between two to five zones per second.

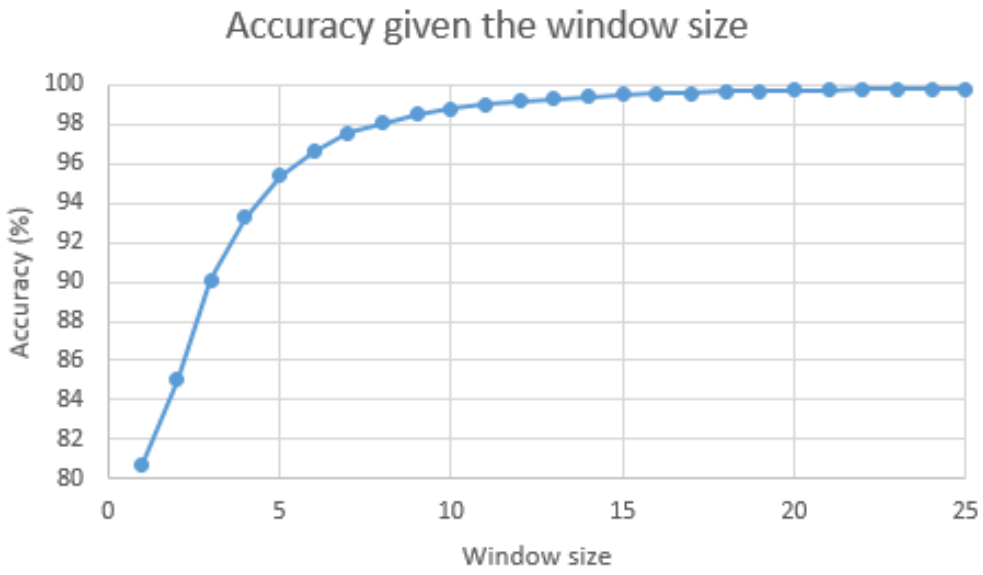

Figure 2 Accuracy of the J48 classifier for a window size varying from 1 to 25 RFID vectors

Apart from the accuracy, another impact of the windowing is the size of the produced tree. A window of 5 results in a tree of 10563 nodes, while a window of 10 produces 4667 nodes and a window of 25 results in a J48 tree of 1965 nodes (of which 848 are leaves). This suggests that most zones have a distinct signature that is easier to detect when more readings are aggregated. This also suggests that some features might not have a significant impact in the classification has they have less to gain from more data. The max and the min RSSI are the features that have to most to gain from a bigger window as more data

increase the chances of finding new extreme values. If there are no extreme values and instead they are similar for each vector, measures of central tendency should not be much impacted by more data. The next experiment aims to find what features are the most significant during the classification.

Selection of features

The previous experiments showed that the feature approach helps to improve the accuracy of classification algorithms on a noisy dataset of RFID readings collected from a real-life smart home. The goal of the indoor localization system presented here is to build an ADLs recognition system based on the tracking of objects. This means that for a real-life settings, the system will need to track near hundred of objects in real-time. At this scale, every computation becomes significant. With that thought in mind, we used several methods to try to reduce the number of features. There a 9 statistical features and 8 global features applied on each sliding window.

We first tried the principal components analysis (PCA) method as a wrapper over a J48 tree. The result was a tree of more than 40 000 nodes. PCA formed 25 new features using most of the original ones. In fact, it used at least once every statistical feature, meaning we could not remove any of them. We then used the information gain wrapper retaining only the features that offered a gain ratio superior to zero. There were 80 features that qualified this criterion, once again including all the original features on at least one antenna. The accuracy is also similar to the original feature dataset, with 95.504% on a tree of 10559 nodes. When then tried to limit the number of original attribute used and the number of produced features, only to obtain similar results.

Since the classical feature selection algorithms were of limited use in our context, we had to think of a less mathematical method. We designed the following experiment to find what features are the most used and what is the impact of removing a feature, both in time and in accuracy. First, we take all features and train a J48 tree on it. Then, we list the frequency of all features used in the tree. Finally, we remove

the least used feature and go back to training a new tree. We repeat this process until there is only one feature left in the dataset. For the statistical features, we count them for all antennas. Therefore, if the Min RSSI is used 10 times for antenna 4 and 8 times for antenna 6, the count is 18 for the feature. To us, it made no sense to remove a feature only for a given antenna, as it would not bring any knowledge transferable to an other room or another smart home setting.

Tableau 4 Accuracy of J48 when removing features with a window of 5.

Feature removed	Computation time	Accuracy	Number of nodes	Features used
None	161s	95.445	10563	67 of 188
Global Mean RSSI	157s	95.433	10573	67 of 187
Global Abs Energy	160s	95.437	10589	66 of 186
Standard Deviation	140s	95.457	10577	66 of 166
Global SC	133s	95.480	10553	65 of 165
Global Total Power	135s	95.500	10557	65 of 164
Global Mean StDev	133s	95.490	10581	64 of 163
Global Min RSSI	121s	95.474	10555	63 of 162
Global Count RSSI	121s	95.504	10549	61 of 161
Global Max RSSI	117s	95.135	11029	60 of 160
Variance	96s	95.188	11113	54 of 140
Mean Change	97s	95.322	11023	46 of 120
Abs Sum of Changes	75s	95.471	10919	38 of 100
Abs Energy	57s	95.532	10935	31 of 80
Mean RSSI	46s	95.772	11349	24 of 60
Count Non-Zero	41s	96.054	10987	16 of 40
Min RSSI	35s	95.104	10761	8 of 20

In Table 4 we listed the accuracy resulting at each step of our removal method. We also list the number of nodes in the resulting tree to give an idea of the complexity and execution time of the resulting tree. The computation time is the time it took to compute the features on the dataset, as this was done in an off-line phase. Those times should be considered carefully, as they are only an indication of the computation complexity of computing the features. This task was done on one of our work computer, while other programs where also running. Therefore, they are only given so we can appreciate the

difference between the features. The last columns indicate how many features where used in the classification tree versus how many where present at this stage of the experiment. Given that the RFID antennas are placed to cover the whole smart home and not only the kitchen, some of them never recorded any reading. Therefore, it is normal to see that they were not used in the tree as we can see in the last column.

As we can see in the table, removing features does not have a big effect on the accuracy. In fact, the accuracy varies around 95.5% for most classifiers. The real difference between each row are the computation time and the number of nodes. In order the have the fastest system possible, both need to be minimized. However, the number of nodes generally grows as we remove features. Still, computing only the Max RSSI takes about a fifth of the time needed for all the feature while producing a tree a fifth bigger. Nevertheless, since trees have a search complexity of $\log_2 n$, the increase in the size of the tree is not enough to slow the classification (it adds about 0.2 node in average). The conclusion of the experiment whose results are in Table 4 is that the most useful feature is the Max RSSI, when using a window of 5. With only this feature, we can built a classifier (J48) that is about as fast and as accurate as those presented in the second experiment. This feature is also the most significant when we use a window of 25, with an accuracy of 99.757% (although a window of 25 represents 5 seconds of data collection and can only be applied to fixed objects).

Individual features

The iterative process presented above allows isolating to most significant feature by removing a feature at the time. It shows that the max RSSI alone offers a good accuracy. Still, we decided to evaluate the other features to see if alone they can perform well. Table 5 presents the results. From this table, it appears that the mean, the min and the max RSSI are the three features that contribute the most to the

accuracy. This was to be expected since the min RSSI is the last feature removed in the previous experiment and the max RSSI is the remaining one.

Tableau 5 Individual features with a window of 5.

Feature	Accuracy	F-Measure
Count Non zero	20.490%	N/A
Absolute Energy	84.877%	0.8494
Abs Sum of Change	19.358%	0.1906
Mean Change	12.861%	0.1255
Mean RSSI	**88.617%**	**0.8860**
Minimum	**90.671%**	**0.9068**
Maximum	**95.104%**	**0.9568**

Another J48 tree was trained on a dataset containing the four best features identified in Table 5 (absolute energy, mean RSSi, maximum RSSI and minimum RSSI). It outperforms most classifiers from Table 4 with an accuracy of 95.853% (8532 incorrectly classified instances). The weighted F-Measure is 0.958 and the tree has 10495 nodes in it. For a generation time of 78 seconds, it can be seen as maybe the best trade-off between accuracy, features computation time and execution time. It is worth noting that those features are the ones with highest information gain ratio.

Conclusion

In this paper, we proposed in indoor positioning system based on passive RFID tags. We model the indoor positioning problem as a classification problem, were we aim to find in what zone a tag is instead of at what coordinates. We considered RFID readings as a time series on which we computed several statistical features and built classifiers on those features. Then, we examined what features are the most useful in our context. Our results shows that using statistical features can help improve the localization accuracy by more than 15% when compared to using the raw RSSI readings. The time series are formed by aggregating several consecutive readings by using a sliding window. The size of the window has an impact

on the accuracy. With a window bigger of 25, the accuracy is more than 99.8%, but it is only usable with fixed objects as it takes 5 second to collect these readings. Our last experiment showed that simple statistical features are as useful as more complex ones. In fact, using only the max RSSI value of the window offers an accuracy comparable to using all the 17 features we first tried. Still, the absolute energy, the mean RSSI and the minimal RSSI values for each antennas also offer an increase in the accuracy compared with the raw dataset. Together, those four features creates one of the best J48 tree of the whole experiment with an accuracy of 95.853%. Those are the features we intend to keep in our upcoming real-time activity recognition system.

All those experiments were made possible by the gathering of a new large dataset of RFID readings. The readings where collected in a full-scale smart home while daily living activities where performed and while other radio-frequency technologies (Wi-Fi, ZigBee, Z-Wave) where in use. This makes this dataset suitable for real-life usage as it already includes many natural form of noise we expect to find in a smart home. The dataset of 4 100 000 data is freely available for all researchers to use. Free data in that quantity is hard to find as smart home settings are still uncommon and costly to build.

The indoor localization system presented offers an accuracy of about 95% with a window of 5 readings. However, our method requires a tedious work of fingerprinting the smart home that must be done for each new home. Future work could focus on automating this process or on analyzing what precision is required for different categories of task to reduce the number of needed zones. A direct follow up of this work would be to use the system as a building block for a tracking system or for an indoor activity recognition system based on the position of objects. The localization system can also provide positioning for a large range of IoT applications. As the proposed methods of statistical features only uses RSSI as an input, it should be replicable with other radio-frequency technologies. Future work might also wish to explore if those other technologies offer the same accuracy.

Acknoledgements

This work was made possible by the financial support of the Natural Sciences and Engineering Research Council of Canada.

References

Al-Shaqi, Riyad and Mourshed, Monjur and Rezgui, Yacine. «Progress in ambient assisted systems for independent living by the elderly.» Édité par Springer. *SpringerPlus* 5, n° 1 (2016): 624.

Azizyan, Martin and Constandache, Ionut and Roy Choudhury, Romit. «SurroundSense: mobile phone localization via ambience fingerprinting.» *Proceedings of the 15th annual international conference on Mobile computing and networking.* ACM, 2009. 261-272.

Belley, Corinne and Gaboury, Sebastien and Bouchard, Bruno and Bouzouane, Abdenour. «An efficient and inexpensive method for activity recognition within a smart home based on load signatures of appliances.» Édité par Elsevier. *Pervasive and Mobile Computing* 12 (2014): 58-78.

Bergeron, Frédéric and Bouchard, Kevin and Gaboury, Sébastien and Giroux, Sylvain and Bouchard, Bruno. «Tracking objects within a smart home.» Édité par Elsevier. *Expert Systems with Applications*, 2018.

Bouchard, Kevin and Bouchard, Bruno and Bouzouane, Abdenour. «Guidelines to efficient smart home design for rapid AI prototyping: a case study.» *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments.* ACM, 2012. 29.

Bouchard, Kevin and Bouchard, Bruno and Bouzouane, Abdenour. «Spatial recognition of activities for cognitive assistance: realistic scenarios using clinical data from Alzheimer's patients.» Édité par Springer. *Journal of Ambient Intelligence and Humanized Computing* 5, n° 5 (2014): 759-774.

Bouchard, Kevin and Ramezani, Ramin and Naeim, Arash. «Features based proximity localization with Bluetooth emitters.» *Ubiquitous Computing, Electronics \& Mobile Communication Conference (UEMCON), IEEE Annual.* IEEE, 2016. 1-5.

Bouchard, Kevin. «Statistical Features for Objects Localization with Passive RFID in Smart Homes.» *International Conference on Smart Objects and Technologies for Social Good.* Springer, 2017. 21-30.

Cook, Diane J and Crandall, Aaron S and Thomas, Brian L and Krishnan, Narayanan C. «CASAS: A smart home in a box.» Édité par IEEE. *Computer* 46, n° 7 (2013): 62-69.

Fortin-Simard, Dany and Bilodeau, Jean-Sébastien and Bouchard, Kevin and Gaboury, Sebastien and Bouchard, Bruno and Bouzouane, Abdenour. «Exploiting passive RFID technology for activity recognition in smart homes.» Édité par IEEE. *IEEE Intelligent Systems* 30, n° 4 (2015): 7-15.

Gutmann, J-S and Schlegel, Christian. «Amos: Comparison of scan matching approaches for self-localization in indoor environments.» *Advanced Mobile Robot, 1996., Proceedings of the First Euromicro Workshop on.* IEEE, 1996. 61-67.

Hightower, Jeffrey and Borriello, Gaetano. «Location systems for ubiquitous computing.» Édité par IEEE. *Computer* 34, n° 8 (2001): 57-66.

Hsu, Yu-Liang and Chou, Po-Huan and Chang, Hsing-Cheng and Lin, Shyan-Lung and Yang, Shih-Chin and Su, Heng-Yi and Chang, Chih-Chien and Cheng, Yuan-Sheng and Kuo, Yu-Chen. «Design and

Implementation of a Smart Home System Using Multisensor Data Fusion Technology.» Édité par Multidisciplinary Digital Publishing Institute. *Sensors* 17, n° 7 (2017): 1631.

Krishnan, Narayanan C and Cook, Diane J. «Activity recognition on streaming sensor data.» Édité par Elsevier. *Pervasive and mobile computing* 10 (2014): 138-154.

Meng, Philipp and Fehre, Karsten and Rappelsberger, Andrea and Adlassnig, Klaus-Peter. «Framework for Near-Field-Communication-Based Geo-Localization and Personalization for Android-Based Smartphones—Application in Hospital Environments.» *Stud Health Technol Inform* 198 (2014): 9-16.

Ni, Lionel M and Liu, Yunhao and Lau, Yiu Cho and Patil, Abhishek P. «LANDMARC: indoor location sensing using active RFID.» Édité par Springer. *Wireless networks* 10, n° 6 (2004): 701-710.

Pahlavan, Kaveh and Krishnamurthy, Prashant and Geng, Yishuang. «Localization challenges for the emergence of the smart world.» Édité par IEEE. *IEEE Access* 3 (2015): 3058-3067.

Pigot, Hélène and Giroux, Sylvain. «Living labs for designing assistive technologies.» *E-health Networking, Application \& Services (HealthCom), 2015 17th International Conference on.* IEEE, 2015. 170-176.

Subbu, Kalyan Pathapati and Gozick, Brandon and Dantu, Ram. «LocateMe: Magnetic-fields-based indoor localization using smartphones.» Édité par ACM. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, n° 4 (2013): 73.

Tekdas, Onur and Isler, Volkan. «Sensor placement for triangulation-based localization.» Édité par IEEE. *IEEE transactions on Automation Science and Engineering* 7, n° 3 (2010): 681-685.

World Health Organization. *World report on ageing and health.* World Health Organization, 2015.