Symbolic Object Localization Through Active Sampling of Acceleration and Sound Signatures

(Nominated for the Best Paper Award)

Kai Kunze and Paul Lukowicz

Embedded Systems Lab, University of Passau,
Insstr 43, 94032 Passau, Germany
{kai.kunze|paul.lukowicz}@uni-passau.de
www.wearable-computing.org

Abstract. We describe a novel method for symbolic location discovery of simple objects. The method requires no infrastructure and relies on simple sensors routinely used in sensor nodes and smart objects (acceleration, sound). It uses vibration and short, narrow frequency 'beeps' to sample the response of the environment to mechanical stimuli. The method works for specific locations such as 'on the couch', 'in the desk drawer' as well as for location classes such as 'closed wood compartment' or 'open iron surface'. In the latter case, it is capable of generalizing the classification to locations the object has not seen during training. We present the results of an experimental study with a total of over 1200 measurements from 35 specific locations (taken from 3 different rooms) and 12 abstract location classes. It includes such similar locations as the inner and outer pocket of a jacket and a table and shelf made of the same wood. Nonetheless on locations from a single room (16 in the largest one) we achieve a recognition rate of up to 96 %. It goes down to 81 % if all 35 locations are taken together, however the correct location is in the 3 top picks of the system 94 % of the times.

1 Introduction

The location of an object can be interesting for a variety of reasons. Most obvious is the 'where did I put my x' scenario. An example where this scenario is relevant are so called assisted living systems. Such systems use on body devices for behavioral monitoring and assistance for elderly and/or cognitively impaired persons. In such a scenario, an important concern is to make sure that the user has the device with him all the time. This implies checking if the user carries the device and, if not, using for example the TV, the radio or the phone to remind him to pick it up. In particular for cognitively impaired users, it is important to be also able to tell the user where the device is located, in case it was lost.

Another well known example is a mobile phone that knows whether it is in a pocket, on the table, or in the user's hand and adjust the volume accordingly. Generally, we can use the location of 'smart objects' as an indication of the user needs and intentions. Thus if a device is put in the drawer where it is usually stored, it is reasonable to assume that it will not be used in the near future and

J. Krumm et al. (Eds.): UbiComp 2007, LNCS 4717, pp. 163–180, 2007.

it can go into power saving mode. Going even further the location of a set of objects can be an indication of more general user activity and intentions.

Clearly understanding how object location can be used in different applications is a complex topic that needs further research. Nonetheless, the type of considerations sketched above indicates that object location is a useful piece of information. From this motivation we present and systematically evaluate a novel method for object localization. The method provides so called symbolic (sometimes also called semantic) location (e.g. [1]) rather then absolute coordinates. Thus the output of the system is of the type 'on the couch' or 'in the drawer'. The key contribution of our work is to present a method that requires no infrastructure, relies on simple, cheap sensors and still produces useful results.

The method is derived from the observation that the a ringing mobile phone sounds differently depending on where it is located. Whereas a phone in a jacket pocket sounds 'dumped', a phone on a metal cabinet can make the entire cabinet resonate. This is true for a ringing as well as for merely vibrating phone. We thus propose to use sound from a built in speaker and vibration from a built vibromotor to create a mechanical 'excitation' of the environment and analyze the response with an accelerometer and a microphone. In an extensive experimental study (47 locations with total of 1200 data points) we demonstrate that two types of information can be derived from this analysis. First, the system can be trained to recognize specific locations such as the 'kitchen table', or the 'dining room table'. Second, it can recognize more abstract locations based on materials such as a 'wood table', 'a closed metal cabinet', or a 'jacket pocket'. While this leads to less specific positioning, it has the advantage that the system does not need to be trained for each single location. Instead, after being trained on, for example, several wood tables, it will recognize others it has not seen before.

1.1 State of the Art and Related Work

Indoor location is known to be a hard problem (see [2] for an overview). As described above our work aims at the localization of simple objects in environments with no, or only minimal augmentation. This means that many of the more reliable, standard methods are not applicable. This includes ultrasonic location such as the BAT [3] or the MIT cricket systems [4] which both require extensive instrumentation of the environment with ultrasonic transceivers. In addition ultrasonic system require free line of sight and will fail to locate objects in closed compartments. This means that infrastructure free, relative positioning methods based on ultrasonics (see ([5]) are also unsuitable. Cost and effort also make the use of complex time of flight based radio frequency (RF) methods such as the commercial UBISENSE ultra wide band system (www.ubisense.net) infeasible. Similar can be said about RFID (radio frequency identification), which require a reader to be put on every location which needs to be recognized.

Simple Beacon Based Systems. Much work has been put into localization based on simple RF beacons, often based on standard communication systems such as Bluetooth, Zigbee and of course WLAN ([6], [7] and [8]). This includes a wide

body of work on positioning in wireless sensor networks [9]. In particular work based on, low power radio systems is clearly relevant to object localization. However it must be seen as complementary rather then a competing approach. Such systems are virtually all based on signal strength, which is inherently unreliable in complex, indoor environments. As a consequence, they are predominantly used for room level location (determining which room or large room segment a sensor node is in). This is not sufficient for the type of symbolic location targeted by this paper. However knowing approximate physical location can be used to constrain the search space for our symbolic location method.

Indirect Localization with Sensor Signatures. Both sound and acceleration have been previously used in location related research. In [10], the authors present a technique for performing accurate 3D location sensing using off-the-shelf audio hardware. Van Kleek et al. has also done some work in this direction, using sound fingerprints to detect collocation in [11].

The general concept of using acceleration signatures to extract location related information can be traced to the 'Smart-Its Friends' paper, [12]. Building on this idea [13] have demonstrated how to determine if a set of devices is being carried by the same person by correlating their acceleration signatures. Kunze et al. has taken this concept even further to show how the acceleration signature of walking can be used to determine where a user is carrying a device [14].

The most direct relation to the work presented in this paper is a patent by Griffin [15] titled: User hand detection for wireless devices. It proposes to use vibration detected by an acceleration signal to determine if a mobile phone is in the user hand, in a holster or in a holder.

1.2 Paper Contributions and Organization

From the above discussion it can be seen that symbolic localization of objects with no external infrastructure and simple sensors suitable for small, cheap nodes is an open problem. This paper proposes a solution for this problem. In terms of hardware the solution requires only a microphone, an accelerometer, a small speaker capable of emitting 'beeps' and a miniature vibration motor. An important feature of our method is the fact that it can be used on both specific locations (e.g. my 'kitchen table'), and abstract location types.

We discuss the physical principle, key issues, and limitations behind our approach (section 2). We then provide a detailed description of the recognition algorithm, including, feature computation,classification, and classifier fusion (section 3). Finally, we validate our method on an extensive, realistic data set (section 4). The data set contains a total of over 1200 measurements from 35 specific locations (taken from 3 different rooms) and 12 abstract location classes. The location were chosen to include examples that demonstrate the limits of the method such as an attempt to distinguish between the inner and the out pocket of the same jacket and between table and a book shelve both made of identical material. The data points at each symbolic location area taken at a number of randomized spots to ensure representativity.

Despite such challenging evaluation our method produces promising results. On room bases (16, 9 and 10 locations) we arrive at an accuracy of between 89 % and 93 % with the correct answer being in the to 2 first picks of the classifier between 97 % and 99 % of the time. With all 35 locations from the 3 rooms in one data set the recognition goes down to 81 %. However we still get the correct answer in the top 2 picks of the classifier 91 % and in the top 3 94 % times.

2 Approach Overview

2.1 The Method

Procedure Description. The proposed method consists of two parts, each of which can be used alone or in combination with the other.

The first part is based on vibrating the device using a vibration-motor of the type commonly found in mobile phones. During the vibration, which last a couple of seconds, motion data is recorded with an accelerometer and sound with a microphone. The motion and sound signals are used separately for an initial location classification using standard feature extraction and pattern recognition methods. The final classification is obtained through appropriate fusion of the two classification results.

The second part is based on sound sampling. The device emits a series of beeps, each in a different, narrow frequency spectrum. The microphone is positioned is such a way that it receives only little energy directly from the speaker. Instead a significant part of the energy comes from reflections from the immediate environment (see section 2.2 for a more detailed discussion). For location recognition the sound received from the different beeps is compared.

When the two parts are used together, the corresponding results are fused using an appropriate classifier fusion method.

General Principles Behind the Recognition. In abstract terms the above method is about analyzing the response of the environment to a mechanical 'excitation' with different frequencies. By vibrating the device we provide a low frequency (a few Hz) relatively high intensity (as compared to sound) source of excitation. By emitting fixed frequency 'beeps' we generate different, low intensity high frequency stimuli. The accelerometer detects the low frequency response (in our case up to 15Hz due to sampling frequency of the used device limited at 30Hz), the microphone the high frequency part.

The response to the above stimuli falls into several categories. First we get a low frequency response that directly mechanically couples to the vibrating object and is detected by the accelerometer. This response can range from a more or less complete absorption of the vibration energy (e.g. when the object is lying on pillow) to a resonant response where the surface, on which or device is lying, joins in the vibration. This fact contains information on two things. For one, it can reveal if, and how the device is fixed (in the hand, in a tight pocket, lying freely). In addition it reveals how hard and elastic is the surface on which the device is placed. This information can be expected to reliably distinguish between soft

surfaces such as a sofa and hard ones like a table. Distinction between several similarly hard surfaces (e.g. metal and stone) is difficult.

Second, we get a high frequency response to the vibration, which is essentially a sound from the device hitting the surface. Assuming that placement of the device does lead to this kind of response (it will not, if the device is in a soft pocket or say hanging), it is quite location specific. The sound depends not only on the surface material but also on the overall structure. Thus a small, solid cube will produce a different sound then a large thin surface, even if both are made of the same material. Finally, objects light and close enough to the device to be influenced by the vibration (e.g. a key chain) might also contribute to the sound. In general, this is a source of noise rather then usable information. Figures 1 show two different vibration spectra.

Third, we get a high frequency response from the beeps which is given by the absorption spectrum of the environment. [1] Clearly this response is only useful if it comes from the immediate vicinity of the device. This can either be the surfaces on which the device is lying or, if the semantic location is a closed compartment, the walls of this compartment (see next section for a discussion of microphone placement issues). It is well known that the acoustic absorption spectrum is a distinct material property. The topic has been extensively studied in the context of musical instruments and sound isolation in construction ([16]). Typically the absorption is given at discrete frequencies as a fraction of the perfect absorption at an open window (lack of any reflecting surface) of equal area. As an example we consider the following coefficients from [16]

frequency	128 Hz	256 Hz	512 Hz	1,024 Hz	2,048 Hz	4,096 Hz
concrete unpainted	0.010	0.012	0.016	0.019	0.023	0.035
brick wall painted	0.012	0.013	0.017	0.020	0.023	0.025
carpet on concrete (0.4inch)	0.09	0.08	0.21	0.26	0.27	0.37

The above clearly demonstrates that, in principle, even seemingly similar materials can be separated with a small number of discrete frequencies.

Applying the Method: Specific Locations vs. Location Classes. The above description shows that our method provides information on abstract properties such as surfaces material as well as information on properties characteristic of a single specific location (e.g. a solid cube vs. large surface with several legs). As a consequence this paper investigates two different usage modes of our method:

1. 'Specific Location Mode'. In this mode we train the system on concrete locations such as a specific table or a specific chair. The advantage of this approach is that the user is provided with exact location information. The main disadvantage is the effort involved in training each individual location. In addition, there is the question being able to distinguish a large enough number of locations to satisfy relevant applications.
2. 'Abstract Location Class'. In this mode we divide locations into abstract classes. The two main criteria are the surface material and being open (e.g.

[1] Note that the absorption also influences the sound caused by the device vibration.

tabletop) or closed (e.g. inside a cupboard). In this mode the system is trained on several instances of each class. It is then able to recognize arbitrary other instances of this class. Thus the training problem is avoided, as the system can pre-trained at 'production time' and given to users without the need for further training. The disadvantage lies in the less exact location information, which has to be further interpreted and/or combined with additional information to find out where the object is actually located.

2.2 Issues to Consider

Microphone and Speaker Placement. As described above for the analysis of the absorption spectrum we must ensure that the sound emitted by the loudspeaker is reflected from the surface on which the device is lying and/or, in case of the symbolic location being a closed enclosure, from the enclosure walls. The second part is trivial. The first implies an appropriate placement of the microphone and the speaker. Optimally the speaker and the microphone should be located close to each other on the side of the device, preferably (but not necessarily) facing downwards with a sound proof barrier blocking the direct sound path between them. The main problem in implementing this type of setup is the definition of 'on the side' and 'downwards'. In the worst case we could be dealing with a cubic or round object with no preferred 'down' or 'side'. For such object two loudspeakers located at a 90 degree angle would have to be used to ensure that there is always a sidewards facing one. Our experiments (see section 4) indicate that the position of the microphone is less critically and we achieved good results despite the microphone facing upwards, so that one microphone might suffice.

Variations within Symbolic Locations. Many symbolic locations such as 'table' or 'desk' have considerable physical dimensions. This means that the response to the mechanical stimuli may be subject to spatial variations. Thus for example the low frequency response to vibration (acceleration data) may be different over the leg then in the middle of a large table. Similarly, on a table adjacent to the wall, the response to the 'beeps' will vary depending on how close to the wall the device has been placed. As a consequence both for training and testing a sufficient number of random physical locations must be sampled for each symbolic location (as has been done in experiments described in section 4).

Number of Relevant Locations. Clearly there are limits to how many locations can be reliably recognized. At the same time, in every day environments such as home or office, there are many places where objects can be put. The question is, whether the number of locations that can be distinguished is sufficient to be useful in relevant applications. An authoritative answer to this question can only be found through an analysis of specific applications. As stated in the introduction this is a technology, not an application paper and we make no

claim to such an answer. Instead, exploring the technology side, we demonstrate and argue the following:

1. Our system shows reasonable recognition performance even using the combined data set of 35 locations. In our experiments these are collected from 3 rooms. It seems unlikely that this would not be sufficient to cover all relevant symbolic locations in a single room. At the same time , as has been discussed in the introduction, room level location of RF enabled sensor nodes is a manageable problem.
2. Provided that a adequate number of sufficiently abstract classes is chosen, the number of locations issue is avoided by the 'abstract location classes' usage mode. In the experiments we demonstrate near perfect recognition for 7 and reasonable results for 12 classes. The type of classes used in the experiments '(open wood surface', 'closed wood cabinet' etc.) is clearly abstract enough to describe a large number of locations.

Sensor Requirements. In the introduction we have stated our aim of developing a method suitable for smart objects. Accelerometers and a microphones are among the most widely used components in small sensor nodes. Small loudspeakers capable of emitting beeps are also commonly integrated in sensor nodes. As will be described in section 3 we work with frequencies between 500 and 4000 Hz, which can be handled by small, cheap speakers and microphones. Finally, although vibration motors have so far not been used in sensor nodes, they are available in sizes around 1cm and smaller (see figure 3a) and cost a few dollars.

In summary it can be said that the proposed sensor configuration is compatible with the target domain of small, cheap smart objects.

Complexity. Any method that is to be deployed on low end sensor nodes and smart objects needs to be resource conscious. However, when considering the method proposed in this paper it is important to remember, that it is not meant for continuous tracking of a moving device. Instead we assume that the method would be run once after the acceleration sensor has detected that the device has been moved and then let to rest. Thus there is no need to deal with speed and consider the power efficiency of the algorithm. We just need to show that with typical resources available in such nodes it is feasible to either perform the required computation or transmit the data to a remote server for processing. For the sake of simplicity we restrict ourselves to the communication requirements of the raw data. With 16 bit resolution and the sampling rates given in section 3 we have a data rate of about 130Kbps for the sound and a about 05Kbps for the acceleration. These have to be sustained for total of 13 seconds.

With respect to online execution we merely point to related work by our group in which we have studied implementations of sound and acceleration based activity recognition (e.g. [17]). With sampling rates, features and classifiers similar to the ones proposed in this paper we were able to demonstrate power efficient execution on nodes using the TI MSP 430 microcontroller with less then 100K of RAM. This leads us to believe that executing the proposed method, or at least

computing most of the features (in particular FFT) to avoid transmitting the raw sound data on a low power sensor node would also be feasible.

3 Recognition Method

As described in section 2, our approach can be divided into two distinct methods, mechanical vibration and sound sampling.

Table 1. Selected features used for frame-by-frame classifications

Feature Name	Description
Standard, simple Features	Zero Crossing Rate, median, variance, 75% percentile, inter quartile range
Frequency Range Power	computes the power of the discrete FFT components for a given frequency band.
Sums Power Wavelet Determinant Coefficient	describes the power of the detail signals at given levels that are derived from the discrete wavelet transformation of the windowed time-domain signal. This feature has successfully been used by [18].
Root Mean Square (RMS)	$\sqrt{\frac{1}{N} * \sum_i x_i^2}$, with N the number of samples in a sliding window, and x_i the i'th sample of the window.
Number of Peaks	The number of peaks in the window with different thresholds, low medium and high.
Median Peak Hight	The median of the peak hight.

3.1 Vibration

During the vibration phase the device itself records the sound and the acceleration. Classification is performed separately on each signal and the information of the two modalities is combined on classifier level (see 3.3).

Vibration Sound Processing. For the vibration sound some 30 individual features were calculated over a 500 msec. sliding window (250 msec. overlap). From those we picked 5 based on initial tests and plots of the data: the zero crossing rate, the frequency range power, 75%Percentile, sums power wavelet determinant coefficient and the median. On these features we trained common machine learning algorithms, e.g. K-NN, Naive Bayes, C 4.5. We found C 4.5 to be the most robust and best (however only by a narrow margin). The frame-by-frame output provided by the C 4.5 classifier is smoothed using a majority decision over the entire length of a single vibration phase. We have also performed experiments using Hidden Markov Models either on the features calculated in the 500ms windows or on the classifier output of the frame by frame classifier. Since none of the above produced significant improvement, we have opted for the less computationally intensive majority decision.

Vibration Acceleration. The process described above for the vibration sound, is essentially repeated for the acceleration. The only differences are the length of the window (1 sec with 0.5 sec. overlapping) and the final feature set (variance, the RMS, number of peaks, median peak hight, the 75%Percentile, inter quartile range). Again C 4.5 has proven to be the best classifier and HMM has showed no advantage over the majority decision.

3.2 Sound Sampling

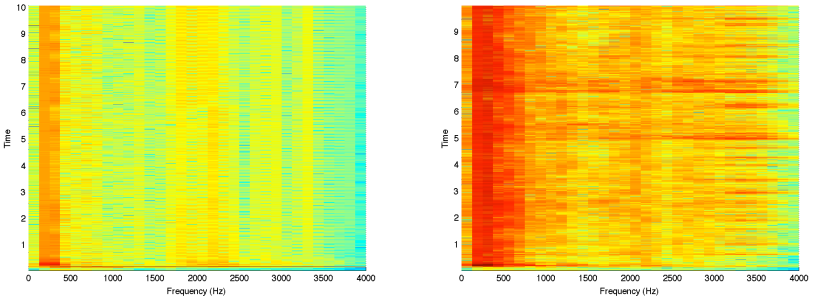
The active sound sampling procedure differs from the vibration method in several ways. We know from literature (see section 2) that few discrete frequencies between a few hundred and a few thousand Hz are enough to separate a large range of material in terms of their absorption coefficients. Therefore, we have selected 8 discrete, equidistant frequencies between 500 and 4000. The frequency range choice was dictated by the performance of small, cheap speakers (not capable of very low frequency tones) and the need for a reasonable sampling rate. From the recorded beeps we first isolate 8 frequency prints using a variable threshold. As features we have empirically selected RMS, frequency range power and the sums power wavelet determinant coefficient. These are calculated again 30 features in 200 msec. sliding windows (150 msec. overlapping).

The features of all 8 frequency prints are combined into one feature set. This means that a feature instance contains the calculated RMS etc. of each frequency band. The rest of the procedure is identical with the vibration recognition (frame by frame classification using C 4.5 and majority decision).

3.3 Fusion

The two main approaches to fusion are signal/feature level and classifier level fusion. Feature level fusion works best for features that are computed at the same sampling rate (sliding window size). This is not the case for the three recognition modalities described above. As the different window sizes were determined heuristically to produce best results for each modality, dropping them for the sake of fusion make little sense. As a consequence no direct feature level fusion was investigated. However we have investigated a fusion approach based on the results of the frame by frame classification. This can be viewed as kind of feature level fusion, since this result is input to the majority decision. Thus we have computed the majority decision for an event over the frame by frame results from all three modalities put together, instead of computing it for each modality separately.

In terms of classifier fusion we have opted for a Bayesian Belief Integration method (see. [19] for an overview of classifier fusion methods). The method uses the confusion matrix obtained from testing the classifiers on the training data set to determine class probabilities as for different combinations of classifier outputs. This allows the system to take into account the peculiarities of each classifier. With just 3 classifiers and a constrained number of classes it is also computationally tractable. If the number of classes and/or is increased the method would could be replaced by for example logistic regression.



(a) (b)

Fig. 1. The vibration sound spectrum for: (a) carpet. (b) desk.

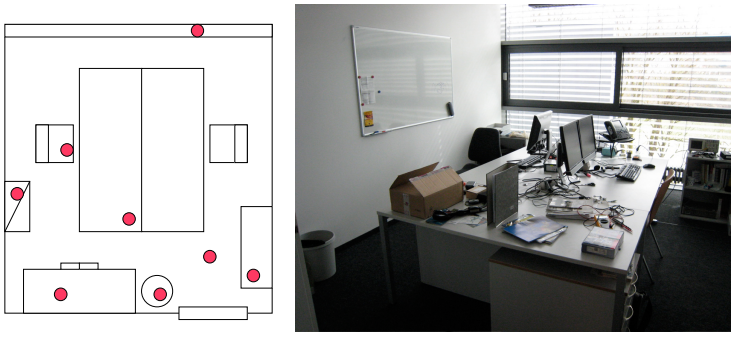
4 Experimental Validation

4.1 Validation Scenarios

Specific Location Mode. As basis for our study we have picked three scenarios: an office, a living room, and a one room student apartment. In each scenario a set of obvious locations for placing objects was selected. These included the furniture present in this room (both open such as table or sofa and closed such as cupboards), the floor, the window ledges and additional things such as the stereo. In the office scenario we have also included three pockets (two different pockets from a jacket and a jeans pocket), the inside of a backpack and a suitcase as well as a the trashcan. A full listing of the investigated location is given in table 2 and illustrated in Figure 2. There are 16 locations in the office, 9 in the living room and 10 in the apartment (total of 35).

We recorded 30 experimental runs on each specific location (a total of over 1000 events), each time randomly varying the exact position of the recording. The object was placed according to positions drawn randomly from a uniform distribution. From the 30 runs, 10 are randomly picked to train the classifiers, the remaining 20 are used as test set. Evaluation is performed first on each individual scenario (under the assumption that room level location could be obtained from other means). To see how our method behaves as the number of location increases we have also done an evaluation on a data set containing all the locations from the three scenarios.

Abstract Location Type Mode. The abstract location types were defined according to the surface material and the location being open (e.g. a table) or closed (e.g a cabinet or a drawer). As shown in table 2 this has lead us to 12 classes that include most typical surfaces (wood, glass metal stone, poster). To get a sufficient number of different instances of each class we have recorded the data in a furniture store. For every abstract class we have picked 6 different furniture. Two recordings were done on each specific piece of furniture leading to 12 data points per abstract class and a total of 144 events. For the evaluation



(a) (b)

Fig. 2. The semantic locations we try to detect are marked in red for the office in (a). In (b) you can see the actual office we conducted the experiments in.

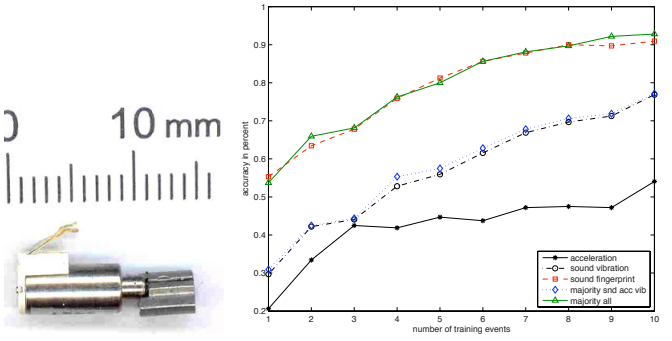
Table 2. Chosen symbolic locations and abstract location classes. The letter in front is the identification for the individual confusion matrix plots presented later in the paper. The letter in brackets behind the 3 scenarios concerned with the symbolic location, is the identifier for the confusion matrix plot over all 35 locations. In j. , o. j. and tr. pocket stand for inside jacket, outside jacket and trousers pocket.

Office		Living room	Appartment	Surfaces
a. backpack(a)	k. in j. pocket(C)	a. desk(h)	a. bath carpet(f)	a. polster open
b. cupboard(z)	l. tr. pocket (e)	b. floor(u)	b. bed(p)	b. glass open
c. suitcase(w)	m. cartbox (F)	c. sofa(n)	c. chair(b)	c. iron open
d. drawer(t)	n. ledge (H)	d. table(A)	d. desk (wood) (l)	d. stone closed
e. desk(D)	o. chair (v)	e. chair(c)	e. radiator(d)	e. wood closed
f. top drawer(E)	f. drawer (m)	f. ledge(k)	f. glass closed	
g. cabinet (x)	p. shelf (i)	g. ledge (G)	g. carpet floor(B)	g. iron closed
h. o j. pocket(j)		h. stereo (s)	h. cupboard(g)	h. metal open
i. trashcan(I)		i. tv (j)	i. drawer(q)	i. polster closed
j. carpetfloor(r)			j. wardrobe (o)	j. stone open

two pieces of furniture from each class (four events per class) were picked for training and 4 (8 events per class) were retained for testing. This is consistent with the envisioned application mode where the user would be given a device 'factory pre-trained' for each class and use it to recognize instance of the class not seen by the system before.

4.2 Experimental Procedure

Setup. For the experiments, we use the Nokia 5500 Sport. It is a mobile of Nokia's third S60 series, equipped with an accelerometer and an extra loudspeaker. The mobile is able to run C++, Java and python code. For the first experiments, we coded a C++ application to record the sensor values. Yet, we soon swapped to Python, as it is much faster for prototyping, less error-prone



(a) (b)

Fig. 3. A common vibration motor (Figure (a), picture from Ulf Seifert under the licence GNU FDL). On the right (b) is the Classification accuracy depending on the number of training events for the office scenario depicted.

debugging using an interactive bluetooth shell and still not lacking low level sensor access, through easy extensibility using C extensions. The evaluation is done in batch processing using a mixture of Python, Matlab scripts and Java code, mainly the Weka machine learning package.

Data Acquisition. An experimental run consists of the following steps. First the mobile is placed on a random spot on a particular location. A python script is used to determine this spot. Then the measurement is started. While the mobile vibrates for 5 sec. lying face up on the surface, a python script running on the mobile records the sound and acceleration simultaneously. The sound is sampled with 8000 Hz, the acceleration with 30 Hz. After the vibration measurement is done the mobile plays the sound sample consisting of 8 tunes in distinct frequencies from 500 to 4000 Hz in 500 Hz steps. Each tune is 1 sec. long. While the mobile plays this using the extra loud-speaker, the python script records the sound with 8000 Hz over the inbuilt mobile microphone. The loud speaker faces the surface, as depicted in Figure 3. We get around a problem of accessing full-duplex mode in python on the Nokia phone by using the music player and the extra speaker.

4.3 Experimental Results

The recognition performance for different scenarios experiments and recognition modalities are summarized in figures 4a (for the three individual scenarios of the specific location mode and the abstract location class) and in 4b (combining all 3 locations and second/third best voting). In addition examples of confusion matrices are visualized for the office, scenario, the combination of all three specific location mode scenarios and the abstract location type mode in figures 6a, 6c, and 7 respectively.

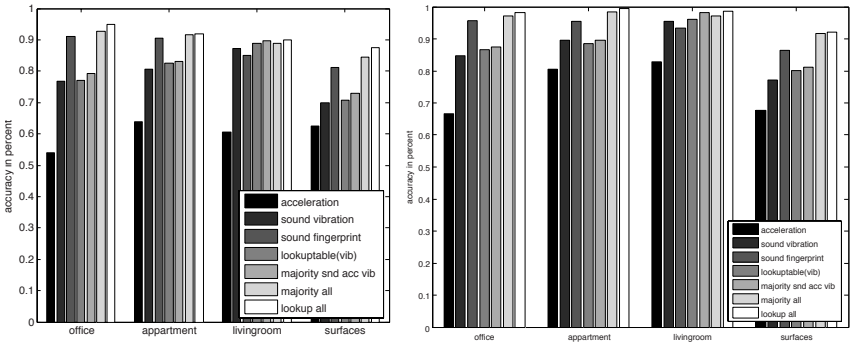


Fig. 4. Barcharts for living room, office, apartment, and abstract classes using just the first result of the classification (a) and allowing the 2nd best vote (b)

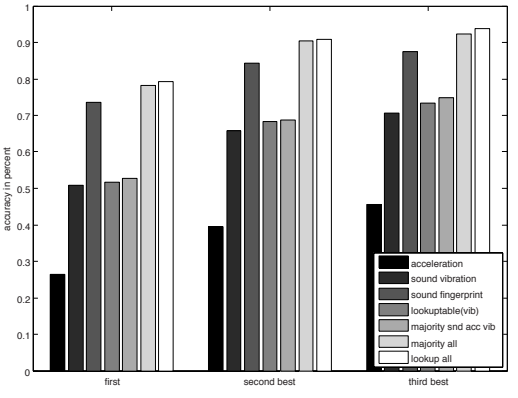


Fig. 5. Barchart for office living room, apartment and all combined including 1st 2nd 3rd best

In the more detailed discussion of the results given below and the some of the figures we at times discuss '2nd best evaluation' or '3rd best evaluation'. This refers to the percentage of cases where the correct class is among the 2 (3) first picks of the classification system.

Office. In the office scenario, 14 of the 16 locations can be classified near perfect accuracy. The single biggest confusion is between the pocket on the inside of the jacket with the one on the outside of the jacket. This is plausible and was expected. An unexpected result is the poor recognition of the metal window ledge. It is confused with the cartbox top the shelf and the chair.

The classification accuracy is 54% using the event-based acceleration classifier, 77% for vibration sound, 91% for the sound sampling, 77% and 79% for the vibration fusion cases, up to 93-94% for the majority decision and lookup-table

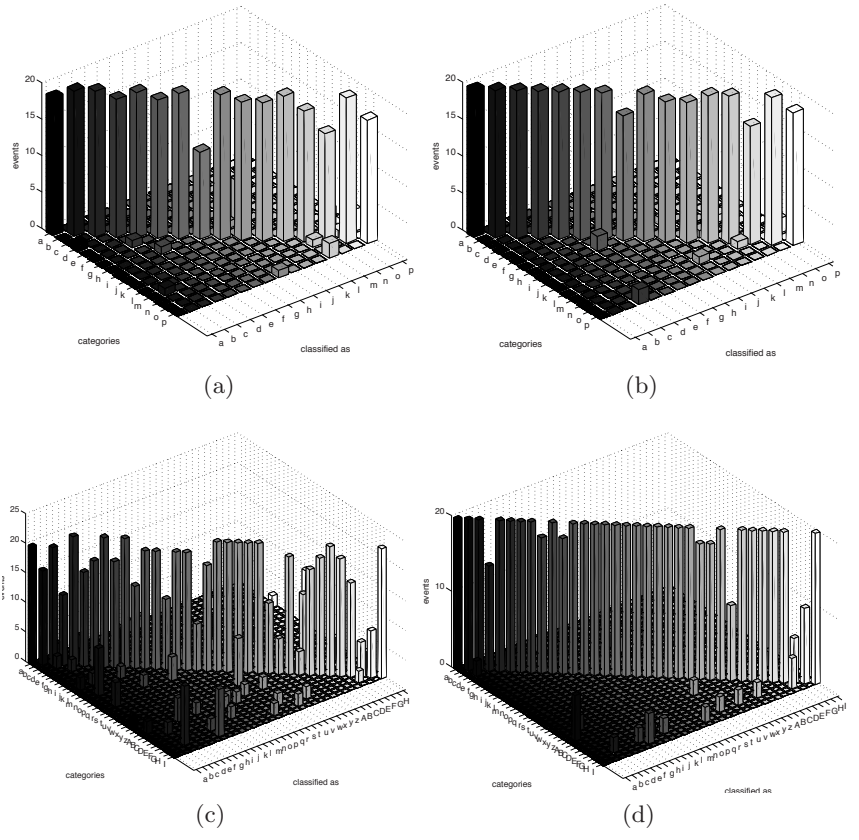


Fig. 6. The confusion matrix (a) of the office using the lookup-table fusion compared with the confusion matrix in (b) using the second best locations in addition to the lookup-table. The same is depicted, below only for all the 35 different semantic locations. Figure (c) shows the classification of the lookup-table fusion, whereas Figure (d) shows the lookup-table fusion considering up to the 3rd best.

fusion using all modalities. The sound sampling is the best non-fusion method with 91%. The '2nd best evaluation' pushes the correct classified up to 96%.

Living room. In the living room scenario, most of the samples from 7 of the 9 locations can be classified correctly. A lot of the sofa instances are confused with the chair, as the chair is also padded. This is the worst confusion. Again the classifiers perform poorly for window ledge category. The living room classification accuracy starts with 60% for acceleration alone, and goes up to 87% for the vibration sound. In this scenario, the sound sampling is worse than the vibration methods at 85%. This explains also why the fusion methods on top of the vibration work so well and are nearly as good as the fusion over all methods, at 88 and 89% respectively. The fusions over all methods just 0.5 % better. Only one/two events are corrected by this fusion. In the '2nd best evaluation' the

accuracy ranges from 66% for acceleration alone, up to 97% for the lookup-table fusion over all methods. Here also the acceleration and sound vibration fusion do extremely well with 93% and 96%.

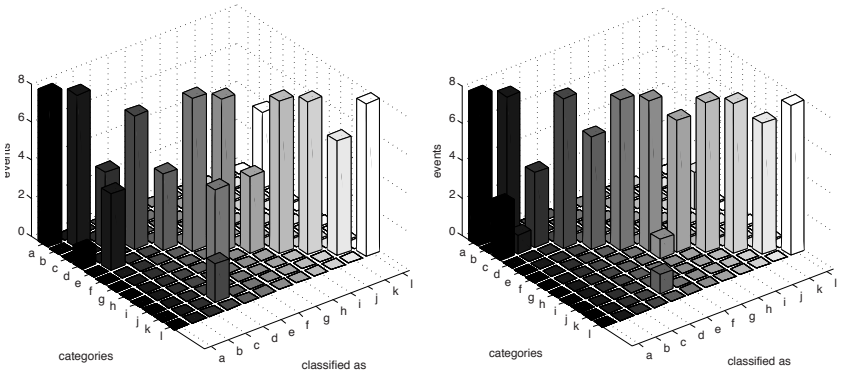
Appartment. In the appartment case, the worst miss classification happens in the cupboard category, which is confused with the desk. Both are made out of the same wood. The radiator class is also confused with several other classes. Here the acceleration accuracy is at 65%, the vibration sound at 81%, sound sampling 90%. The fusion using just the vibration method is at 82 and 84% respectively, as with all the fusion examples the lookup-table is slightly better. Finally, the fusion techniques on all 3 modalities are all over 90%. Taking a look at the '2nd best evaluation', there the accuracy ranges from 80 % for acceleration to up to 99% for the look-up table over all three classifiers.

Combined over all rooms (35 classes). As already seen in the single scenarios, the ledge classes perform poorly; even in the 2nd and 3rd best evaluation. Also one of the table classes does badly and is confused with several other classes. The classification accuracy over all 35 semantic locations is expectably lower that those of the single scenarios, ranging from 26% for acceleration, 51% for vibration sound, 74% for sound sampling, over 52% for the vibration fusion, up to 78% for the fusion of all methods. The 2nd and 3rd best evaluations look considerably better. Second best is up to 90%. Third best reaches 94%.

Abstract Location Classes. For the abstract classes, the iron and wood classes are easily confused, as well as the stone and glass. Acceleration classification alone performs reasonably well compared to the other scenarios with 63%. Sound vibration is better with 69%. As nearly always, sound sampling performs better compared to the vibration method, with 81% accuracy. For the fusion techniques, also nothing surprising. The vibration fusion majority decision is at 70%, the vibration lookup-table around 71% accuracy. The two fusions based on all methods are on 83% for the simple majority decision case and 86% for the lookup-table. Allowing the second best classification method, one can stem up the performance to 92% for the lookup-table fusion method.

4.4 Lessons Learned and Implications

Overall Performance. The performance of the system is extremely inhomogeneous with respect to the classes. There is a large proportion of classes for which the classification is perfect or near perfect, and a small one with very poor performance (see confusion matrices in figures (6a, 6b, 6c and 6d)). As a consequence the overall recognition accuracy figures are strongly influenced by few classes that the system has problems with. This is best exemplified by the abstract location type confusion matrix and 3rd best evaluation of the combined specific location classes. As can be seen in the plots the former has 8 perfect or near perfect classes, 1 reasonably good ad 3 very poor. The latter has 31 perfect to very good (27 perfect), 1 mediocre and 3 very poor classes.



(a) (b)

Fig. 7. Confusion matrix (a) of the abstract classes compared with the corresponding 2nd best confusion matrix in (b)

Class by Class Performance. For some of the classes such as the inside and outside pocket poor performance was expected, as they were included to test the limits of the system. In fact the recognition for this locations is better then expected. Better then expected recognition has also been achieved in a number of locations that were included as 'hard cases' such as the backpack and the trousers pocket. Surprising is the poor performance of the window ledge and the radiator. At this stage we have no verified explanation. On possibility is a spatial inhomogeneity of those symbolic locations. On the ledge the sound sampling is certainly different depending when the speaker faces the window and when it faces away from the window.

Value of the 2nd and 3rd Best Evaluation. The performance of the system is particularly appealing for applications that can live with choice of two or three most probable locations as system output. This has already been mentioned for the case of 3rd best evaluation of the 35 combined symbolic locations. For the other data sets even the 2nd best produces close to perfect recognition for the vast majority of classes.

Value of Different Classification Modalities. While from the discussion in 2 it was to be expected that sound sampling will produce the best results and acceleration the poorest, the difference between the two is larger then we expected. In particular the fact that in most cases little is gained by adding acceleration and vibration sound to the sound fingerprint is surprising. On the other hand combining vibration sound and acceleration is often produces significant gains.

Significance of Training set Size. For the specific location mode the user needs to train the system for every single relevant location. Thus the training effort is a significant issue. As shown on the example of the office scenario in figure 3b

the system starts to display significant recognition performance from around 5 training examples and stagnates at about 10. We have found this behavior to be typical for all the specific location mode scenarios.

5 Conclusion and Future Work

Summarizing the discussion from section 2.2 and the experimental results from section 4.3 we conclude the following

1. The proposed method is well suited for low end, simple sensor nodes and smart objects and requires no additional positioning infrastructure.
2. The key source of information is sound sampling. Thus if size is critical the vibration motor can be left out.
3. The system can reliably (90% and more accuracy) resolve a sufficient number of specific locations to cover one room a or a small flat. It is advisable to combine our system with room level positioning
4. The performance of the system is extremely inhomogeneous with respect to the classes, with most classes being recognized with high accuracy and few 'rogue' lasses showing very poor performance.
5. Settling for the two or three best picks instead of a crisp single classification greatly increase the number of locations that are reliably recognised and the tolerance towards 'rogue' classes.
6. If training by the user is an issue the abstract location class mode offers a possibility to provide 'pre trained' systems at the price of a more 'fuzzy' location information.

Key points to investigate in the future are improved vibration sampling (using different amplitudes and frequencies to improve acceleration based performance), an investigation of the sources of errors on the problematic classes, more elaborate fusion methods, and a combination with radio signal strength based location methods. In addition in specific application projects, in particular in the area of assisted living, we will work towards real life applications of our method.

References

1. Becker, C., Dörr, F.: On location models for ubiquitous computing. Personal Ubiquitous Comput. 9, 20–31 (2005)
2. Hightower, J., Borriello, G.: Location systems for ubiquitous computing. Computer 34, 57–66 (2001)
3. Want, R., Hopper, A., Falcö, V., Gibbons, J.: The active badge location system. ACM Trans. Inf. Syst. 10, 91–102 (1992)
4. Priyantha, N., Chakraborty, A., Balakrishnan, H.: The Cricket location-support system. In: Proceedings of the 6th annual international conference on Mobile computing and networking, pp. 32–43 (2000)
5. Hazas, M., Kray, C., Gellersen, H., Agbota, H., Kortuem, G., Krohn, A.: A relative positioning system for co-located mobile devices. In: MobiSys '05: Proceedings of the 3rd international conference on Mobile systems, applications, and services, pp. 177–190. ACM Press, New York (2005)

6. Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based user location and tracking system. In: INFOCOM, (2), pp. 775–784 (2000)
7. LaMarca, A., Chawathe, Y., Consolvo, S., Borriello, G., et al.: Place lab: Device positioning using radio beacons in the wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, Springer, Heidelberg (2005)
8. Krumm, J., Cermak, G., Horvitz, E.: Rightspot: A novel sense of location for a smart personal object. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, Springer, Heidelberg (2003)
9. Doherty, L., Pister, K.S.J., Ghaoui, L.E.: Convex position estimation in wireless sensor networks. In: INFOCOM, IEEE, Los Alamitos (2001)
10. Scott, J., Dragovic, B.: Audio location: Accurate low-cost location sensing. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, Springer, Heidelberg (2005)
11. Van Kleek, M., Kunze, K., Partridge, K., Bo Begole, J.: Opf: A distributed context-sensing framework for ubiquitous computing environments. In: Youn, H.Y., Kim, M., Morikawa, H. (eds.) UCS 2006. LNCS, vol. 4239, Springer, Heidelberg (2006)
12. Holmquist, L.E., Mattern, F., Schiele, B., Alahuhta, P., Beigl, M., Gellersen, H.W.: Smart-its friends: A technique for users to easily establish connections between smart artefacts. In: UbiComp '01: Proceedings of the 3rd international conference on Ubiquitous Computing, pp. 116–122. Springer, London, UK (2001)
13. Lester, J., Hannaford, B., Borriello, G.: "are you with me?" - using accelerometers to determine if two devi ces are carried by the same person. In: Ferscha, A., Mattern, F. (eds.) Pervasive Computing (2004)
14. Kunze, K., Lukowicz, P., Junker, H., Tröster, G.: Where am i: Recognizing on-body positions of wearable sensors. In: LOCA'04: International Workshop on Location- and Context-Awareness, Springer, London, UK (2005)
15. Griffin, J., Fyke, S.: User hand detection for wireless devices U.S.Patent 20,060,172,706 (August 3, 2006)
16. Olson, H.: Music, Physics and Engineering. Courier Dover Publications (1967)
17. Stager, M., Lukowicz, P., Troster, G.: Implementation and evaluation of a low-power sound-based user activity recognition system. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 138–141. Springer, Heidelberg (2004)
18. Sekine, M., Tamura, T., Fujimoto, T., Fukui, Y.: Classification of walking pattern using acceleration waveform in elderly people. Engineering in Medicine and Biology Society 2, 1356–1359 (2000)
19. Ruta, D., Gabrys, B.: An overview of classifier fusion methods. Computing and Information Systems 7, 146–153 (2000)