Multimodal Indoor Localization: An Audio-Wireless-Based Approach

Oriol Vinyals
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
e-mail: vinyals@icsi.berkeley.edu
Eladio Martin
Dept. of EECS
University of California, Berkeley
Berkeley, CA 94720
e-mail: emartin@berkeley.edu
Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
e-mail: fractor@icsi.berkeley.edu

Abstract—Location-based services on mobile devices have become a key element in today's wireless and mobile phone infrastructure, which due to their potential for precise personalization offer interesting opportunities for Semantic Computing. However, location information is mostly only available outdoors and current indoor localization schemes are not very accurate. In this paper, we therefore present a novel approach for indoor localization using multiple modalities of information that are easily available indoors on handheld devices. We use the microphones plus the various wireless signals that are sensed by smartphones to serve as input for a novel localization approach. Our proposed approach is computationally lightweight and, by making use of recent machine learning techniques for integrating modalities, achieves greater accuracy than current work in the area.

Index Terms: localization, indoor, audio, wifi, multimodal

I. Introduction

Location-based services are rapidly gaining traction in the online world. An extensive and rapidly growing set of online services is collecting, providing, and analyzing geo-information. Besides major players like Google and Yahoo!, there are many smaller start-ups in the space as well. The main driving force behind these services is the enabling of a very personalized and intuitive experience. Foursquare for example encourages its users to constantly "check-in" their current position, which they then propagate on to friends; Yowza!! provides an iPhone application that automatically locates discount coupons for stores in the user's current geographical area; and SimpleGeo aims at being a one-stop aggregator for location data, making it particularly easy for others to find and combine information from different sources. In a parallel development, a growing number of sites now provide public APIs for structured access to their content, and many of these already come with geo-location functionality. Flickr, YouTube, and Twitter allow queries for results originating at a certain location. Currently, however, GPS is not available indoors or where there is no line of sight with the satellites. So the aforementioned services only work very limitedly. For this reason, research has recently started on inventing indoor localization methods to enable geo-location where it is not regularly available. In other words, indoor localization would fill an important usability gap for many Semantic Computing applications.

In this paper, we present a novel approach for indoor localization using multiple modalities of information that are typically available indoors: The presence of microphones in the devices that we carry in connection with various wireless signals sensed by current smartphones serve to indicate the location with about +/- 3 m accuracy. Our proposed approach is computationally lightweight and, by making use of recent machine learning techniques for integrating modalities, achieves greater accuracy than current work in the area. Also the modalities truly complement each other: wireless signal localization is global as it indicates location in a specific building in the world with about room accuracy, audio localization is local with sub-room accuracy.

The article is organized as follows. We introduce and compare related work in Section II before Section III starts to introduce our proposed algorithm, which is extensively evaluated in Section IV. Section V finally concludes the article and presents future directions.

II. BACKGROUND

Our work is based on the fusion of two modalities that have previously been used for indoor localization individually: wireless signals and acoustics. In the following we therefore present related work in the two individual domains.

A. RSSI-based localization

Radio Signal Strength Indications (RSSIs) can be translated into distances from beacon points by means of theoretical or empirical radio propagation models. The following expression accounts for a general radio propagation model delivering the received power P_r:

$$P_r = P_t \left(\frac{\lambda}{4\pi d}\right)^n G_t G_r$$

Where P_t represents the transmitted power, λ the wavelength of the radio signal, G_t and G_r the gains of the transmitter and receiver antennas respectively, d the distance separating them, and n is the path loss coefficient, typically ranging from 2 to 6 depending on the environment. The two main approaches for the estimation of location making use of RSSI values are: 1) fingerprinting, where a pre-recorded radio map of the area of interest is leveraged to infer locations through

2010 IEEE International Conference on Semantic Computing
978-0-7695-4154-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSC.2010.87

2010 IEEE Fourth International Conference on Semantic Computing
978-0-7695-4154-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSC.2010.87
120
2010 IEEE Fourth International Conference on Semantic Computing
978-0-7695-4154-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSC.2010.87
120

best matching, and 2) propagation-based, in which RSSI values are used to calculate distances through the computation of the path loss. Propagation-based techniques can face errors of up to 50% due to multipath (reverberation), non line-of-sight conditions, interferences and other shadowing effects, rendering this technique unreliable and inaccurate, especially for indoor environments, where multipath is very important. Several authors have tried to improve the efficiency of this technique for indoor environments, introducing new factors in the path loss model to account for wall attenuation, multipath or noise [1], but the hardware and software requirements due to the complexity of the method and the overall poor accuracy achieved makes this approach not feasible for current state of the art smart phones. On the other hand, fingerprinting techniques have already proven to be able to deliver better accuracies [2]. In these techniques, the mobile terminal estimates its location through best matching between the measured radio signals and those corresponding to locations previously registered in the radio map. This process consists of two stages:

1) Training phase, also called offline phase, in which a radio map of the area in study is built. RSSI values from different beacons are recorded at different locations; the separation between these chosen locations will depend on the area in study, and for instance, for indoor environments this separation can be of around a meter [3]. Each measurement consists of several readings, one for each radio source in range [4].

2) Online phase, in which the mobile terminal infers its location through best matching between the radio signals being received and those previously recorded in the radio map. Localization algorithms employed in this case generally make use of deterministic or probabilistic techniques.

Deterministic techniques store scalar values of averaged RSSI measurements from the access points [5]. The most relevant techniques in this group are closest point, or nearest neighbor in signal space [6]; nearest neighbor in signal space-average [5], [7], choosing k nearest neighbors and calculating the centroid of that set; and smallest polygon, selecting several nearest neighbors which will form various polygons, and the centroid of the smallest polygon will be considered as the estimated location [5].

Probabilistic techniques choose the location from the radio map as the one with the highest probabilities, and usually require the storage of RSSI distributions from the different beacons at each location in the radio map [6]. Fingerprinting techniques are especially appropriate for the range of frequencies in which GSM and WiFi networks operate (aprox. 850 MHz to 2.4 GHz) because of two main reasons [4]: the signal strength at those frequencies presents an important spatial variability, and also a reliably consistency in time (despite the variable nature of radio signals).

Considering GSM as an example for cellular communications technology, although it makes use of power control both at the mobile terminal and base station, the data on the Broadcast Control Channel (BCCH) is transmitted at full and constant power, making this channel suitable for fingerprinting [4]. Several authors have tried this approach for localization but it requires dedicated and complex hardware. In order to improve the accuracy of this approach, a selection among all the measured signals is recommended, rejecting those which are either too noisy, too stable across all areas or simply do not provide enough information [3].

Regarding WiFi technology, several research groups have already tried to leverage RSSI fingerprinting for localization:

- Radar [8]: represents the first fingerprinting system achieving the localization of portable devices, with accuracies of 2 to meters.
- Horus [9]: based on the Radar system, it manages a performance improvement making use of probabilistic analysis.
- Compass [10]: applies probabilistic methods and leverages object orientation to improve precision, claiming errors below 1.65 meters.

Besides cellular communications and WiFi technologies, the

RSSI fingerprinting technique for localization can be utilized with other radio frequency technologies including:

- Bluetooth, which despite the extra infrastructure requirements in comparison with WiFi, it can achieve accuracies in the range of 1.2 meters.
- Conventional radio, can also be used for localization. However, the requirement of dedicated hardware and the fact that devices can be located only down to a suburb, represent important drawbacks.
- Digital TV signals have also proved to be suitable for localization, but subject to dedicated hardware requirements and low resolutions.
- Zigbee technology can also be applied for localization through fingerprinting [11], achieving accuracies of approximately 2 meters. However, this technology also requires extra hardware for a correct implementation, constituting a major drawback.

Nevertheless, all the existing approaches use dedicated and complex hardware, making them not feasible for direct implementation in current state of the art smart phones.

B. Audio-based localization

Different approaches have been developed to analyze a scene using acoustics, such as those derived from Computional Auditory Scene Analysis (CASA). A very important practical method is to measure the time-delay-of-arrival on different microphones and thus leveraging the travel time differences of the signals to microphones in different positions, such as the work presented in [12], [13]. The current standard method for computing the time-delay-of-arrival features is the so-called GCC PHAT algorithm [14]. The idea is to correlate the signals from the different microphones under the assumption that they are identical but phase-shifted. A very important advantage of the technique is that the microphone positions

Fig. 1. Our localization application on two different Android-based phones.

do not have to be known. In [15] the authors present a novel audio-visual approach for unsupervised speaker localization. Using recordings from a single, low-resolution room overview camera and a single far-field microphone, a state-of-the-art audio-only speaker localization system (traditionally called speaker diarization) is extended so that both acoustic and visual models are estimated as part of a joint unsupervised optimization problem [16]. The speaker diarization system first automatically determines the speech regions, the number of speakers, and estimates "who spoke when". Then, in a second step, the visual models are used to infer the location of the speakers in the video. The experiments were performed using 4.5hours of real-world meetings. However, the system assumes stationary microphones and a stationary camera.

Like most audio and visual localization methods, this method focusses on a specific scenario. The idea of the work in this article is to enhance wireless localization to sub-room accuracy by adding acoustic localization, independent of a scenario.

To the best of our knowledge, our proposed indoor localization application is the first one to not only accurately leverage RSSI fingerprinting on smart phones but to also include acoustic processing as a second modality.

III. Proposed Approach

In the following section we explain our approach by modality before discussing the multimodal integration step.

A. Wireless

We have studied the possibilities offered by three available resources in smart phones: WiFi radio, cellular communications radio and accelerometer, with the aim to build a multimodal approach for localization. As will be explained later in this Section, the consideration of WiFi radios represents the most reliable approach for indoor localization in our experimental setup in buildings across the University of California, Berkeley campus. In fact, RSSI information from WiFi beacons deployed within buildings allows us to obtain a radio map of different locations via fingerprinting. We can estimate locations through the comparison of the current RSSI measurements with those stored in the radio map. Different

Fig. 2. Visualization of the result of applying acoustic-only localization as described in Section III-B. The numbers indicate different positions of speakers and the dots represent the localization results. Audio-only localization already shows promising accuracy and requires no calibration. However, it needs further information to infer absolute geo-location.

attempts to obtain RSSI-based indoor localization without fingerprinting show an important loss of accuracy [17]. Also, many fingerprinting-based localization systems make use of dedicated hardware for the collection of data in the training phase. Then, in the measurement phase, the actual mobile device used for localization is different resulting in an error called signal reception bias [18] caused by differences in antenna characteristics and measurement acquisitions schemes between different equipment. Our experiments showed an average difference of approximately 10 dB between RSSI values measured with a Dell Latitude laptop and those measured with a Motorola Droid cell-phone. Consequently, we have integrated both the training and measurement phases into the same mobile device. The feature vectors extracted consist of the RSSI values at each second for each of the stations (see Figure 1). To the best of our knowledge, our application is the first one following this approach with smart phones. Moreover, the way the fingerprints are taken in the training phase should reproduce as accurately as possible in the way the measurements will be carried out in the localization phase. In this sense, the orientation of the phone (obtainable from accelerometer and magnetometer data) helps enhance the localization accuracy.

B. Audio

The overall idea of integrating audio is that most mobile devices, such as cell-phones or laptops have at least one microphone. Via networking, these devices can then communicate and form a microphone array which can be used to aid and perform localization, as discussed in Section II-B.

We have explored several possibilities to use audio for localization. Given several microphones distributed in a room and assuming the positions of the microphones are fixed, the first and most straightforward approach would be to measure the energy at each microphone, thus inferring the distance of the audio source to each of the microphones. However, this

Method	Room
Closest Point	85 %
Nearest Neighbor	78 %
Smallest Polygon	84 %
NN in signal and AP averages	87 %

TABLE I
COMPARISON OF ACCURACIES OF DIFFERENT WIFI-ONLY APPROACHES FOR LOCATION ESTIMATION (TABLE SHOWS % SUCCESS) WITH ABOUT ROOM-ACCURACY (9 m^2).

is not very robust as acoustic signals tend to be fairly sparse, and instead we use the cross correlation between each audio channel, and the corresponding delay. With a sampling rate of 48000 samples per second, and taking 340m/s as the speed of sound, we get that one sample of delay is approximately 1 centimeter of displacement. Thus, we can accurately detect changes of fractions of a meter, which in our scenario allows us to achieve good accuracies. Note that since the speed of sound can be assumed constant, no calibration step is needed for this localization approach. The delays between each channel is computed with BeamFormIt [19], and the delays are treated as feature observations in a standard Hidden Markov Model (HMM), where each state corresponds to one particular position in the room. Figure 2 shows the distinct positions in the room, with the features from the delay features projected onto a 2 dimensional PCA space (note that we have 4 microphones and, thus, our original features space of delays is 3 dimensional).

C. Multimodal integration

When integrating the WiFi measurements and the acoustic delay features several issues must be addressed. First, the sampling rate of the extracted delay features is 100 samples per second, whereas the wireless measurements are taken once every second. Thus, we consider observations every second by averaging the delay features to match the wireless measurements. The next consideration is offset synchronization between both modalities. In our case, we assume that both modalities are aligned (i.e. we manually align them), although in a fully automated system this may not be trivial as each cell phone and the recording devices for the microphones must be synchronized with precision of at least half a second. One possibility to add robustness to this approach is to perform speaker diarization [16] (see Section II-B) on the audio track, to smooth the state transition of the HMM according to speaker turns. Since in our scenario, only four positions far apart were considered, we did not perform this step. However, both in terms of complexity and integrability, this can be trivially added depending on the specific needs of the task.

To leverage the information from both modalities, we consider a discriminative version of a sequence model, similar to a Hidden Markov Model (HMM). Even though more elegant solutions such as the Max Margin Markov Networks (M3N) [20] have been proposed, we use an approach that consist of two simple steps. It performs similarly and is easier

Modality	Accuracy (SVM)	Accuracy (SVM+CRF)
Random	25%	25%
Audio	86%	91%
$Wifi_1$	84%	87%
$Wifi_2$	33%	35%
Audio + $Wifi_2$	86%	92%

TABLE II
COMPARISON OF THE ACCURACIES OF DIFFERENT MODALITIES FOR LOCATION ESTIMATION OF FOUR CORNER POINTS IN A 9 m^2 ROOM (TABLE SHOWS % SUCCESS).

to implement. The details are described in [21]. The first step is a training step using two Support Vector Machine (SVM) classifiers, which are trained separately on the sensor output of each modality. The output values of the SVMs are interpreted as confidence values between -1/1 for each class and are trained using a separate training set. Then, using the binned confidence values for both modalities as binary features in a Conditional Random Field (CRF) the sequential temporal behavior of the data is modeled.

IV. EXPERIMENTAL RESULTS

In the following section we present accuracy measurements that provide evidence for the accuracy of the localization, as well as the value gained from the multimodal integration.

A. WiFi-only Baseline

First, we carried out tests to measure different radio frequency signal strengths within the Cory building in the University of California, Berkeley campus. WiFi technology offers a reliable approach for indoor localization in a building, i.e. room-accuracy, because of the important deployed infrastructure of WiFi access points providing coverage in the whole building. For the measurement of the signals and practical implementation of our localization application, we have used smart phones running on Android, in particular the G1 and the Droid. The sensitivities of the Android phones range from -45 dBm to -104 dBm. It must be noted that values below -85dBm are generally too inconsistent to be leveraged as reference. Consequently, RSSI values above -80dBm are desirable in order to obtain reliable results. Processing this information statistically, we have built an Android application for localization, and we have tested it in locations where 25 WiFi radios in average were listened (approximately 40% of them with RSSI above -80dBm), obtaining acceptable accuracies to discriminate between rooms, as shown in Table I.

It must also be noted that as the number of WiFi radios (and their RSSI values) that can be listened in a specific location decreases (e.g. only 2 WiFi radios with RSSI values above -75dBm), our application's location accuracy drops.

B. Multimodal Approach

In our experimental setup, each WiFi access point has 5 radios (each represented by a MAC address). RSSI values (in dBm) from the same access point can show important standard deviations in between consecutive scans (within the same

radio) and also in between different radios within the same access point. Consequently, averaging of values both within the same access point and over time provides much more stable values that can successfully be used as a fingerprint component of each particular location. We call this approach Nearest Neighbor in signal space and access point averages, and the results summarized in Table I show that our approach can outperform existing deterministic techniques (the resolution metric, in percentage, accounts for the number of true positives obtained during localization tests).

In order to test the most demanding application, that is to detect and track people within the same room, we performed the audio experiments in a controlled scenario, more specifically in UC Berkeley's Tele-Immersion laboratory [22]. This is a well-defined space of 3 meters by 3 meters with four microphones situated on the corners of the space. We measured the audio delays at the exact same points as the WiFi experiments using the techniques described in Section III-B. One of the main challenges imposed by the Wifi data is that it depends on external factors such atmospheric conditions or position of objects in the room [23]. Therefore, measurements performed close to each other (i.e. within the virtual space) vary greatly. Therefore, either the WiFi subsystem needs to be retrained or the accuracy drops dramatically. Adding audio to the system effectively help to adjust for this effect.

In our first experiment, separate SVM classifiers were trained for audio and Wifi on the four corners of the virtual space. The accuracy drops from 84%, to 33% (barely above the baseline 25% accuracy of random choice) if the WiFi measurements are taken on a different day rather than within minutes of each other (this refers to $Wifi_1$ and $Wifi_2$

respectively in Table II). We assume that calibrating the WiFi subsystem so often defeats the purpose of this experiment. Therefore we only focus on the most challenging (and therefore least accurate) scenario where the test measurements were taken on a different day. This results in much less temporal correlation due to the time variance of the channel.

In a second experiment we added the use of a Conditional Random Field (CRF) for temporal smoothness. This helps in all the cases. However, the gain of having the WiFi modality is quite modest since the baseline accuracy of the system in this accuracy range is quite low, due to the challenging fact that the conditions between training and testing on the WiFi signal strength is dependent on external, non-controllable factors such as weather. To better train our models, differences between signal strengths rather than absolute values could be used, as well as exploring possible adaptations/corrections of the models depending on varying conditions (or training several models for each of them). We expect the accuracy to drop slightly, when microphones are not stationary but as described in Section III-B, the audio localization does not require the knowledge of the positions of the microphones and therefore requires only limited stationarity.

In summary one can see that the multimodal approach performs much better then the WiFi-only method even when compared to only room-accuracy.

Fig. 3. Localization application in the Droid showing location information as multimedia messages.

V. Conclusion and Future Work

This article described a novel indoor localization method based on the integrated use of WiFi signals and acoustic signals picked up through microphones. We presented an implementation (see also Figure 3) and compared it to related work.

Our work indicates that a multimodal approach is feasible and that the integrated use of two modalities readily available on any mobile device benefits the accuracy. While for this initial experiment stationary microphones were used, the idea is to combine the microphones of different mobile devices and have them synchronize through network communication. Also, we assume that in the future RSSI fingerprints can be stored in central databases (as currently done in Google Maps) and/or communicated peer-to-peer among different handheld devices.

We believe indoor localization will enable location-based services to be much more intuitive since it allows these service to properly work in places where GPS signals are currently not present. In other words, location based services could work in university buildings, malls, airports, and inside museums and enable the next generation of intuitive and personalized Semantic Computing experiences.

Acknowledgements

This research was supported partly by a stipend of the Spanish Ministry of Science and Innovation (MICINN) and by a US NGA NURI grant. Opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the supporters. We thank Prof. Ruzena Bajcsy for making the UC Berkeley Tele-Immersion laboratory available for this research.

References

[1] R. Singh, M. Gandetto, M. Guainazzo, D. Angiati, and C. Ragazzoni., “A novel positioning system for static location estimation employing WLAN in indoor environment,” in Proceedings of the IEEE PIMRC, 2004, pp. 1762–1766.

[2] P. Brida, P. Cepel, and J. Duha, “Geometric Algorithm for Received Signal Strength Based Mobile Positioning,” in Proceedings of the Czech Slovak Technical Universities and URSI, 2005.

[3] A. Varshavsky, E. Lara, H. J., A. Lamarca, and V. Otsason, “GSM indoor localization,” Pervasive and Mobile Computing, vol. 3, no. 6, pp. 698–720, 2007.
[4] V. Otsason, A. Varshavsky, A. Lamarca, and E. De Lara, “Accurate GSM Indoor Localization,” in Proceedings of UbiComp, 2005, pp. 141–1582.
[5] A. Roxin, J. Gaber, M. Wack, and A. Nait-Sidi-Moh, “Survey of Wireless Geolocation Techniques,” in Proceedings of Globecom Workshops, 2007, pp. 1–9.
[6] A. Dempster, B. Li, and I. Quader, “Errors in deterministic wireless fingerprinting systems for localization,” in Proceedings of ISWPC, 2008, pp. 111–115.
[7] M. A., H. N., Y. J., and W.-S. S., “Indoor Localization Using Multiple Wireless Technologies,” in Proceedings of MASS, 2007, pp. 1–8.
[8] P. Bahl, V. Padmanabhan, and A. Balachandran, “Enhancements to the RADAR user location and tracking system,” in Technical Report MSR-TR-00-12, Microsoft Research, 2000.
[9] M. Youssef, “A WLAN-based indoor location determination system,” in Ph.D. Dissertation, University of Maryland, 2004.
[10] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg, “COMPASS: A Probabilistic Indoor Positioning System Based on 802.11 and Digital Compasses,” in Proceedings of 1st WiNTECH, 2006, pp. 34–40.
[11] A. Noh, W. Lee, and J. Ye, “Comparison of the Mechanisms of the Zigbee’s Indoor Localization Algorithm Software Engineering,” in Proceedings of 9th ACIS Int. Conf., 2008, pp. 13–18.
[12] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop, 2007.
[13] J. Pardo, X. Anguera, and C. Wooters, “Speaker diarization for multiple-distant-microphone meetings using several sources of information,” IEEE Transactions on Computers, vol. 56, no. 9, pp. 1212–1224, 2007.
[14] M. Brandstein and H. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” in in Proceedings of IEEE ICASSP, 1996.
[15] G. Friedland, C. Yeo, and H. Hung, “Visual speaker localization aided by acoustic models,” in Proceedings of ACM Multimedia. ACM, October 2009, pp. 195–202.
[16] G. Friedland, H. Hung, and C. Yeo, “Multi-modal speaker diarization of real-world meetings using compressed-domain video features,” in ICASSP ’09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4069–4072.
[17] X. Li, “Ratio-based zero-profiling indoor localization,” in Proceedings of the IEEE 6th Int. Conf. MASS, 2009, pp. 40–49.
[18] C. Hsu and C. Yu, “An Accelerometer based approach for indoor localization,” in Proceedings of the Symposia and Workshops on UIC’09 and ATC’09 Conferences, 2009, pp. 223–227.
[19] J. Pardo, X. Anguera, and C. Wooters, “Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features And Inter-Channel Time Differences,” Proceedings of Interspeech, pp. 1337–Thu1A1O.5, 2006.
[20] B. Taskar, C. Guestrin, and D. Koller, “Max-margin markov networks,” in Advances in Neural Information Processing Systems 16, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.
[21] G. Hoefel and C. Elkan, “Learning a two-stage svm/crf sequence classifier,” in CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management. New York, NY, USA: ACM, 2008, pp. 271–278.
[22] J.-M. Lien, G. Kurillo, and R. Bajcsy, “Multi-camera tele-immersion system with real-time model driven data compression,” The Visual Computer, vol. 26, no. 1, pp. 3–15, January 2010.
[23] P. Pathanawongthum, N.; Cherntanomwong, “Empirical evaluation of RFID-based indoor localization with human body effect,” in Proceedings of the 15th Asia-Pacific Conf. on Communications, 2009, pp. 479–482.