Montage: Combine Frames with Movement Continuity for Realtime Multi-User Tracking

Lan Zhang*†, Kebin Liu*‡, Yonghang Jiang*, Xiang-Yang Li†‡§, Yunhao Liu*‡, Panlong Yang¶
*School of Software, Tsinghua University
†Department of Computer Science and Technology, Tsinghua University
‡TNLIST, Tsinghua University
§Department of Computer Science, Illinois Institute of Technology
¶PLA University of Science and Technology

***Abstract*—In this work we design and develop Montage for real-time multi-user formation tracking and localization by off-the-shelf smartphones. Montage achieves submeter-level tracking accuracy by integrating temporal and spatial constraints from user *movement vector* estimation and distance measuring. In Montage we designed a suite of novel techniques to surmount a variety of challenges in real-time tracking, without infrastructure and fingerprints, and without any a priori user-specific (*e.g.*, stride-length and phone-placement) or site-specific (*e.g.*, digitalized map) knowledge. We implemented, deployed and evaluated Montage in both outdoor and indoor environment. Our experimental results (847 traces from 15 users) show that the stride-length estimated by Montage over all users has error within $9cm$, and the moving-direction estimated by Montage is within 20^o. For real-time tracking, Montage provides meter-second-level formation tracking accuracy with off-the-shelf mobile phones.**

I. Introduction

Tracking the spatial-temporal formation of multiple mobile users plays an important role in many applications, *e.g.*, real-time team-formation tracking for team-sports strategy study, animal community monitoring for behavior analysis, and virtual-reality interactive games. When users are outdoor, localization and tracking could be solved by GPS. The accurate indoor tracking/localization in realtime is still challenging and has attracted considerable research efforts.

One category of existing methods are based on fingerprints, *e.g.*, [6], [14], [20], [27], which achieve room-level (meter-level) accuracy. Those methods, however, are typically labor intensive and environment restrictive during fingerprint collection stage. Many dedicated systems with specialized hardware, *e.g.* sensors [28] and RFID [11], can achieve high accuracy, but are not applicable for phones. Another category of approaches are range-based using different metrics. The acoustic based methods on commercial mobile handset address the issue of meter-level *pair-wise* ranging, *e.g.*, [15], [16]. Some other solutions use code division multiple access (CDMA) acoustic telemetry to simultaneously monitor the movements of numerous individual users, *e.g.*, [1], [13]. Those schemes, however, require either accurate synchronization or a synchronized hydrophone array which is quite difficult to be implemented on commercial phones. Dead reckoning based approaches, *e.g.* [2], suffer from accumulated errors. Most of the exiting indoor tracking solutions need a pre-knowledge or at least three anchors.

978-1-4799-3360-0/14/$31.00 © 2014 IEEE

There are many challenges in achieving high accurate multi-user tracking due to the highly dynamic and continuously evolving movement pattern of mobile users. Acoustic-based ranging can be used to obtain the frame snapshot of multi-user formation. With commercial phones, the accurate acquisition of audio tones is difficult due to the attenuation, distortion, interference, and multi-path effect. Besides, for multiple dynamic users, the required small ranging delay and the narrow available acoustic band make the multi-user ranging even more difficult. As the detectable distance by the audio tone is limited, the ranging results of some frame snapshots may be ambiguous, leaving some users still nonlocalizable. Even when ranging results can produce snapshots of team formation, the continuous movements of individuals are hard to obtain without anchor nodes. We need accurate information about the moving distance and moving direction of users to combine these scattered frames to achieve continuous tracking. The movement continuity may also help to remove ambiguities from each frame. Previous schemes estimate the moving distance and direction by dead-reckoning [18]. But special devices or pre-knowledge are usually required, *e.g.*, [17], [26], and absolute positions are also require to fix the accumulated errors.

To address above issues, we propose **Montage**, to track the realtime formation and movements of multiple users. This design uses the coded acoustic signal for simultaneous multi-user ranging and inertial sensors for accurate moving distance/direction estimation. Combining the ranging results and moving estimations, **Montage** provides meter-second-level formation tracking with off-the-shelf mobile phones and requires no pre-knowledge or synchronization services. It achieves accurate localization using merely **one** anchor node. The contributions of this work are as follows:

- We design coded audio tones with which the instantaneous distances among multiple mobile users are accurately estimated when they generate tones simultaneously, in the presence of high noise, multi-path effect and Doppler Shift.
- We present innovative step stride-length and walking direction estimation methods to achieve a very accurate moving trace estimation without any priori knowledge (such as the stride-length, phone-placement and indoor map).
- We connect successive localization snapshots to refine the range-based localization and generate continuous moving

traces, by leveraging the accurate moving distance and direction estimation. It provides better disambiguation and estimates the real trace of users without anchor nodes.
- We design, develop, and deploy **Montage** in both indoor and outdoor environment to evaluate its performance. 847 traces from 15 volunteers are collected and analyzed. The results show that the estimated stride-lengths over a variety of users have errors within $8.9cm$ and the mean error is $4.3cm$. The estimated moving-direction is within 20^o of the real direction. For real-time single-user indoor tracking, the mean deviation of 847 traces is about 0.87 meter, and 90% deviations are less than 2 meters. For real-time multiuser indoor experiment, the maximum deviation is about $1m$ while the mean deviation is about $0.5m$ using both inertial sensors and acoustic ranging.

The rest of the paper is organized as follows. We present problem formation and baseline method in Section II, and novel multiuser ranging with coded audio tones in Section III. In Section IV we discuss our techniques of accurate estimation of moving distance and direction. Our evaluation results are presented in Section V. We review the related work in Section VI and conclude the paper in Section VII.

II. Our Approach

Assume that there is a group of n mobile users $A = \{a_1, \ldots, a_n\}$ in proximity. At time t, the *location* of user a_i at earth coordinate is $P_i^e(t) = (x_i^e(t), y_i^e(t))$. If we record the location of user a_i according to the time vector $T = \{t_0, t_1, \cdots, t_M\}$, the moving trace of a_i can be represented by a sequence of locations $\{P_i^e(t_0), P_i^e(t_1), \cdots, P_i^e(t_M)\}$. For simplicity of presentation, besides the earth coordinate system, we introduce the *translation coordinate* system in which each location has a constant offset from that of earth coordinate, *i.e.*, the origin of a *translation coordinate* system is moved but the directions of both axes remain the same. For example, let $P_1^e(t_0) = (x_1^e(t_0), y_1^e(t_0))$ be the origin of a *translation coordinate* system, noted as $P_1(t_0) = (0, 0)$. If the position of a_i at the translation coordinate is $P_i(t_0) = (x_i(t_0), y_i(t_0))$, then its earth location is $P_i^e(t_0) = P_i(t_0) + P_1^e(t_0) = (x_1^e(t_0) + x_i(t_0), y_1^e(t_0) + y_i(t_0))$.

A. Main Idea

Our goal is to design a scheme for precise mobile user tracking without pre-deployed infrastructures. Our scheme exploits coded acoustic signals to simultaneously measure the distances among users. The ranging results expose multi-users' distances at a certain timestamp and thus indicate a logical topology of the network. The logical structure, normally, lacks orientation information and may not be rigid [25]. The second component is the *movement vectors* detection which leverages information from various sensors on the smartphones. The *movement vectors* connect locations of the same user at consecutive timestamps. With the ranging results and *movement vectors*, **Montage** dynamically calculates the *distance vectors* to measure the Euclidean distance between different users. Using these vectors, we can easily reassemble the real topology and continuously track users' movement traces. In **Montage**, the localization and tracking are at a translation coordinate system in the absence of anchor nodes. As the translation coordinate has a fixed offset from the earth coordinate, given an arbitrary anchor point $P_i^e(t_j)$, our approach can determine the traces and locations at the earth coordinate.

B. Baseline Approach for Localization

User a_i moves from location $P_i^e(t_u)$ to $P_i^e(t_v)$ during period t_u to t_v. The *movement vector* is

$$M_i(t_u, t_v) = P_i^e(t_v) - P_i^e(t_u) = P_i(t_v) - P_i(t_u),$$

which is independent of the coordinate system and only determined by its magnitude/distance d and orientation θ. The *movement vector* can also be represented by a two-tuple $(r_i^{uv}, \theta_i^{uv})$ in the polar coordinate system. Then, the trace of a single user a_i can be recorded by a sequence of *movement vectors* $\{M_i(t_0, t_1), M_i(t_1, t_2), \cdots\}$. At the time t_u, the *distance vector* between user a_i and a_j is defined as

$$R_{ij}(t_u) = P_i^e(t_u) - P_j^e(t_u) = P_i(t_u) - P_j(t_u).$$

The magnitude of the *distance vector* can be measured by the ranging result between a_i and a_j, say $r_{ij}(t_u)$. As shown in Fig. 1(a), $M_i(t_0, t_1)$ and $M_j(t_0, t_1)$ are movement vectors of user a_i and a_j. $R_{ij}(t_0)$ and $R_{ij}(t_1)$ are distance vectors at time t_0 and t_1.

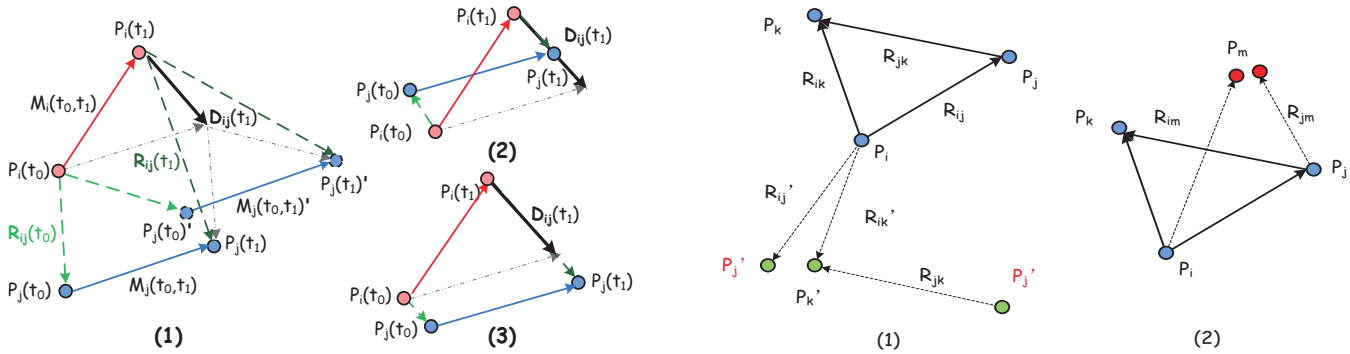
Given ranging results, which are the magnitude of distance vectors, only a formation of user locations can be derived at a time if the topology is rigid. The orientation of the formation is uncertain, thus we cannot derive the traces of users' movement from consecutive formations. The output of the trace detection scheme is represented as a sequence of movement vectors for each single user, but the locations of points in the trace are undetermined. We propose to combine the ranging results and movement vectors to localize all users at each sample time and to acquire continuous user traces. Our approach is based on the observation that the distance vectors and the movement vectors meet the following equation:

$$\begin{aligned} D_{ij}(t_{u+1}) &= R_{ij}(t_{u+1}) - R_{ij}(t_u) \\ &= M_j(t_u, t_{u+1}) - M_i(t_u, t_{u+1}). \end{aligned} \tag{1}$$

Here $D_{ij}(t_{u+1})$ is defined as the *difference vector*. Let the two-tuple of the difference vector $D_{ij}(t)$ be $(d_{ij}(t), \theta_{ij}(t))$. As illustrated by Fig. 1(a), $D_{ij}(t_1)$ is the difference vector. When the movement vectors $M_i(t_0, t_1)$ and $M_j(t_0, t_1)$ are known, $D_{ij}(t_1)$ is determined. And, we have

$$\begin{cases} r_{ij}(t_1)\cos\theta_{ij}(t_1) - r_{ij}(t_0)\cos\theta_{ij}(t_0) = d_{ij}(t_1)\cos\theta_{ij}(t_1) \\ r_{ij}(t_1)\sin\theta_{ij}(t_1) - r_{ij}(t_0)\sin\theta_{ij}(t_0) = d_{ij}(t_1)\sin\theta_{ij}(t_1) \end{cases} \tag{2}$$

Given the ranging results $r_{ij}(t_0)$ and $r_{ij}(t_1)$, each solution for $\theta_{ij}(t_0)$ and $\theta_{ij}(t_1)$ determines a possible assignment of a_i and a_j's positions at time t_0 and t_1. When $r_{ij}(t_1) + r_{ij}(t_0) > d_{ij}(t_1)$ and $r_{ij}(t_1) - r_{ij}(t_0) < d_{ij}(t_1)$, there exist two solutions. As illustrated in Fig. 1(a), both the position groups $\{P_j(t_0), P_j(t_1)\}$ and $\{P_j(t_0)', P_j(t_1)'\}$ satisfy the constrains of distance vectors and movement vectors. When $r_{ij}(t_1) + r_{ij}(t_0) = d_{ij}(t_1)$ or $r_{ij}(t_1) - r_{ij}(t_0) = d_{ij}(t_1)$, there is only one solution, as shown in Fig. 1(a). There exists a special case that the movement vector $M_i(t_0, t_1)$ of user a_i and $M_j(t_0, t_1)$ of user a_j are equal, *i.e.* they move in the same direction at the same speed. In this case, $r_{ij}(t_1) = r_{ij}(t_0)$ and there are infinite groups of solutions.



(a) Baseline method for Montage

(b) Localize users in sequel

Fig. 1. Baseline team formation tracking based on movement vector and ranging results.

Based on the above calculation, each distance vector may have one, two or infinite possible solutions. For the first case, the distance vector is determined. For the third case, we cannot decide the value of distance vector and require further information. The most common situation is that there are two possible values for the distance vector with the same magnitude while different orientations. In this case, we leverage the neighboring information to eliminate the ambiguity. In the above example, assume that user a_i and a_j both have ranging results to a third user a_k, then we can get the two possible solutions of R_{ik} and R_{jk} as well. Clearly, locations of the user a_i, a_j and a_k form a triangle (called *ranging triangle*), and thus theoretically the value of three distance vectors must meet the following equation.

$$R_{ij} + R_{jk} - R_{ik} = 0 \tag{3}$$

As each vector has two potential solutions, there are 8 combinations in all. For example in Fig. 1(b)(1), the distance vectors in solid lines meet the equation constraint and the combination in dashed lines is a wrong answer because it leads to two ambiguous locations of user a_j. In practice, we select the combination which minimizes the absolute value of Equation (3).

C. Vector Based Multi-User Tracking

We will further discuss the full-featured user tracking approach. In the first step, we select an arbitrary ranging triangle and determine the three distance vectors (edges) of this triangle using the algorithm discussed in previous subsection. Here we prefer to select the start triangle whose vertices have more ranging neighbors. Then we put all three users in this triangle into a set denoted as *localized set* which keeps all the distance vectors as well.

In the second step, we iteratively add more users to the localized set by determining distance vectors from the new user to neighboring users in the localized set. As illustrated in Fig. 1(b)(2), user a_m has ranging results with a_i and a_j. According to the aforementioned baseline algorithm, we get one or two possible solutions for each of R_{im} and R_{jm}. We simply drop the results of zero solution or infinite solutions, because the distance vector cannot be determined according to them. For the two-solution case, based on the observation that R_{im}, R_{jm} and R_{ij} form a triangle, and theoretically we have $R_{ij} + R_{jm} - R_{im} = 0$. To address the ranging errors, we will select the pair of R_{im}, R_{jm} values that minimizes $R_{ij} + R_{jm} - R_{im}$. After that, the distance vectors from two users in localized set to a_m have been determined. We put a_m into the localized set and keep both distance vectors. We calculate the distance vectors from a pair of neighboring users instead of separated ones to a new user, for the purpose of avoiding cascading errors. The above process iterates until no new user can be added. These vectors corresponding to users in the localized set specify the relative locations of users and if we assign location (*e.g.* at earth or translation coordinate system) for any one of them, all the other users can be located at the specified coordinate system.

Now we have localized all users (obtain distance vectors and rebuild the topology) at time t_0 and t_1, in the coming timestamp t_2, the localization process can be significantly simplified. Later in Section II-D we will show how to calculate the distance vector based on Eq. (1). With knowing the value of the distance vector in prior timestamp, $R_{ij}(t_1)$ in the example of Fig. 1(a), the distance vector $R_{ij}(t_2)$ can be directly calculated using $R_{ij}(t_1)$ + $D_{ij}(t_2)$. With this method, Montage conducts localization in consecutive timestamps and rebuilds topology snapshots over time.

After rebuilding the topology at translation coordinate for each timestamp, we connect these topology snapshots and form integrated user movement traces. As the movement vectors connect locations among continuous timestamps, Montage leverages them to connect consecutive topology snapshots and locate users continuously at the same coordinate system to provide movement traces. Here we select an arbitrary user a_i and set its position at time t_0 as origin, then all other users' locations at time t_0 can be determined. At time t_1, we calculate the new position of a_i using its movement vector. We can get different positions of a_i through applying different users' movement vectors to connect the two topology snapshots. In order to avoid the impact of measurement error in single movement vector, we use the mean value as the new position of a_i. Then the topology can be determined at the same coordinate system as t_0. This process iterates until the movement traces of all users are determined.

D. Design Issues

Three key issues need to be discussed in this approach. First, the order of adding new user into the localized set

can impact the overall performance. In this work, we apply a width-first approach to alleviate the accumulating errors. During each iteration, we firstly select all users that have two ranging neighbors in current localized set. After determine the distance vectors for all these users, we add them to the localized set and thus update the localized set.

Second, the selected distance vectors can deviate from the real value due to the measurement error, and thus lead to ambiguous locations for a user, for example, user a_m in Fig. 1(b)(2). To address this problem, we introduce an integrated optimization algorithm to achieve a globally consistent result. As the acoustic ranging is relatively accurate, we focus on fine-tuning the orientation of distance vectors, which is formalized as an optimization problem with constraints.

$$\min \sum_{i,j=1,i\neq j}^{n} (\delta\theta_{i,j}^2), \textbf{s.t.},$$
$$R_{ij}(r_{ij}, \theta_{ij} + \delta\theta_{i,j}) + R_{jk}(r_{jk}, \theta_{jk} + \delta\theta_{j,k})$$
$$= R_{ik}(r_{ik}, \theta_{ik} + \delta\theta_{i,k}), \forall R_{ij}, R_{ik}, R_{jk} \text{ in a ranging triangle.}$$

Here $R_{ij}(r_{ij}, \theta_{ij}+\delta\theta_{i,j})$ denotes the vector R_{ij} with magnitude r_{ij} and direction $\theta_{ij} + \delta\theta_{i,j}$. In the above optimization, r_{ij} is a known value computed from acoustic ranging, θ_{ij} is a known value computed from Eq. (2) in Subsection II-B. $\delta\theta_{i,j}$ is a variable to be computed. According to the optimization results, we rotate each distance vector with angle $\delta\theta$ and finally get a consistent localization result.

Third, we consider the situation that the new user only has one ranging neighbor in the localized set. A special scenario for this case is that there are only two users in the network and we want to determine the relative position between them. In the case of single ranging neighbor, we present a multi-stage scheme using the temporal correlation among candidate locations to eliminate ambiguities. Assume that at time t_1, we get two movement vectors $M_i(t_0, t_1)$ and $M_j(t_0, t_1)$ of user a_i and a_j, we can calculate two possible solutions of $R_{ij}(t_0)$. Then at timestamp t_2 the users report $M_i(t_1, t_2)$ and $M_j(t_1, t_2)$. We have $M_i(t_0, t_1) + M_i(t_1, t_2) + R_{ij}(t_2) = M_j(t_0, t_1) + M_j(t_1, t_2) + R_{ij}(t_0)$.

If we put in the values of movement vectors and two candidate values of $R_{ij}(t_0)$, we get two solutions of $R_{ij}(t_2)$. As we have the ranging results of $R_{ij}(t_2)$, we can distinguish these two solutions and determine the right answer. In most cases, the ranging metric works well and can successfully find the right solution. However, two candidate solutions of $R_{ij}(t_2)$ may have the same length which cannot be distinguished. To address this issue, we wait for some period and use new movement vectors. Then users with at least one ranging neighbor in the localized set can be included and the final localized set contains all users that have at least one ranging path to the initial triangle.

III. Multi-user Ranging by Coded Audio Tones

The distance vectors among users provide information to determine their locations in translation coordinates. When users are all dynamic, it is difficult to estimate the orientation of a distance vector at the earth coordinate. In our multi-user tracking approach, as presented in the previous section, only the magnitude of the distance vectors are required for multi-user localization and tracking. There are some exiting works dedicating to acoustic signal based accurate ranging between a pair of mobile phones, *e.g.* the ETOA protocol [15]. But it is still a challenging problem to measure the distances among *multiple mobile* users. As users walking at a speed about $2m/s$, *i.e.* a round of multi-user ranging must be completed within a short period to capture the simultaneous locations of multiple users at a high sampling rate. A Frequency Division Multiplexing (FDM) seems a good solution to improve the delay for multi-user ranging. The detectable frequency range of most commercial mobile phones is 0 to 22kHz. The audio signal with frequency below 15kHz is audible to people and the frequency above 20kHz suffers a severe distortion and attenuation, which leaves us a usable frequency range 15kHz to 20kHz. When users are moving, the Doppler shift must be taken into consideration. For example users are walking at a speed $1.5m/s$, and the emitted signal is 19kHz, at least 350Hz gap between two consecutive frequencies is required to avoid the interference. Thus, there remain very limited usable channels. Besides, a simple audio tone cannot resist environment noises, *e.g.* the honk of a car. To address this issue, we propose a method using coded audio tones to range multiple users simultaneously.

A. Coded Audio Tones

In our scheme, to separate different users, a set of codes are used to encode the audio tones. A code is a binary sequence $\mathcal{C} = \{C(0), C(1), C(2), \ldots, C(N-1)\}$, with N chips $C(k)$. These chips can have 2 values -1/1 (polar), *i.e.*'0'/'1'(logical). each user a_i owns a code C_i of the same length, then modulates a carrier at frequency ω_c with his/her code. The transmitted audio tone of user a_i is

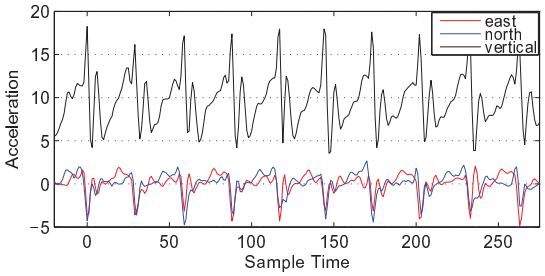
$$T^i(t) = C_i(t) \cdot \cos(\omega_c t). \quad (4)$$

1) Code Selection: Code selection has a large impact on the performance of multi-user ranging. The coding method of the audio tone should be deterministic to make sure every user is able to independently generates the same code book. The cross-correlation and out-of-phase auto-correlation must be low enough to resist interference among multiple users and self-interference due to multi-path propagation. Besides, the code must have a proper period: long enough to discriminate a large number of users, but short enough for small delay.

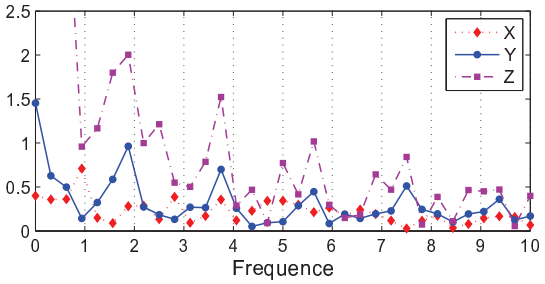
We leverage pseudo-noise (PN) codes [21] to design our multi-user ranging approach. There are three typical PN codes: maximal length sequence (m-sequence), Gold codes and Kasami codes. All these sequences have the maximum possible period $N = 2^r - 1$. We choose Gold code as it provides us a good tradeoff among auto-correlation and aperiodic correlations.

2) Coded Tones Generation: Before the tracking starts, we can detect the background noise of the current environment and select the most clean frequency between 15kHz and 20kHz as f_c, via simple spectral analysis. Our extensive sampling tests show that, frequency space between 15kHz and 20kHz has less noise even in a very loud environment. Then we have $\omega_c = 2\pi f_c$.

Then we need choose the parameter r to generate a set of Gold codes. r determined the period $2^r - 1$ of the codes

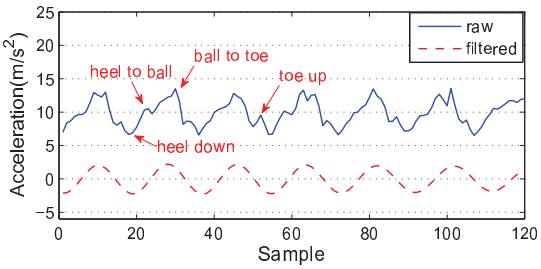


(a) Vibration

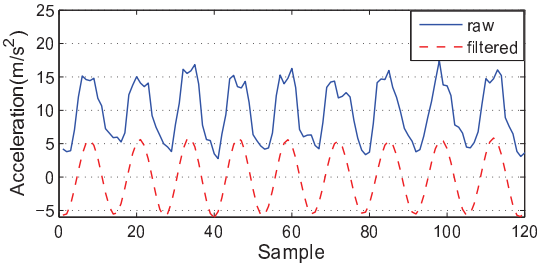


(b) Waggling

Fig. 2. a) the raw data of acceleration of a walking user; b) the FFT of accelerations along the walking orientation (Y), perpendicular (X) and to the sky (Z).



(a) The raw vertical acceleration and filtered vertical acceleration while user is walking at a normal pace.



(b) The raw vertical acceleration and filtered vertical acceleration while user is walking at a fast pace.

Fig. 3. The raw vertical acceleration and filtered vertical acceleration when a user walks at different paces.

and the size $2^r + 1$ ($r \not\equiv 0 \mod 4$) of the code set. The supported sample rate of most commercial mobile phones is 44100Hz. When each chip is s samples long, the length of the audio tone is $\frac{s}{44100}(2^r - 1)$ seconds. On one side, the longer the period, the greater the delay; on the other side, tracking n users requires $2^r - 1 > n$. Considering both the delay and user number requirements, a proper r and s can be determined. For example, when $n = 20$, then the selection $r = 5$ and $s = 100$ will produce 72 ms audio tones.

After the set of Gold codes and the length of a chip are determined, each user a_i is assigned a unique code from the set and generate his/her own tones according to Equation (4). As soon as received a ranging command via a radio channel, each user emits his/her coded tone. For a continuous tracking task with a update interval δt, each user emits his/her coded tone periodically for every δt after the first emission. For different applications, δt varies from tens of milliseconds seconds to tens of seconds.

3) Coded Tones Acquisition: For an emitted tone $T^i(t)$, the received signal $R(t)$ comprises $T^i(t)$, the interfering tones $I(t)$ and white noise $n(t)$. Then we have $R(t) = T^i(t) + I(t) + n(t)$. When the receiver captures a sequence of acoustic signal, he/she uses a narrow frequency bandpass filter to clean most of the background noise and get $T'(t)$. For example, in a walking scenario, the passband could be $[f_c - 500, f_c + 500]$. To recover the code stream, the receiver multiplies $T'(t)$ by the reference carrier $\cos(\omega_c t)$. Then $T'(t) \cdot \cos(\omega_c t) = 0.5C_i(t) + 0.5C_i(t) \cdot \cos(2\omega_c t) + (I(t) + n(t)) \cdot \cos(\omega_c t)$. After the multiplication, a lowpass filter is used to remove the ω_c and $2\omega_c$ component and get $C'(t)$.

If multiple users emit tones simultaneously, $C'(t)$ is the sum of all their codes. To acquire the code of user a_i, a sliding window, whose size is $\frac{s}{44100}(2^r - 1)$, is used to detect the peak of the correlation between $C_i(t)$ and the $C'(t)$ in the window. When a peak exceeding a threshold is detected, the start sample of the current window will be stamped as the arrival time of a_i's tone. Then collecting the time line of all participants, the range between each pair of user can be calculated according to [15].

IV. Movement Vector Detection

Movement vector is the key to connect successive localization snapshots to achieve disambiguated multi-user tracking. There are two major categories of methods for determining the user's movement vector with a commercial smart phone. One category uses the integration of horizontal acceleration, which is impractical due to the large error caused by double integration of sensor drift and noise. The other category detect the steps of user by pattern recognition and use the multiplication of step number and average step length to estimate the distance. Those methods require some user measurements and inputs in advance, which can hardly adapt to different users and different paces of the same user. Without the help of GPS, there are no effective accurate methods to detect the moving orientation of a user with the off-the-shelf smart phone, *e.g.*, the error spans about 60° in [17].

A. Understanding the Acceleration

Considering a user could hold the phone in any position, we convert the realtime acceleration from the phone coordinate system to the earth coordinate system,*i.e.* north, east, gravity. In this work, we use the earth coordinate system as an inertial coordinate system for localization.

To understand the cause of the error of existing distance and orientation estimation approaches, we analyze the accelerometer data from a commercial smart phone. We observed

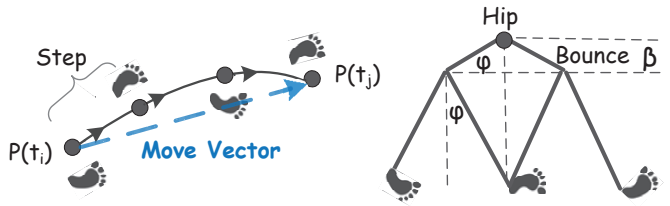


Fig. 4. Movement vector and walk model.

the following phenomena. Even when the phone is static, there exists huge drifts of acceleration at three orientations, which cause more than 10 cm displacement within 10 seconds by double integration. The drift is much severer when the phone is in a mobile status, exceeding a meter in 10 seconds. As shown in Fig. 2(a), the various springs of acceleration of walking are caused by diverse walking habits of different persons, or changing paces of the same person, or different positions and attitudes of the phone. A very important cause is that the acceleration of walking is not only caused by moving forwards, but also by waggling left and right as well as the vertical movement. Fig. 2(b) presents the spectrum distribution of walking accelerations at three orientations. It shows that there is a great energy from the movement perpendicular to the walking orientation, whose frequency is half of the walking frequency. The perpendicular component could result in great error of the integration and the misunderstanding of the moving orientation.

These observations inspire us to design a method achieving a good movement vector estimation we need first extract the pure acceleration caused by walking from raw acceleration values. In our system, we filter the acceleration using a bandpass filter with a narrow window of the walking frequency, $pb = [\frac{3f_w}{4}, \frac{3f_w}{2}]$, where f_w is the walking frequency. With a simple step detection, given the sample rate of the accelerometer, the current walking frequency f_w can be determined by counting the sample number of the current step. The filtering eliminates the high-frequency noise from the vibration of the phone and the low-frequency noise from the left and right waggling. Fig. 3 presents example raw vertical acceleration data and filtered acceleration data when the user moves at different paces. As we can see, the filter also removes the large zero-frequency component, *i.e.* gravity component.

B. Magnitude of Movement Vector

To estimate the moving distance, we combine dead reckoning and the stride length based approach. The challenges come from the changing stride length of different people at different paces. We propose an adaptive stride length estimation method, which requires no user input and no knowledge from digitalized map. Combining the accurate step detection and the stride length estimation, the moving distance is obtained automatically.

Given one step, our adaptive stride length estimation is based on two principles:

1) As shown in Fig. 4, the vertical bounce β (*i.e.*, the maximum vertical displacement of user's hip in one step walking) of a walking person is directly correlated to his/her stride length through an almost equal angle ϕ. Here ϕ is half of the angle between two legs when both feet touch the floor during walking. When a person walking at a constant pace, the angel is constant. So we can estimate the stride length by $2\cot\phi\beta$. Here the bounce β can be computed from double integration of the vertical acceleration $\mathbf{a} - avg$, where $\mathbf{a}$ is current vertical acceleration and avg is the historical average vertical acceleration of this user.

2) For the same person at greater paces, the angle increases. From Fig. 3, we notice that when the pace increases, the ratio $\frac{max-min}{avg-min}$ of the acceleration raw data increases with the stride length. Here max and min is the historical maximum and minimum acceleration data of this user. The spring pattern of the raw acceleration also changes the ratio, as presented in Fig. 3(a), which reflects the difference of different person's step.

Assume that there are T acceleration samples $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_T\}$ within a step. Combining these two principles, we adaptively estimate the moving distance d as $d = k\sqrt{\frac{max-min}{avg-min}\sum_{j=1}^{T}\sum_{t=1}^{j}(\mathbf{a}_t - avg)}$. Here the parameter k is a constant for the same person. In our approach, an initial value of k is given according to the average value of people. Then, according to the online localization with the ranging result, k is calibrated for the first several rounds and fixed for each user respectively.

C. Orientation of Movement Vector

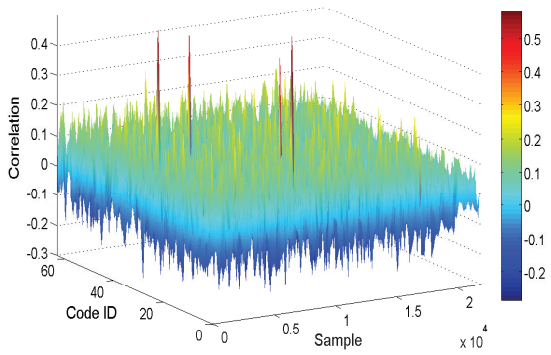
We use the filtered horizontal accelerations along the east and north axes at the earth coordinate system to estimate the orientation of each step. Assume that there are T acceleration samples within a step, the horizontal accelerations within a step are $\mathbf{a}^H = \{\mathbf{a}_1^H, \mathbf{a}_2^H, \cdots, \mathbf{a}_T^H\}$, each $\mathbf{a}_i^H = \sqrt{\mathbf{a}_i^{E\,2} + \mathbf{a}_i^{N\,2}}$. Here $\mathbf{a}_i^E$ is the east component of the i-th acceleration sample, and $\mathbf{a}_i^N$ is the corresponding north component. The maximum horizontal acceleration $\max\{\mathbf{a}_1^H, \mathbf{a}_2^H, \cdots, \mathbf{a}_T^H\}$, is detected for each step, let its index be κ. The orientation of $\mathbf{a}_\kappa^H$ is closest to the moving orientation of this step. As mentioned in [17], even knowing the moving orientation by $\arctan(\mathbf{a}_\kappa^E/\mathbf{a}_\kappa^N)$, it is still difficult to determine the forward and backward orientation. To address this issue we notice that the forward acceleration companies the rising edge of the vertical acceleration. With our approach, the orientation of each step can be determined within 20° error range.

V. Analysis and Evaluation

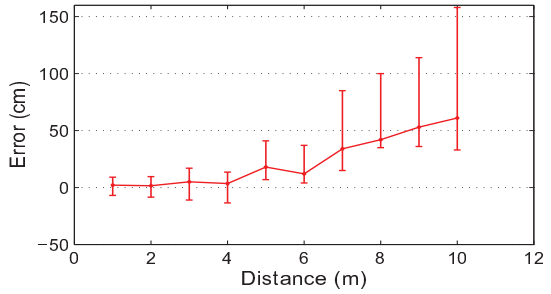
We implement **Montage** on Android phones and examine the performance with extensive experiments in this section.

A. Coded Tone Based Ranging

For the coded tone based multi-user ranging, the delay mainly consists of three parts: the time for tone emission, the time for tone transmission, and the time of coded tone acquisition. The transmission time is decided by the distance, which is usually tens of milliseconds for indoor application. The emission time is determined by the the length of the audio tone, which is $\frac{s}{44100}(2^r - 1)$. In the experiments, we select the set of Gold codes with $r = 7$ as the codebook, 19 kHz as the carrier frequency, and the chip length is 40 samples. As a



(a) Tone acquisition of 4 users in a round of ranging.



(b) Ranging error when there are 4 users.

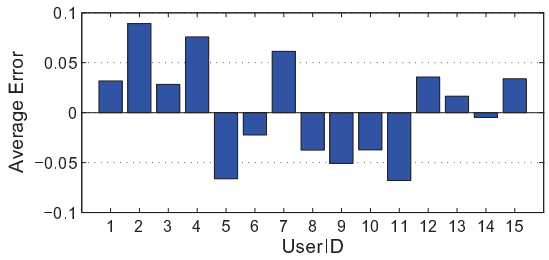
Fig. 5. Coded tone based 4 users ranging.

result, the length of a coded tone is 115 ms, so is the sliding detection window. The step of the sliding window is 4 samples. We test the ranging performance with 4 users. Each user selects a unique code from the set. To exam the interference-resistance property, we design the experiment that will result in larger interference by dividing 4 users into two groups and changing the distance between groups. All users emit their tones as soon as they received a start signal through Wi-Fi. The arrival time of each coded tone is detected by sliding its code to locate the maximum correlation peak. Fig. 5(a) shows the coded tone acquisition result by one of the users in a round of ranging. And Fig. 5(b) presents the ranging results in the hall of an office building. And the delay is less than 200 ms. The result shows that, our coded tone based ranging method achieves sub-meter accuracy when users are about 10 meters apart.

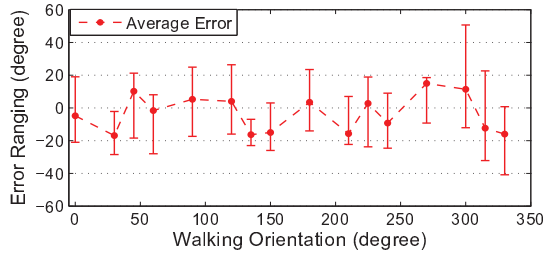
B. Movement Vector Determine

We examine the stride length estimation accuracy of our system. There are 15 participants in the experiments, including 4 female and 11 male persons. Their heights vary from 1.56m to 1.82m and their average stride lengths vary from 53cm to 83cm. Each participant carries the phone arbitrarily and walks at arbitrary paces. Fig. 6(a) shows the average error of the the estimated stride length for each person. The maximum error is 9cm, and the mean error is 4cm.

We also examine the accuracy of the forward orientation estimation. Fig. 6(b) shows the error of the estimated orientations while the walking orientation changes from $-180°$ to $180°$. The mean error of detected orientation by our methods is $\pm 10°$, with 90% errors are within $\pm 20°$, which greatly outperforms the existing orientation estimation work.

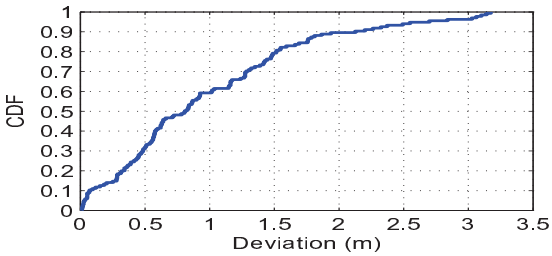


(a) The average error of estimated stride length of 15 users.

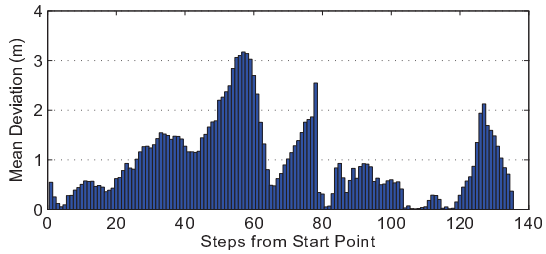


(b) Error range of estimated orientation when a user is walking.

Fig. 6. Estimation of movement vector.



(a) CDF of the deviation of 847 indoor traces.

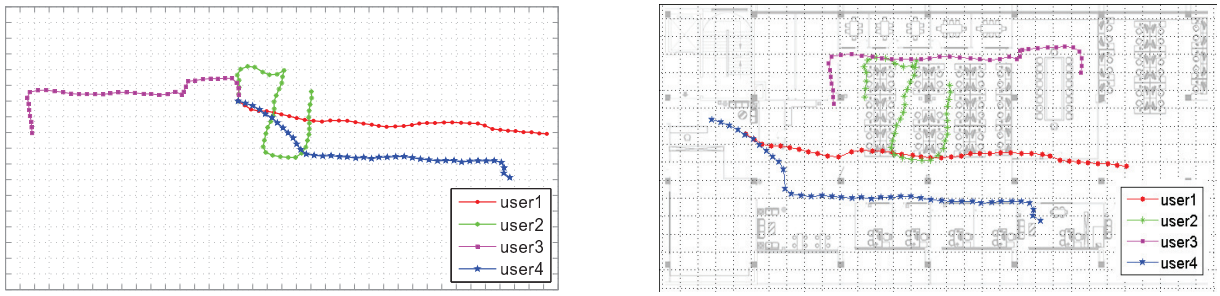


(b) Mean deviation according to the steps from the start points.

Fig. 7. Single users's tracking result by movement vectors.

C. Single User Tracking

With the real-time magnitude and orientation estimated by our approach, a single user's trace can be tracked by a series of movement vectors. To get robust evaluation results of movement-vector-based single user tracking, we have 15 volunteers (4 female and 11 male) installed **Montage** in their smart phones to collect traces. Since there is no GPS signal indoor, to get the ground truth we mark 25 optional traces with diverse lengths, directions and shapes on the floor of our office, which covers 1600 square meter area. Volunteers can walk along any combinations of these traces with free paces and arbitrary phone positions. 847 traces from 15 volunteers are collected. For every step, there is a tracking location, and about 32,000 locations in total. We analyze the deviation of each tracking location, and Fig. 7(a) presents the CDF of the



(a) Fragments of 4 users' movement vectors.

(b) A fragment of indoor tracking results of 4 users.

Fig. 8. A fragment of four users's tracking result.

deviation. The result shows that, the mean deviation is about 0.87 meter, with 90% tacking location have a deviation less than 2 meters. We also explore the deviation change with the distance to the start point, Fig. 7(b) shows that within the initial 20 steps (about 16 meters), the deviation won't exceed 0.5 meter. The deviation increases with the distance to start point and won't exceed 2 meters for 90% time within 140 steps (about 110 meters). But we notice that, a small portion of large deviations (about 3 meters) occur around 60 steps, and we consider the reason as the scale of our office makes most turns happen between 50 and 70 steps.

Then extensive indoor tracking results show that, with only inertial sensors of an off-the-shelf mobile phone, our method can achieve a highly accurate tracking result of walking people. To compare with the state-of-art methods in [2], which achieve a tracking error of 6.9%, **Montage** achieves a tracking error of about 2.5%.

D. Multi-user Tracking

Combing movement vectors and ranging results, we can track the team formation and movement of multiple users. In our experiments, 4 users walk randomly in a the 1600 square meter office. Their movement vectors are detected in real time and distances between every pair of users are calculated periodically. Each time their ranges are obtained, our localization approach introduced in Section II is applied to calculate their locations at a translation coordinate, which takes the initial location of a randomly chosen user as the origin, and calibrate the estimated movement vectors accordingly. Fig. 8 illustrates a fragment of 4 users' team formation tracking. As shown in Fig. 8(a), with estimated movement vectors, the traces of each user can be obtained. However, without anchor nodes we cannot know the team formation of 4 users. Combing the ranging results and the movement vectors, the locations of 4 users are determined. Fig. 8(b) presents the detected team formation with three rounds of ranging results. When the No.4 user knows the location of his start point (the entrance of the office), the other three users' absolute locations are determined as illustrated on the floor plan, which matches the ground truth surprisedly well. In our experiments, in which each user walked for about $1000m$ in the office, the mean deviation of the estimated trace to the ground truth is about $0.5m$ and the largest deviation is about $1m$. With the help of ranging, **Montage** enables formation detection and improves tracking accuracy. With only one anchor position, **Montage** enables accurate indoor localization for multiple users.

VI. Related Work

One popular line of mobile handset indoor localization is fingerprinting. Some systems exploit fingerprints of wireless signals to achieve room-level user localization and tracking, *e.g.*, [6], [20], [27]. Horus [27] designs a WLAN localization system with a meter-level accuracy. [20] presents a GSM indoor localization system that achieves a median accuracy of 4 m. EZ [6] uses the RSSI to indoor APs and yields a median accuracy of 2-7 m with no pre-deployment effort. There are other types of fingerprints or landmarks used to achieve room-level localization, *e.g.*, [7], [19], [22]. Batphone [19] uses an ambient sound fingerprint called the Acoustic Background Spectrum (ABS), and [7] uses Geo-magnetism as fingerprint. Unloc [22] uses identifiable location signatures on one or more sensing dimensions. Most fingerprinting based localization methods cost an effort for site-survey. Some recent systems have incorporated survey by users,*e.g.* [23], [26]. But they still face the problem that different locations may have similar fingerprints. There are also some works using wireless signal to localize people in a dynamic way, *e.g.*, [24]. But it is difficult to track multiple persons simultaneously.

Some schemes perform localization by estimating distances to anchor nodes based on RSSI, time-of-arrival (TOA), time-difference-of-arrival (TDOA) and angle-of-arrival AoA. Peng *et al.* [15] proposed ETOA with centimeter-level accuracy acoustic-based pair-wise ranging method. [16] presents a solution for achieving high speed 3D continuous pair-wise localization using two microphones and one speaker on the phone. Liu *et al.* [12] uses acoustic ranging estimates among peer phones as constraints to reduce the significant errors of WiFi-based method. [9] leverages Doppler Effect of acoustic signal to achieve centimeter-level accuracy. Most of the acoustic based ranging approaches are designed for a pair of users. Some work [15] uses a TDMA scheme for multi-users ranging, that results long delay and lack of identification when tracking multiple users. [3] proposes a FDMA based solution to estimate the number of mobile devices present in an area, however when users are moving, the FDMA methods may fail due to the Doppler effect. Many work,like [13], propose CDMA based systems using a high frequency acoustic signal and a hydrophone array to enable simultaneous sub-meter tracking of multiple targets. [1] proposes an ultrasonic multicode despreader allowing simultaneous acoustic ranging in real time by embedded sensors. These methods require synchronization or hydrophone array which is quite difficult to implemented on the off-the-shelf mobile phones. Besides,

anchor nodes are necessary for positioning too.

Several inertial navigation approaches are proposed to tracking the move trace of a single user. [8] and [10] provide good survey of inertial positioning systems for pedestrians. Most of them use step-and-heading-based dead-reckoning, *e.g.* [18],with special devices and absolute position fixes are required to correct dead-reckoning output. Some work use the inertial sensors of smartphones with indoor maps to track users as they traverse indoor, *e.g.*, [5], [17]. But it requires a map showing the pathways and barriers and the orientation estimation is quite inaccurate. [2] provides single pedestrian tracking using mobile phones to achieve a tracking error of 6.9%. There are other works dedicated to mobile phone based indoor localization/tracking, *e.g.* Virtual Compass [4], OIL [14]. Most of the exiting indoor tracking methods need a pre-knowledge or at least three anchors, and are infeasible to provide the realtime multi-user formation.

VII. Conclusion

In this paper, we proposed **Montage** for realtime multi-user team formation tracking with no anchor node and provide multi-user localization with merely one anchor node. We designed coded acoustic tones for supporting tracking of multi-users with small latency and designed innovative techniques to accurately estimate the moving distance and directions with off-the-shelf smartphones. No pre-setting or pre-knowledge is required by **Montage**. Our extensive evaluations (847 traces from 15 users) showed that **Montage** achieved meter-second-level accuracy. A future work is to investigate whether Doppler effects will result in better performance for multi-user tracking. as we can estimate the relative distance and direction between two users using Doppler effects caused by mobility.

Acknowledgment

The research is supported in part by NSF China Major Program 61190110, NSF China under Grants No. 61003277, No. 61232018, No. 61272487, NSFC\RGC Joint Research Scheme 61361166009, RFDP 20121018430. The research of Li is partially supported by NSF CNS-0832120, NSF CNS-1035894, NSF ECCS-1247944, NSF ECCS-1343306, NSF China under Grant No. 61170216, No. 61228202. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of author(s) and do not necessarily reflect the views of the funding agencies (NSF, and NSFC).

References

[1] Alloulah, M., and Hazas, M. An efficient cdma core for indoor acoustic position sensing. In *IEEE IPIN* (2010).

[2] Alzantot, Moustafa and Youssef, Moustafa. UPTIME: U-biquitous pedestrian tracking using mobile phones. In *IEEE WCNC* (2012).

[3] Ananda, A. L., and Peh, L.-S. Low cost crowd counting using audio tones. In *ACM Sensys* (2012).

[4] Banerjee, N., Agarwal, S., Bahl, P., Chandra, R., Wolman, A., and Corner, M. Virtual compass: relative positioning to sense mobile social interactions. In *IEEE PerCom* (2010).

[5] Cheng, B., Li, X.-Y., Jung, T., Mao, X., Tao, Y. and Yao, L. SmartLoc: Push the Limit of the Inertial Sensor Based Metropolitan Localization Using Smartphone. In *ACM MobiCom Poster* (2013).

[6] Chintalapudi, K., Padmanabha Iyer, A., and Padmanabhan, V. Indoor localization without the pain. In *ACM MobiCom* (2010).

[7] Chung, J., Donahoe, M., Schmandt, C., Kim, I., Razavai, P., and Wiseman, M. Indoor location sensing using geo-magnetism. In *ACM MobiSys* (2011).

[8] Harle, Robert. A survey of indoor inertial positioning systems for pedestrians. *IEEE Communications Surveys & Tutorials*, (2013).

[9] Huang, W., Xiong, Y., Li, X.-Y., Lin, H., Mao, X., Yang, P. and Liu, Y. Shake and Walk: Acoustic Direction Finding and Fine-grained Indoor Localization Using Smartphones. *IEEE INFOCOM*,(2014).

[10] Jahn, J., Batzer, U., Seitz, J., Patino-Studencka, L. and Gutiérrez Boronat, J. Comparison and evaluation of acceleration based step length estimators for handheld devices. *IEEE IPIN*,(2010).

[11] Jin, Y., Soh, W., and Wong, W. An indoor localization mechanism using active RFID tag. *IEEE SUTC*, (2006).

[12] Liu, H., Gan, Y., Yang, J., Sidhom, S., Wang, Y., Chen, Y., and Ye, F. Push the limit of wifi based localization for smartphones. In *ACM MobiCom* (2012).

[13] Niezgoda, G., Benfield, M., Sisak, M., and Anson, P. Tracking acoustic transmitters by code division multiple access (cdma)-based telemetry. *Hydrobiologia 483*, 1 (2002), pp. 275–286.

[14] Park, J., Charrow, B., Curtis, D., Battat, J., Minkov, E., Hicks, J., Teller, S., and Ledlie, J. Growing an organic indoor location system. In *ACM MobySys* (2010).

[15] Peng, C., Shen, G., Zhang, Y., Li, Y., and Tan, K. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *ACM Sensys* (2007).

[16] Qiu, J., Chu, D., Meng, X., and Moscibroda, T. On the feasibility of real-time phone-to-phone 3d localization. In *ACM SenSys* (2011).

[17] Rai, A., Chintalapudi, K. K., Padmanabhan, V. N., and Sen, R. Zee: Zero-effort crowdsourcing for indoor localization. In *ACM MobiCom* (2012).

[18] Robertson, P., Angermann, M., and Krach, B. Simultaneous localization and mapping for pedestrians using only foot-mounted inertial sensors. In *ACM UbiCom* (2009).

[19] Tarzia, S., Dinda, P., Dick, R., and Memik, G. Indoor localization without infrastructure using the acoustic background spectrum. In *ACM MobiSys* (2011).

[20] Varshavsky, A., de Lara, E., Hightower, J., LaMarca, A., and Otsason, V. Gsm indoor localization. *Pervasive and Mobile Computing 3*, 6 (2007), pp. 698–720.

[21] Viterbi, Andrew J and others CDMA: principles of spread spectrum communication. *Addison-Wesley Reading*, (1992).

[22] Wang, He and Sen, Souvik *etc.*. No Need to War-Drive: Unsupervised Indoor Localization . *ACM MobiSys*, (2012).

[23] Wu, C., Yang, Z., Liu, Y. and Xi, W. WILL: Wireless indoor localization without site survey. *TPDS*, 24 (2013), pp. 839-848.

[24] Xi, W., Zhao, J., Li, X.-Y., Zaho, K., Tang, S., and Liu, X. and Jiang, Z. Electronic Frog Eye: Counting Crowd Using WiFi. *IEEE INFOCOM*, (2014).

[25] Yang, Z., Liu, Y., and Li, X.-Y. Beyond trilateration: On the localizability of wireless ad hoc networks. *IEEE/ACM TON 18*, 6 (2010).

[26] Yang, Z., Wu, C., and Liu, Y. Locating in fingerprint space: wireless indoor localization with little human intervention. In *ACM MobiCom* (2012).

[27] Youssef, M., and Agrawala, A. The horus location determination system. *Wireless Networks 14*, 3 (2008).

[28] Zhao, J., Xi, W., He, Y., Liu, Y., Li, X.-Y., Mo, L., and Yang, Z. Localization of Wireless Sensor Networks in the Wild: Pursuit of Ranging Quality. *IEEE/ACM Netw 21*, 1 (2013).