EV-Loc: Integrating Electronic and Visual Signals for Accurate Localization

Boying Zhang [†*], Jin Teng [†*], Junda Zhu [‡], Xinfeng Li [†], Dong Xuan [†] and Yuan F. Zheng [‡]

[†]Department of Computer Science and Engineering, [‡]Department of Electrical and Computer Engineering
The Ohio State University, Columbus, Ohio, USA, 43210
{zhangboy, tengj, lixinf, xuan}@cse.ohio-state.edu, {zhuj, zheng}@ece.osu.edu

ABSTRACT

Nowadays, an increasing number of objects can be represented by their wireless electronic identifiers. For example, people can be recognized by their phone numbers or their phones' WiFi MAC addresses and products can be identified by their RFID numbers. Localizing objects with electronic identifiers is increasingly important as our lives become increasingly "digitalized". However, traditional wireless localization techniques cannot meet the fast growing needs of accurate and cost efficient localization. Some of these techniques require expensive hardware to achieve high accuracy, which is impractical for massive deployment. Others, such as WiFi RSSI based localization, are inaccurate and not robust to environmental noise. In this paper, we propose a new localization technique called *EV-Loc*. In EV-Loc, we use visual signals to help improve the accuracy of wireless localization. Our technique fully leverages visual signals' high accuracy and electronic signals' pervasiveness. To effectively couple these two signals, we design an E-V match engine to find the correspondence between an object's electronic identifier and its visual appearance. We implement our technique on mobile devices and evaluate it in real-world scenarios. The localization error is less than 1 m. We also evaluate our approach using large scale simulations. The results show that our approach is accurate and robust.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless Communication

General Terms

Algorithm, Design, Experimentation

Keywords

Localization, Wireless Devices, Visual, Matching

1. INTRODUCTION

1.1 Motivation

Mobile devices have proliferated in recent decades. Almost everyone carries one or more such devices, e.g., smartphones and laptops. RFID technology is also developing rapidly. RFID applications like product tags are becoming an integral part of our daily lives. Following this trend, our lives are becoming increasingly "digitalized". We are living in and connecting with a new interactive "web" that involves almost everything around us. It is imaginable in the near future, our intelligent house will respond to our needs automatically. For example, the door will open and the lights will turn on when and only when the house owner approaches and wants to enter. In supermarkets, if we want to find something, a robotic assistant will come up to help localize or even retrieve the product (with an RFID tag) amidst racks of merchandise.

One key enabling technique for the digital life described above is accurate localization of an object, such as a human, a product, a robotic assistant or even a phone. In a digitalized world, every object can be assigned an electronic identifier, e.g., we can represent a person by his phone number or a product by its RFID number. Therefore, we are interested in the following localization problem: given an electronic identifier, how can we localize the object bearing this electronic identifier accurately? Here accuracy is critical. In the above examples, with large localization errors, the back door may be opened when the house owner is at the front door, or the robotic assistant may bring back the wrong product. As we consider the electronic identifier and its bearer as one logical point in the space, we will use 'localizing the electronic identifier' and 'localizing the object with the electronic identifier' interchangeably hereafter.

To localize an object with an electronic identifier, people naturally use wireless technologies to capture and measure wireless signals. Different wireless localization techniques have been proposed in recent years [19]. However, the performance of these localization techniques (shown in Table 1) is still not satisfactory. Some techniques like Cricket [25] or Pinpoint [32] provide accurate localization results, but their accuracy relies on special hardware, which is often costly. These techniques are impractical for civilian use. Other techniques, e.g., Virtual Compass [5], do not rely on special hardware, but their localization results have relatively low accuracy, because these techniques rely on such measurements as signal strength, which are very vulnerable to the noise in the electronic environment.

[*]The two authors are co-primary authors.

MobiHoc'12, June 11–14, 2012, Hilton Head Island, SC, USA.

System	Technologies	Accuracy
Pinpoint [32]	RF TOA	1.3 m
Cricket [25]	TDOA(Ultrasound+RF)	5 cm
RADAR [3]	WiFi RSSI	5.9 m
Horus [31]	WiFi RSSI	2 m
TIX [15]	WiFi RSSI	5.4 m
Virtual Compass [5]	WiFi RSSI	3.2 m

Table 1: Accuracy of representative localization techniques

1.2 Our Contributions

Given the limitations of wireless localization, we consider using visual signal as an auxiliary tool for accurate and practical localization. Compared with electronic signals, visual signals are relatively accurate in localization (less than 1 m) and less affected by noise [30]. The proliferation of camera phones and commercial cameras make visual localization affordable. Therefore, visual signals are a good candidate to help traditional wireless localization.

However, visually localizing an object does not mean knowing what the object is, hence is not enough. For example, we can visually localize a human, but we may not find out who he or she is. It is well known that *recognizing* a human is very hard just based on his or her visual appearance [14]. However, if we know the correspondence (or mapping) between a human's visual appearance and electronic identifier, we can localize the human with the electronic identifier (as well as recognizing him or her) based on the visual localization result and the correspondence. Hence, with the visual localization result, the key problem is to find out such correspondence or mapping. In some cases, it is possible to build up the correspondence library based on *a priori* information. But inputting the precise visual appearances of every object to be localized is cumbersome and undesirable. Moreover, even with the precise visual appearances, as mentioned above, recognizing a very specific human or object among others is still very difficult and time consuming.

In this paper, we propose EV-Loc, a technique integrating electronic and visual signals for accurate localization. In EV-Loc, we assume that each object has an electronic identifier, such as a smartphone's WiFi MAC address, is given. We want to localize an object associated with the electronic identifier. Specifically, EV-Loc takes electronic identifiers as the input and automatically corresponds/maps them with their visual signal counterparts extracted from the images. Then we can leverage visual localization for more accurate localization results. We claim the following contributions:

- We propose a methodology that leverages the accuracy of visual localization to help wireless localization.

- We propose an effective approach for E-V matching, i.e., corresponding an object's electronic identifier with its visual appearance. We also propose a novel method using a distance based location descriptor to model the uncertainty in E-V matching. We derive appropriate thresholds to produce correct matching results.

- We prototype the EV-Loc system and conduct real-world experiments. We can achieve high accuracy for cellular phone localization with median error at ∼0.5 m and 90 percentile error at ∼1 m. We further perform large scale simulations to evaluate EV-Loc's performance. The simulation results show that our approach is efficient and robust.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed localization technique, including cases where electronic and visual signals are indistinct and missing. Section 4 presents our experimental and simulation results. Section 5 discusses practical issues related to EV-Loc. Finally, Section 6 concludes the paper.

2. RELATED WORK

Wireless localization has been an active research area in recent years. Many researchers have conducted extensive work to advance core localization technologies and systems. In this section, we summarize representative work in this area. Generally, these works can be classified into three classes: range-based, range-free, and fingerprinting based.

Range-based localization uses electronic signals to measure the distance or angle between neighbor nodes and performs trilateration/triangulation to estimate an object's position. Based on ways for measuring electronic signals, different localization techniques are proposed. For example, Pinpoint [32] relies on time of arrival (TOA) for localization. Cricket [25] implements time difference of arrival (TDOA) using ultrasound and RF. Virtual Compass [5] uses RSSI (received signal strength indicator) for relative positioning by combining the Bluetooth and WiFi RSSI readings.

Range-free localization does not rely on measurement of distance or angles. Instead, it assumes nodes can estimate distances between each other and the anchor nodes' positions are known. For example, the Centroid algorithm [6] and APIT [16] use area estimation to estimate an object's position. DV-Hop [23] and Amorphous [22] estimate the minimum hop count between the unknown target node and the anchor node as well as the average hop distance, then calculate distance based on these estimations.

Fingerprinting based localization differs from the previous techniques. It fingerprints each location in a scene with a vector of RSSIs from various transmitters, e.g., WiFi APs and GSM towers. Then an object's position is estimated by matching the observed RSSI readings with the closest *a priori* location fingerprints. RADAR [3], Horus [31], and TIX [15] are representative works in this category.

Our work is also closely related to visual tracking and sensor fusion techniques. Yilmaz et al. overview this field in [30]. It has many real applications in a variety of areas such as robotics or bioinformatics. Smith et al. survey the use of multisensor fusion for tracking [27]. SurroundSense [2] uses WiFi RSSI, sound, light, etc. for indoor localization. There is a recent work [28] using combinations of electronic and visual signals for object identification. This work focuses on object filtering over a long period of time in which objects' locations change. Thus, it is not a localization problem.

3. EV-LOC DESIGN

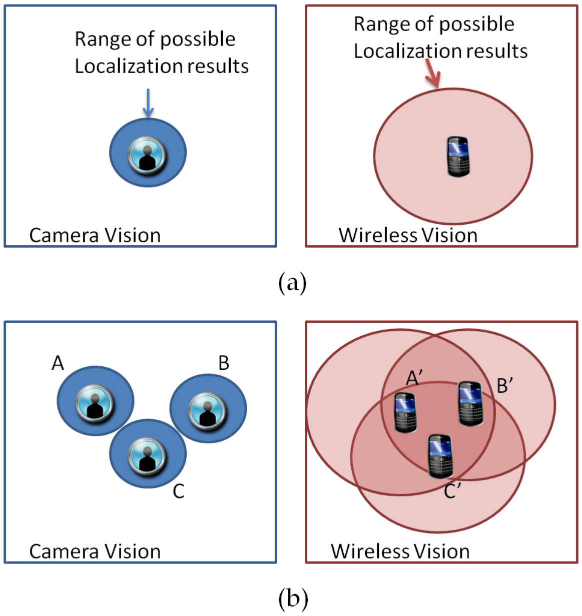
In this section, we present the design of EV-Loc. We start with an overview of EV-Loc, then introduce its workflow, followed by its core component, the E-V match engine.

3.1 Design Overview

As discussed in Section 1, our goal is to improve the localization accuracy of an electronic identifier with the help of video cameras. This can be achieved by corresponding a wireless device, i.e., its electronic identifier, with its shape or owner in the video and then incorporating the visual localization results. In reality, the electronic identifiers and their visual signal counterparts can take any form. In this paper, we take smartphone localization as an example; the visual object corresponding to a smartphone is its owner.

In order to find the correspondence, we estimate the wireless device's location with traditional wireless localization, and match this location to a human's location. If the match is successful, i.e., we find the smartphone's owner, we can fuse the electronic and visual location results to get a more accurate result than can be obtained from purely electronic or visual localization.

However, the wireless localization result can be highly inaccurate. The circles in Fig. 1a give the possible localization result range. The circle in the wireless "vision" is much bigger than that in the camera vision.Consider Fig. 1 as an example. If there is only one smartphone in the wireless vision and one person in the camera vision (Fig. 1a), corresponding the smartphone and person is straightforward. However, if there are several people close to each other (Fig. 1b), corresponding smartphones with people is very hard. In this case, we cannot directly fuse the location results, as a mismatch will result in a large localization error. The correspondence difficulty increases as the number of smartphones and people in our vision increases.



(a)

(b)

Figure 1: Associating wireless devices with their owners

Though inaccurate wireless localization makes correspondence difficult, we can still infer the correspondence statistically. In Fig. 1, if person A really corresponds to smartphone A', the average geometric distance between their location estimates from wireless and visual localization should be the smallest in the long run, e.g., smaller than the average distance between A and B'. Given enough time for location estimation, we can have a certain degree of confidence in determining the correspondence between people and smartphones.

We have two ways to describe the distance between two objects. The traditional way is to use the Euclidean distance between the location estimate points. However, we introduce here the concept of a *location descriptor*. A location descriptor of an object is a tuple of distances between this object and other objects, including the wireless detectors, i.e., access points (APs). We can define the distance between two tuples using any reasonable measure. The advantage of using location descriptors is two-fold. First, location descriptors capture the geometric topology among the objects, and carry more information than an estimated location point. In fact, the location estimate point is calculated from the topology among the object and several APs, so it cannot carry more information than raw distances. Furthermore, we can add the distance between any pair of objects into the descriptor to increase its power and flexibility. Second, the descriptor's error distribution is easier to model than pure estimated location points. The elements in each tuple are distance measurements, whose distribution can be analyzed or at least approximated in closed form. On the other hand, the error distribution of the final point localization result is almost impossible to derive. Though we can use the Cramér-Rao lower bound to estimate the variance [9], this bound is not typically useful in our case. As we will compare many pairs of locations, we need a finer distribution of the error than mere variances.

Note that using the descriptor brings extra benefits if the object to be localized is moving, e.g., humans with smartphones or robotic assistants. Normally, mobility is thought of as a curse for localization, as we cannot average away the noises from multiple measurements. However, if we consider the topological information of all objects to be localized, mobility can be useful to distinguish one object from the others. Specifically, a local position change of one object will change all the distances, i.e., the whole topological information. This more significant change in the topological information can help us uniquely identify and localize this object. For example, when three objects are very close, they cannot be distinguished. But if one of them moves farther, it can be easily identified from the topological change. The faster the topology changes, the faster the objects can be distinguished. Nevertheless, it should be pointed out that our scheme works with both mobile and static objects. While object mobility helps the matching converge faster, multiple readings of a static object help to cancel the noise, which also leads to faster convergence of the matching. Our EV-Loc scheme fully exploits any statistical opportunity for both mobile and static objects.

3.2 Workflow

Fig. 2 describes the main components and workflow for realizing our proposed localization technique.

The data collector has two parts: the electronic signal measuring unit and the visual signal measuring unit. The electronic signal measuring unit is installed on APs or mobile devices used as reference points. It periodically measures RSSI readings from smartphones. The visual signal measuring unit is installed on a central camera, which continuously records time-sequenced visual snapshots containing all objects and their background scene information.

The signal processor conducts two following tasks: signal error modeling and signal to distance conversion. The signal error modeling module is to find a statistical model for

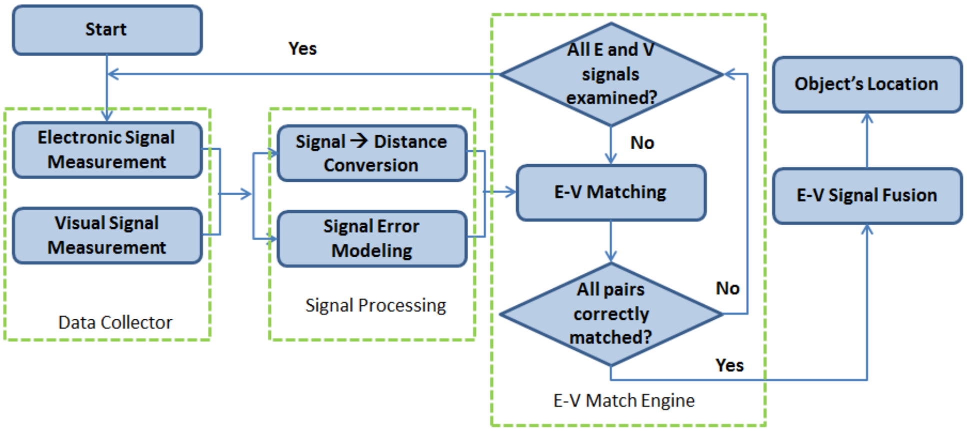

Start
Yes
All E and V signals examined?
No
Object's Location
Electronic Signal Measurement
Visual Signal Measurement
Data Collector
Signal → Distance Conversion
Signal Error Modeling
Signal Processing
E-V Matching
E-V Signal Fusion
All pairs correctly matched?
No
Yes
E-V Match Engine

Figure 2: Workflow

the random observation of the electronic and visual signals collected from each scene. The signal to distance conversion uses site-specific environmental information to calibrate the parameters of the log-distance path loss model and the visual coordinate system. Then it converts each object's electronic and visual signals to estimated distances from the object to all surrounding APs.

The E-V match engine corresponds an object's electronic signals to its visual signals. Its inputs are the estimated distance (converted from electronic and visual signals) and the signal error model. Since the aforementioned E-V correspondence is not known *a priori*, we need to compare each pair of electronic and visual signals. To facilitate this comparison, we use a cost matrix to represent the similarity between each pair of converted distances from the electronic and visual signals. Based on this matrix, the match engine generates a "best match" result with highest similarity between each pair.

The workflow of our approach is as follows. Given an object's electronic identity, e.g., its WiFi MAC address, we first use the data collector to gather RSSI readings from different APs and visual snapshots from the central camera. All collected data are sent to a back-end server for further processing. After processing, the statistical characteristics of the collected signals are generated with the signal error model; the RSSI readings and visual appearances are converted to estimated distances between the object and different APs. Using this information as inputs, the E-V match engine can generate a best match between an object's electronic and visual location descriptors. To ensure the matching result is correct with high confidence, we repeat the matching process using the processed electronic and visual signals at other time points until a certain threshold is reached. Finally, when the match result is stable, we leverage visual localization to estimate the object's location.

It is worth noting that we use highly modular function blocks for adaptability. For example, instead of using a range-based location descriptor, we may also use range-free localization results for E-V matching. However, we need a new error model for each different localization method.

3.3 E-V Match Engine

Now we introduce the core of the above workflow, the E-V match engine. For simplicity, we assume here that the electronic signals and visual signals are complete with no false negatives or positives, i.e., there are no "ghost" or missing objects. Also, we assume the visual signals are distinct, i.e., we can distinguish people in different frames. More practical considerations, e.g., indistinct visual signals or incomplete signals, will be discussed in Section 3.5.

Suppose we have n wireless devices and n people. Each person carries a wireless device, which is uniquely identifiable. The set of electronic location descriptors is $\mathbf{x} = (x_1, \ldots, x_n)^T$, and the set of visual location descriptors is $\mathbf{y} = (y_1, \ldots, y_n)^T$. π_i is a permutation of the sequence $(1, 2, \ldots, n)$, and $\mathbf{y}_{\pi_i} = (y_{\pi 1}, y_{\pi 2}, \ldots, y_{\pi n})^T$ is a permutation of the original vector $\mathbf{y} = (y_1, \ldots, y_n)^T$. Then we can formulate the following best match problem:

$$\arg\min_{\pi_i} \sum_{i=1}^{n} \|x_i - y_{\pi_i}\| \tag{1}$$

$$z_i = \alpha x_i + \beta y_{\pi_i} \tag{2}$$

The problem defined in (1) can be understood as we first find a permutation of $\mathbf{y}$ to match $\mathbf{x}$. (1) can be solved with the standard Hungarian algorithm [18]. After finding such a permutation π_i, we fuse the locations acquired wirelessly and visually into $\mathbf{z} = (z_1, \ldots, z_n)^T$. α and β are the coefficients that reflect the measurement confidence. If the measurement is inaccurate, i.e., the standard deviation is large, we give the location estimate less weight, and vice versa. Suppose every $x_i \in \mathbf{x}$ and $y_i \in \mathbf{y}$ have standard deviations σ_1 and σ_2, respectively. σ_1 and σ_2 are determined by the equipment used and the experiment environment. They remain relatively stable throughout the time. Then we can let $\alpha = {\sigma_1}^{-2}/({\sigma_1}^{-2} + {\sigma_2}^{-2})$ and $\beta = {\sigma_2}^{-2}/({\sigma_1}^{-2} + {\sigma_2}^{-2})$. These are the two optimal coefficients in the maximum likelihood sense given the two standard deviations σ_1 and σ_2. $\|\cdot\|$ is the norm operation. If we take x_i and y_i as coordinates, $\|\cdot\|$ can be the Euclidean distances or squared Euclidean distances. As we note that it is often far more convenient to deal with the squared Euclidean distance, we will stick to squared Euclidean distance in the following part, especially in Section 3.5.

So far, we have assumed that x_i, y_i, and z_i are static coordinates. We can extend them to a function of time, i.e., $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$, $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))^T$ and $\mathbf{z}(t) = (z_1(t), \ldots, z_n(t))^T$. We correspondingly define the norm $\|x_i(t) - y_j(t)\| = \int |x_i(t) - y_j(t)|\mathrm{d}t$, where $|\cdot|$ is the Euclidean distance (or squared Euclidean distance). With these adaptations, we can use (1) and (2) to solve the dynamic version of the problem. Note that we need to adapt the Hungarian algorithm to this dynamic inflow of information. After a new frame arrives at time t_i, we need to recompute the distance and re-run the Hungarian algorithm. However, we notice that we can keep the final matrix obtained in the last round of the Hungarian algorithm and increase the distance based on that matrix. We can consider this an incremental version of the Hungarian algorithm.

In practice, the accuracy of visual localization is far better than that of wireless localization, so it is possible that α is close to 1 and β close to 0. Under this circumstance, it is reasonable to take the visual localization result alone as the final fusion result.

3.4 Derivation of Matching Threshold

The E-V match engine provides a best match and a potential correspondence between each object's electronic identifier and visual appearance. A remaining problem is guaranteeing the correctness of the matching and the ensuing localization results. To address this issue, we derive a matching threshold based on the deviation of the estimated distances.

In general, the location descriptors for the same object from the electronic and visual sides should be the same. (In our cases, they are sets of distances.) But the electronic and visual descriptors actually differ because of noise. We model the noise in the visual distance reading as Gaussian. We also model the signal strength readings as Gaussian [26]. This means the converted distance from signal strength has a log-normal distribution, as signals attenuate exponentially with distance. The above measurements fluctuate around the mean values. If we average the measurements, the result is unlikely to deviate far from the mean values by the central limit theorem. Thus we can bound the deviation and determine with a certain confidence that the electronic descriptor and the visual descriptor do not belong to the same object if their deviations are too large. In the following, we give a rigorous mathematical description of this process.

First, we define the electronic and visual location descriptor as $x_i = (x_i^{AP_1}, x_i^{AP_2}, x_i^{AP_3})$ and $y_i = (y_i^{AP_1}, y_i^{AP_2}, y_i^{AP_3})$, where $x_i^{AP_j}$ or $y_i^{AP_j}$ is the measured distance between the i-th object and the j-th AP. For brevity, we will write x_{ij} and y_{ij} hereafter. In the above descriptor, we only consider the topological relation of objects with APs. It can be easily extended to cases where more comprehensive topology information is used. We can simply append the tuple with the distances to reference objects other than APs in a consistent manner, i.e., all the tuples should include the distance to these objects. Also, we use three APs here for illustration purposes, but we could easily extend the tuple to accommodate the additional AP measurements.

With the above transformation, we can model the variance of each x_i and y_i. We model x_{ij} as a log normal variable. According to [26], we can model $P(d)$, the RSSI reading at distance d, as

$$P(d) = P_0 - a\log_{10}\frac{d}{d_0} + P_n, \tag{3}$$

where P_0 is the original transmission power (known), a is the attenuation coefficient (known), d_0 is a reference distance (known), and P_n is the noise, which can be modeled as a normal random variable, $N(0, \sigma'_x)$. So $x_{ij} = d_0 \cdot \exp\{\ln 10 \cdot (RSSI - P_0 - P_n)/a\}$. We can choose the reference distance d_0 to have unit length and define the exponent as a random variable $N(\mu'_{x_{ij}}, \sigma'_x)$. Here, $\mu'_{x_{ij}}$ can be considered the accurate path loss reading $RSSI - P_0$ for object x_i by AP_j; σ_x^2 is the measurement variance of the AP in that wireless setting, which remains invariant. Let $\mu_{x_{ij}} = \ln 10 \cdot \mu'_{x_{ij}}/a$ and $\sigma_x = \ln 10 \cdot \sigma'_x/a$. Then $x_{ij} \sim \log N(\mu_{x_{ij}}, \sigma_x)$. It is worth noting that, if $x \sim \log N(\mu, \sigma)$, then $E[x] = \exp\{\mu + \sigma^2/2\}$, and $Var(x) = \{\exp(\sigma^2/2) - 1\}\exp\{2\mu + \sigma^2\}$.

On the other hand, y_{ij} can also be modeled as a normal variable, $y_{ij} = y \sim N(\mu_{y_{ij}}, \sigma_y)$. For the same object i, the distances measured wirelessly and visually without noise should be the same, i.e., $\exp\{E[x]\} = E[y]$.

Suppose x_i and y_i represent the same object. Then we can have $\exp\{\mu_{x_{ij}}\} = \mu_{y_{ij}}$. The squared Euclidean distance between $x_i = (x_{i1}, x_{i2}, x_{i3})^T$ and $y_i = (y_{i1}, y_{i2}, y_{i3})^T$ is written as:

$$\Delta_i = \sum_{j=1}^{3}(x_{ij} - y_{ij})^2. \tag{4}$$

If x_i and y_i represent the same object, Δ_i, the distance between x_i and y_i should be lower-bounded by a threshold. If Δ_i is larger than the threshold, we can say with high confidence that the matching is wrong. Now we will calculate the threshold and the confidence.

Let us fix j and look at a single term $M_{ij} = (x_{ij} - y_{ij})^2$. We can bound M_{ij} within $[0, (\mu_{y_{ij}}\exp\{3\sigma_x\} + 3\sigma_y)^2]$ with the 3-σ rule. Though the 3-σ rule is just an approximation, it accurately reflects the fluctuation range of RSSI readings. From our large amount of empirical data and many previous research data, e.g., [20, 3], we find that the 3-σ rule holds in general. Then we apply the Hoeffding inequality, which states that if we have n independent variables $X_1, \ldots, X_n$, $\Pr(X_i \in [a_i, b_i]) = 1$ and $\bar{X} = \sum_i X_i/n$, then

$$\Pr(\bar{X} - E[\bar{X}] \geq t) \leq \exp\left\{-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}. \tag{5}$$

Let each M_{ij} be a random variable and substitute it in (5). We find that

$$\Pr(\Delta_i - E[\Delta_i] \geq t) \leq \exp\left\{-\frac{2t^2}{\sum_j(\mu_{y_{ij}}(\exp\{3\sigma_x\} - 1) + 3\sigma_y)^4}\right\}. \tag{6}$$

It can be seen that $E[\Delta_i] = (\mu_y^2 \cdot (\exp\{2\sigma_x^2\} - 2\exp\{\sigma_x^2/2\} + 1) + \sigma_y^2)$. From (6), we know that if we have $\Delta_i > t + E[\Delta_i]$, then the possibility that x_i is not y_i is at least $1 - \exp\{-2t^2/\sum_j \mu_{y_{ij}}(\exp(3\sigma_x) - 1) + 3{\sigma_y}^4\}$.

The above inequalities form the basis for evaluating the quality of the matching. A good matching must satisfy the following conditions. If x_i is matched with y_j, Δ_{ij}, which is the squared Euclidean distance between the x_i and y_j tuples, must be below a certain threshold with a certain confidence, and Δ_{ik} ($k \neq j$), the distance between x_i and any other y_k tuples than y_j, should be larger than a threshold with a certain confidence. Specifically, $\Pr(\Delta_{ij} - E[\Delta_{ij}] \leq th_1) \geq c_1$, and $\forall k \neq j$, we have $\Pr(\Delta_{ik} - mE[\Delta_{ik}] \geq th_2) \geq c_2$. Empirically, we can set c_1 very large, e.g., 99.9%, and let $c_2 \in [90\%, 99\%]$. We can then compute thresholds th_1 and

th_2 and use them to decide the appropriateness of the matching.

We note that we do not actually know $\mu_{x_{ij}}$ and $\mu_{y_{ij}}$ in our calculation. In practice, we may take the visual measurements as approximate $\mu_{y_{ij}}$s as they are relatively accurate. With $\mu_{y_{ij}}$, a simple logarithmic operation can give us $\mu_{x_{ij}}$.

3.5 Extensions to Practical Settings

In this subsection, we discuss practical issues beyond our baseline cases given above. We examine the cases where some objects' visual signals are indistinct, or some objects' electronic or visual signals are missing, i.e., false negatives in the detection. It is worth noting that false positives with respect to electronic sensing can be considered as false negatives with respect to visual sensing, and vice versa. For example, if we detect an irrelevant visual object without wireless devices or a post is mistakenly identified as a person, a false positive takes place. We can view this false positive object "missing" an electronic identifier. If our E-V match engine is robust enough, we will not associate any electronic object with this irrelevant visual object.

To handle the missing object cases, we look deeper into the Hungarian algorithm. There are several implementations of the Hungarian algorithm. One of them is based on maximum flow. It finds a maximum flow in a bipartite graph, regardless of the number of nodes in each graph bipartition. This means that such an implementation can have different numbers of x_is and y_is and still finds a best match. We can use this Hungarian algorithm implementation to handle the missing objects cases. We will evaluate the robustness of our proposed approach when facing the indistinct and missing visual signal cases in Section 4.

With indistinct visual objects, we cannot distinguish among people in different visual frames. We need to determine this correspondence. We first define some terminology before formulating the problem. Suppose we have m frames and n wireless and visual objects in each frame. Then we have $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))^T$ as wireless location descriptors and $\mathbf{y}(t) = (y_1(t), \dots, y_n(t))^T$ as visual location descriptors, where $t = t_1, \dots, t_m$. Here we notice that $x_i(t_1)$ is the same object as $x_i(t_2)$, but $y_i(t_1)$ is not necessarily the same object as $y_i(t_2)$. We only know that there are n visual objects at times t_j and t_k. We randomly place the visual location descriptors in the descriptor tuple. Thus, we can only say that $y_i(t_1)$ is some $y_j(t_2)$, but we do not know which y_j. However, we have a distance matrix containing the distances between each pair of visual objects in different images. We want to find a permutation π_i for each time point, i.e., $\pi_i(t)$, that minimizes the total sum of location differences and visual object distances.

We can visualize the problem with the help of Fig. 3. We want to match one x_i with one y_j at each time point $t_1, \dots, t_m$. We have a location distance matrix XY_i between $\mathbf{x}$ and any one of the m $\mathbf{y}(t_i)$s and a visual distance matrix Y_{ij} between each pair of $\mathbf{y}(t_i)$ and $\mathbf{y}(t_j)$. Y_{ij} details the visual dissimilarity between an object in $\mathbf{y}(t_i)$ and an object in $\mathbf{y}(t_j)$. By associating an E signal with one or two V signals from two different visual frames, we actually pick a cost in the distance matrix. A natural way to formulate the problem is to find $\pi_i(t)$ to minimize the total association cost from each XY_i and Y_{ij}. However, we face two difficulties here.

First, for comparisons of every pair of visual objects in different frames (m frames in total), the visual distance itself has space complexity $O(n^m)$, which is exponential in m. Then the time complexity is at least exponential in m. If m is relatively small, e.g., $m \leq 10$, we may try enumeration. But as m increases, brute force enumeration will soon fail.

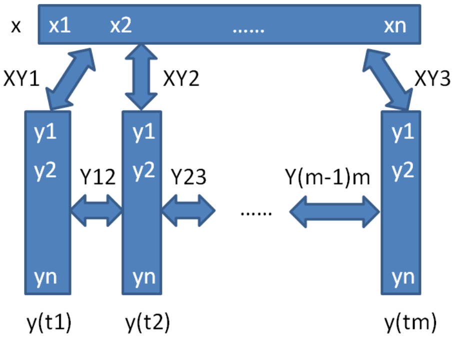


Figure 3: Problem formulation

Second, the above problem is in fact a multi-dimensional assignment problem [7]. It is NP-hard and even inapproximable when the dimension is at least 3. Even if we only consider visual tuples at three time points, finding a best match to minimize the cost is infeasible for large n.

In order to cope with the above two difficulties, we take the following strategies:

First, we only consider the distance between two frames neighboring in time. This practice makes sense, because two frames far away in time may have low correlation, and the distance of objects therein can be heavily distorted. We may extend the processing scope to three or more frames neighboring in time. In essence, we only consider a finite number of frames close in time, not the entirety of frames. Then, we need to solve the following problem:

$$\arg\min_{\pi_j(t)} \sum_{i=1}^{m} \sum_{j=1}^{n} |y_{\pi_j}(t_i) - y_{\pi_j}(t_{i+1})|, \tag{7}$$

where, for convenience, we write $y_{\pi_i}(t)$ to denote $y_{\pi_i(t)}(t)$, and define $|y_{\pi_j}(t_m) - y_{\pi_j}(t_{m+1})| := 0$.

Second, we perform visual object matching first. After getting the matching, the objects' visual appearances can be considered distinct, and then we run the E-V match engine discussed in Section 3.

In this paper, we leverage visual tracking techniques to generate a similarity matrix between every pair of consecutive visual frames. Then we perform $m-1$ rounds of the Hungarian algorithm to find the best match between every pair of neighboring visual frames. After this step, we can use the E-V match engine to localize an object as discussed in previous section. The performance of this approach is evaluated in Section 4.

4. EVALUATION

In this section, we evaluate the performance of our proposed approach. To evaluate the localization accuracy in real-world settings, we conduct both indoor and outdoor experiments using mobile phones and laptops. To evaluate the efficiency and robustness of our proposed approach at large scale, we perform simulations with different environmental settings and population sizes.

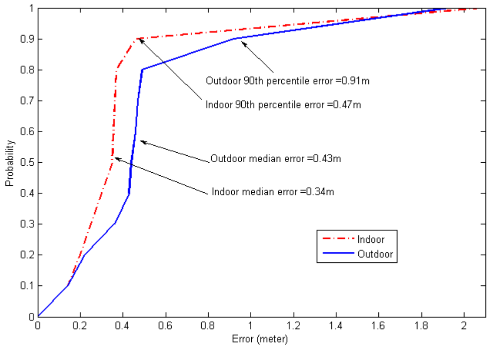


Figure 4: Cumulative distribution function (CDF) of the error distance

4.1 Real World Experiments

We conduct experiments to localize mobile phones in both indoor and outdoor settings. The indoor experiments are conducted in a 10 m × 10 m area inside a research building. The outdoor experiments are conducted in a 20 m × 20 m area outside the research building. As mobile phones are too small to see in the image, we associate a discernible object with each phone. In this case, we choose its owner as the visual object, and we view the identifier and the person as an integral entity, i.e., we do not consider the impact of human body on the phone signals. So we let the phones lie on the ground with the owners. We will discuss the impact of human bodies in the next section. In all experiments, we have five colleagues with WiFi-equipped mobile phones as target objects. We performed 10 indoor experiments and 10 outdoor experiments. In each experiment, we recorded 15 electronic frames and 15 visual frames. The RSSI readings recorded in each frame are averaged over 100 measurements. The time for APs to conduct 100 measurements varies between 5 to 20 seconds.

To capture target objects' visual information, we set up a camera shooting from above and covering the entire area. The surveillance camera is calibrated with known intrinsic parameters including focal length, lens distortion, and the relative rotation and translation with respect to the scene. The target objects' visual appearances are detected using the HoG pedestrian detector [12] and the correspondences across frames are established using the detection scheme in [1]. A person's planar coordinates in the scene are calculated by the pixel location of the bottom center of each bounding box. To capture target objects' electronic information, we use three laptops on the ground as APs to detect nearby mobile devices' WiFi MAC addresses. The WiFi on target objects' mobile phones is turned on, enabling them to be continuously detected.

The result of all experiments is shown in Fig. 4. For the indoor experiments, we achieve a median error of 0.34 m and 90 percentile error for 0.47 m. For the outdoor experiments, the median error is 0.43 m and the 90 percentile error is 0.91 m. Specifically, our E-V match engine can achieve a 90% (indoor) and an 80% (outdoor) success rate for correctly pairing all target objects' WiFi MACs and visual appearances. On average, the E-V match engine uses 6 per 15 frames to converge indoors and 8 per 15 frames to converge outdoors. The parameter settings of our matching threshold are $\sigma_x = P \cdot 5/\sqrt{K}$ and $\sigma_y = 0.3$, where P is the tuning parameter for RSSI to distance conversion and is learned from each scenario. K records the number of RSSI measurements (100) for each frame.

4.2 Large-Scale Simulations

Since it is difficult to conduct larger scale real experiments, we conduct extensive simulations to further validate the performance of our proposed approach in different environment settings. In our simulations, we want to localize electronic identifiers, which may be static or moving, e.g., humans with mobile phones or the robotic assistants. All objects are randomly distributed in an area of 10 m × 10 m, and, if they are mobile, they move with a constant speed under the random waypoint model [8]. Depending on the coordinates of a given object within the area, its RSSI as received by the APs is simulated, and a perspective distortion determined from its distance to the camera is applied. For each location, we obtain 10 RSSI readings, and the standard deviation of RSSI is set to $5/\sqrt{10}$.

The visual appearances of objects are simulated by taking human-shape samples from the INRIA database [11] and scaling the sample in each frame according to its relative location with respect to the camera. (Whether the objects are human or not is not important, so long as the visual objects can be extracted from the images.) Samples of different poses from the same objects are randomly selected, and image noise is added to the visual appearances to simulate appearance changes in different frames. The HSV histogram feature with 8×8×4 bins is extracted for each visual object, and the similarity between two visual objects is determined using the Bhattacharya distance between their histograms [12, 24].

We measure the performance of our proposed scheme using three aspects: (1) The accuracy of our proposed approach as compared to localization using only electronic signals. We measure it by examining the variation of the localization accuracy when a number of frames are captured in a single scenario; (2) The efficiency of our proposed approach. We evaluate it by examining the number of frames needed to reach a given localization accuracy by changing the number of APs, the number of localized objects, and the motion speed of localized objects; (3) The robustness of our proposed approach when there are missing and indistinct

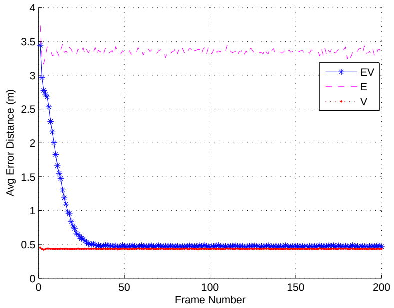
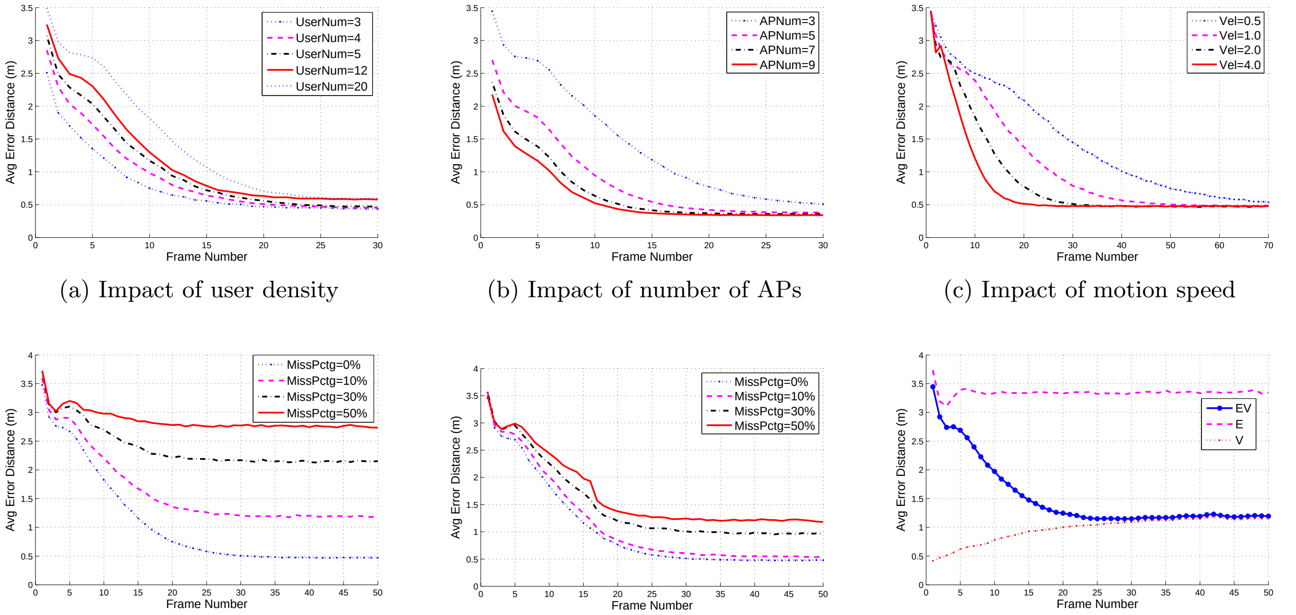


Figure 5: Localization performance comparison among different approaches



(a) Impact of user density

(b) Impact of number of APs

(c) Impact of motion speed

(d) Impact of missing visual objects

(e) Impact of missing electronic signals

(f) Impact of indistinct visual objects

Figure 6: Efficiency and robustness of our approach

visual signals. All simulation results are reported in Fig. 5 and 6. All results are averaged over 5000 runs.

Fig. 5 shows the accuracy of our proposed approach with respect to the number of input electronic and visual frames. We compare our approach with pure wireless and visual localization. For wireless localization, we use RSSI-based distance estimation plus multilateration to estimate the position of localized objects, as this technique is more practical and more widely available than other wireless localization techniques. We let 10 objects move with a speed of 2 m/s and deploy 3 APs on the boundary to form an isosceles triangle. There are three curves in the figure. The top curve shows the error distance using RSSI localization, the middle curve is the error distance of our approach, and the bottom curve is the error distance of visual localization. We can see that the wireless localization consistently has an error of around 3.4 m, while our approach has a decreasing error distance as the frame number increases. This is because more and more objects are corresponded between their electronic identifiers and visual appearances. After around 40 frames, our approach can fully leverage the high accuracy of visual localization to achieve an average error about 0.5 m.

In Fig. 6a, we evaluate the impact of the number of localized objects on the localization error distance. The initial position of each object is random, and their movements are independent. The figure shows that we need more frames to reach a certain error distance when there are more objects. However, according to the trend of the displayed curves, we can also discern that the number of frames needed to reach a certain accuracy does not remarkably increase with more objects. Eventually, they all reach a very high accuracy. This result shows that our approach is efficient even if the density of localized objects is high.

In Fig. 6b, we evaluate the impact of the number of APs on the localization error distance. All the APs are deployed along a regular polygon on the boundary. When the number of APs exceeds 8, we deploy them in a grid shape. As more APs are deployed, the error distance of our approach decreases rapidly. This shows that additional APs improve the accuracy of EV-Loc. Eventually, the error distance becomes stable at around 0.5 m.

In Fig. 6c, we evaluate the impact of the motion speed of the localized objects on EV-Loc's error distance. The trend of the curve shows that we need less frames to reach a certain accuracy when the localized objects move faster. This result confirms that the combination of objects' movement and their topological relationship can effectively reduce the convergence time of the E-V match engine.

In Fig. 6d, we evaluate the error distance of EV-Loc in the case where objects are not all visually detected (misses). Generally, the localization accuracy decreases when the objects' visual miss rate increases. We also observe that the localization error distance is decent at around 1.2 m even if the miss rate is as high as 10%.

In Fig. 6e, we evaluate the error distance of EV-Loc in the case where objects are not all electronically detected (misses). In this situation, some objects' visual appearances simply do not have corresponding electronic identifiers. According to the simulation result, we find the localization accuracy decreases when the objects' electronic miss rate increases. However, such decline does not significantly impact the accuracy of EV-Loc. Its error distance remains around 1 m even when 30% of objects are missing.

In Fig. 6f, we evaluate the localization error distance in the case where the objects' visual appearances are indistinct. To achieve this effect, we randomly permute the order of visual appearances in different frames. Generally, the localization accuracy decreases when the objects' visual appearances become indistinct. This is because the object that fails to find a match will simply use wireless localization results. As is shown in the figure, our solution for handling the object's visual indistinction can achieve similar performance to the case with distinct visual appearances.

5. DISCUSSIONS

Though our paper focuses on introducing a new idea and enabling it, this section discusses some problems that the conceptual EV-Loc system may encounter in practice.

The EV-Loc system requires two sensing systems working at the same time, one wireless network, and one camera network. Compared with traditional wireless localization, there will be additional costs and coordination efforts in deployment. However, we believe these issues do not pose a major hurdle to deployment. Cameras are becoming increasingly affordable: common webcams normally cost under 100 dollars. Moreover, we can use existing surveillance networks, as they are increasingly pervasive for public or private usage. The camera network needs to be calibrated for localization, i.e., the transformation matrix from an image point to a real location should be known. This can be done during installation. The cameras also need to be synchronized with the wireless network, as we need to match objects from both sensors. Here a small temporal drift of up to 1 s for synchronization is usually tolerable. We can achieve it via the NTP protocol. Other relevant infrastructure issues for EV-Loc include sensor deployment and data management. For cost efficiency, we want to cover the whole monitored area with as few sensors as possible. This is the optimal sensor deployment problem. We plan to follow the research of [4, 33] and design optimal AP and camera deployment patterns as a part of our future work. On the other hand, in real implementation, we may need a database and a mechanism, such as [29], for recording, managing and searching through electronic and visual signals and the final location results.

In the paper, we have implicitly assumed that there is little visual occlusion. We have shown that EV-Loc works under this assumption. However, in reality, there might be serious occlusions when there are many objects in the monitored area. One way to deal with occlusion is to follow the method proposed in Section 3.5 and treat erroneous detections due to visual occlusion as missing IDs. Another more advanced method is to enable visual tracking. Currently, visual tracking technologies allow continuous tracking when occlusion occurs in some circumstances. Using multiple cameras and tracklets are two major approaches [1, 21]. Though it is well beyond the scope of this paper, we plan to incorporate these technologies in our future work.

Environmental interference can also impact EV-Loc's accuracy. For example, human body can cause a blocking effect on transmitted electronic signals [34]. To examine the impacts of such interference, we conduct a preliminary experimental study in an outdoor environment. There are five experiment participants, all holding mobile phones in their hands and standing in random positions. The purpose of such settings is to consider the human body's blocking effects on electronic signals. Based on our experimental result, we find that the median error of EV-Loc is around 0.5 m, while the 90 percentile error is around 2.5 m. The elevated 90 percentile error distance indicates a simple error distribution model of electronic signals may not suffice to describe human body interference. Based on this observation, we plan to investigate the impact of the human interference on EV-Loc's localization accuracy in our future work.

The EV-Loc technique can easily work with other electronic and visual localization methods. We can simply input the location estimation results obtained via other localization methods as location descriptors into our workflow (with a well-defined distance measure), model the error distribution of each localization method separately, and then run our E-V match engine. For example, to accommodate range-free localization, we take coordinates-based location descriptors as input to our E-V match engine and set the cost function of associating electronic and visual objects appropriately. By properly selecting the error distribution model (e.g. [17]), we can match an object's electronic and visual signals and then apply the proposed technique to localize it. With more accurate localization estimates on the electronic side, the association time is expected to be shorter.

EV-Loc is largely centralized, like most current existing localization algorithms [10, 3, 2, 5, 13]. In some circumstances, a distributed algorithm is desirable for efficiency and privacy reasons. We can achieve different levels of distributed processing under the EV-Loc framework. In the simplest distributed version, electronic and visual sensors perform part of the computations, e.g., the cameras complete human detection and location estimation on their own rather than delegating the tasks to the central server. In a more complicated version, a hierarchy can be established in the EV-Loc system. We partition the EV-Loc sensors, both electronic and visual, into clusters and assign cluster headers. This is useful for large area deployment as global information is rarely useful for localizing an object that is unlikely to move much. Finally, in a fully distributed version of EV-Loc, each sensor performs its own object localization, but sufficient information exchange needs to happen in a properly large geographical area, as is required for E-V matching. The fully distributed version introduces further scheduling and error tolerance issues, which we aim to explore in future work.

6. FINAL REMARKS

In this paper, we presented a technique called EV-Loc for accurate localization based on electronic and visual signals. Given an object's electronic identifier, we aimed to accurately localize the object with the help of visual signals. In order to achieve this, we proposed the E-V match engine that can accurately and efficiently correspond an object's electronic and visual signals. Following the matching results, we used visual localization to precisely estimate the given object's position. We also considered practical situations, e.g., missing or indistinct electronic and visual signals, and devised schemes to eliminate their impacts. We implemented EV-Loc on mobile devices and conducted real world experiments and large scale simulations to evaluate our proposed approach. The results showed our approach can achieve high localization accuracy and the underlying matching algorithm is efficient and robust.

Acknowledgments

This work was supported in part by China 973 Project No. 2011CB302800, the National Science Foundation of China (NSFC) under grant No. 61070221, and the US National Science Foundation (NSF) under Grant No. CNS0916584, CNS1065136. Any opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

We would like to thank Adam C. Champion, Gang Li and Qiang Zhai for their help in the discussions and experiments. We would also like to thank the anonymous reviewers and our shepherd for their constructive comments which help to bring the paper to its current state.

7. REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking -by-detection and people-detection -by-tracking. In *Proc. of IEEE CVPR*, 2008.

[2] M. Azizyan, I. Constandache, and R. R. Choudhury. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proc. of ACM MobiCom*, 2009.

[3] P. Bahl and V. N. Padmanabhan. RADAR: an in-building rf-based user location and tracking system. In *Proc. of IEEE INFOCOM*, March 2000.

[4] X. Bai, C. Zhang, D. Xuan, J. Teng, and W. Jia. Low-connectivity and full-coverage three dimensional networks. In *Proc. of ACM Mobihoc*, 2009.

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

[6] N. Bulusu, J. Heidemann, and D. Estrin. Gps-less low cost outdoor localization for very small devices. *IEEE Personal Communications Magazine*, 7(5):28–34, October 2000.

[7] R. E. Burkard and E. Cela. *Handbook of Combinatorial Optimization, Volume A*. Kluwer Academic Publishers, 1999.

[8] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing*, 2(5):483–502, 2002.

[9] C. Chang and A. Sahai. Cramér-rao-type bounds for localization. *EURASIP Journal on Applied Signal Processing*, pages 1–13, 2006.

[10] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan. Indoor localization without the pain. In *Proc. of ACM MobiCom*, pages 173–184, 2010.

[11] N. Dalal. INRIA person dataset. `http://pascal.inrialpes.fr/data/human/`, 2005.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of IEEE CVPR*, pages 886–893, June 2005.

[13] M. Ding, F. Liu, A. Thaeler, D. Chen, and X. Cheng. Fault-tolerant target localization in sensor networks. *EURASIP J. Wirel. Commun. Netw.*, 2007(1):19–19, Jan. 2007.

[14] A. Ferencz, E. Learned-Miller, and J. Malik. Learning hyper-features for visual identification. In *Proc. of NIPS*, pages 425–432, 2005.

[15] Y. Gwon and R. Jain. Error characteristics and calibration-free techniques for wireless lan-based location estimation. In *Proc. of ACM MobiWac*, September 2004.

[16] T. He, C. Huang, B. Blum, J. Stankovic, and T. Abdelzaher. Range-free localization schemes for large scale sensor networks. In *Proc. of ACM MobiCom*, pages 81–95, 2003.

[17] A. Karbasi and S. Oh. Distributed sensor network localization from local connectivity: performance analysis for the hop-terrain algorithm. *SIGMETRICS Perform. Eval. Rev*, 38(1):61–70, June 2010.

[18] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[19] H. Liu, H. Darabi, P. Banerjee, and J. Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Trans. Syst., Man, Cybern. C: Applications and Reviews*, 37(8):1067–1080, November 2007.

[20] Y. Liu, Z. Yang, X. Wang, and L. Jian. Location, localization, and localizability. *Journal of Computer Science and Technology (JCST)*, 25(2):274–297, 2010.

[21] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Proc. of IEEE CVPR*, pages 1988–1995, june 2009.

[22] R. Nagpal. Organizing a global coordinate system from local information on an amorphous computer. In *A.I. Memo 1666*. MIT A.I. Laboratory, August 1999.

[23] D. Niculescu and B. Nath. DV based positioning in ad hoc networks. *Journal of Telecom. Systems*, 2003.

[24] K. Nummiaro, E. Koller-Meier, and L. V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21(1):99–110, January 2003.

[25] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. The cricket location-support system. In *Proc. of ACM MobiCom*, pages 32–43, 2000.

[26] S. Y. Seidel and T. S. Rappaport. 914MHz path loss prediction models for indoor wireless communications in multifloored buildings. *IEEE Transactions on Antennas and Propagation*, 40(2):209–217, 1992.

[27] D. Smith and S. Singh. Approaches to multisensor data fusion in target tracking: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1696–1710, December 2006.

[28] J. Teng, J. Zhu, B. Zhang, D. Xuan, and Y. Zheng. E-V: efficient visual surveillance with electronic footprints. In *Proc. of IEEE INFOCOM*, March 2012.

[29] H. Wang, C. Tan, and Q. Li. Snoogle: A search engine for the physical world. In *Proc. of IEEE INFOCOM*, April 2008.

[30] A. Yilmaz, O. Javed, and M. Shah. Object tracking: a survey. *ACM Comput. Surv*, 38(4), December 2006.

[31] M. Youssef and A. Agrawala. The Horus WLAN location determination system. In *Proc. of ACM MobiSys*, June 2005.

[32] M. Youssef, A. Youssef, C. Reiger, A. Shankar, and A. Agrawala. Pinpoint: an asynchronous time-based location determination system. In *Proc. of ACM MobiSys*, pages 165–176, 2006.

[33] Z. Yu, J. Teng, X. Bai, D. Xuan, and W. Jia. Connected coverage in wireless networks with directional antennas. In *Proc. of IEEE INFOCOM*, 2011.

[34] Z. Zhang, X. Zhou, W. Zhang, Y. Zhang, G. Wang, B. Y. Zhao, and H. Zheng. I am the antenna: accurate outdoor AP location using smartphones. In *Proc. of ACM MobiCom*, pages 109–120, 2011.