Multiview and Multimodal Pervasive Indoor Localization

Zhenguang Liu
National University of Singapore
liuzhenguang2008@gmail.com

Li Cheng
Agency for Science Technology and Research
chengli@bii.a-star.edu.sg

Anan Liu
Tianjin University
anan0422@gmail.com

Luming Zhang
Hefei University of Technology
zglumg@gmail.com

Xiangnan He
National University of Singapore
xiangnanhe@gmail.com

Roger Zimmermann
National University of Singapore
rogerz@comp.nus.edu.sg

ABSTRACT

Pervasive indoor localization (PIL) aims to locate an indoor mobile-phone user without any infrastructure assistance. Conventional PIL approaches employ a single probe (i.e., target) measurement to localize by identifying its best match out of a fingerprint gallery. However, a single measurement usually captures limited and inadequate location features. More importantly, the reliance on a single measurement bears the inherent risk of being inaccurate and unreliable, due to the fact that the measurement could be noisy and even corrupted.

In this paper, we address the deficiency of using a single measurement by proposing the original idea of localization based on multi-view and multi-modal measurements. Specifically, a location is represented as a multi-view graph (MVG), which captures both local features and global contexts. We then formulate the location retrieval problem into an MVG matching problem. In MVG matching, a collaborative-reconstruction based measure is proposed to evaluate the node/edge similarity between two MVGs, which can explicitly address noisy measurements or outliers. Extensive experiments have been conducted on three different types of buildings with a total area of 18,719 m^2. We show that even with 30% noisy measurements or outliers, our method is able to achieve a promising accuracy of 1 meter. As another contribution, we construct a benchmark dataset for the PIL task and make it publicly available, which to our knowledge, is the first public dataset that is tailored for multi-view multi-modal indoor localization and contains both magnetic and visual signals.

CCS CONCEPTS

•**Human-centered computing → Ubiquitous and mobile computing systems and tools;**

KEYWORDS

Infrastructure-free; pervasive indoor localization; multiview; multimodal; graph matching

1 INTRODUCTION

Pervasive indoor localization (PIL) is a fundamental problem in location-based services and applications. Accurate and reliable indoor positioning could enable a wide range of applications [19, 23, 33] such as finding a conference room in an unfamiliar building, navigating an individual toward the nearest safety exit in case of fire, guiding a customer in a shopping mall, and routing robots in a fully automated factory [15, 27], etc. Meanwhile, despite the fact that GPS signals have been widely used for navigation in outdoor environments, robust and accurate indoor positioning remains an unsolved challenge [23].

As GPS signals are usually blocked by concrete walls, researchers have explored various possibilities for indoor localization, including using WiFi, Bluetooth, ultrasound, or infrared, to name a few [4, 8, 17, 33]. However, most of existing approaches rely on the deployment of a great many beacons or access points (APs), which are expensive to deploy and difficult to maintain. Moreover, it might be infeasible to deploy these devices in some buildings due to safety or privacy concerns. Hence, there has been a remarkable shift in research efforts towards infrastructure-free indoor localization, which is more scalable and pervasively available. Among these infrastructure-free methods, approaches using dead reckoning (e.g., [19, 23]), vision (e.g., [28, 31]), and magnetic fields (e.g., [7, 22, 33]) have shown great promise. However, few of them can achieve the needed meter-level accuracy. More importantly, their performances are known to be very sensitive to noise, making them unreliable and unsatisfactory for practical usage.

We believe that existing issues are mainly due to the following two reasons: first, the presence of noise is almost inevitable in both the probe and fingerprint measurements collection, due to sensor noise, image blur, and image out-of-focus, among others. Second, conventional approaches rely on a single probe measurement to localize by identifying its best match in the fingerprint gallery. However, a single measurement captures limited location features, and bears the inherent risk of being noisy or even corrupted.

To overcome these challenges, we present an accurate and robust indoor-localization system termed *MviewGraph* for mobile-phone users without using any infrastructure assistance. The only requirement on the user side is to shoot a few photos of the surrounding scenes in different directions. The motivation stems from the fact that a person also locate herself by looking around at scenes in multiple directions. Without loss of generality, usually four sets of measurements are collected, with each focusing on a distinct view (i.e., front-view, right-view, behind-view and left-view) of a location. Each such measurement set is also referred to as a *view*

MM'17, October 23–27, 2017, Mountain View, CA, USA.
© 2017 ACM. 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123436

cluster, as it contains the cluster of measurements for a specific view. A multiview graph (MVG) is then constructed from the four view clusters of a location l, which characterizes l by utilizing comprehensive local features and global attributes. As such, each location is represented as an MVG and the localization problem can then be formulated as a graph matching problem of pairwise MVGs. For MVG matching, we design a novel collaborative-reconstruction scheme to explicitly deal with noisy measurements or outliers.

In our settings, each measurement has two modalities: an image and the simultaneously collected magnetic signal. The reasons for utilizing these two modalities are two-fold. (1) Both, images and geomagnetic signals, are omnipresent and can be easily collected by a common mobile-phone (camera & magnetometer sensors). (2) Images and the geomagnetic field are complementary in resolving positions since images are usually distinguishable across distant locations while magnetic signals are known to be more locally distinctive [26, 33]. It is worth noting that floor identification is a well studied research problem (e.g., [1, 2]), as such here we focus on studying indoor localization on a single floor.

To summarize, the key contributions of this work are:

- We are the first to leverage multimodal multiview measurements for accurate and robust indoor localization. An MVG structure is devised that is capable of capturing local features and global contexts to characterize a location. A collaborative-reconstruction based MVG matching scheme is further developed to explicitly address the issue of noisy measurements.
- Extensive experiments have been conducted on three large and complex buildings with a total area of 18,719 m^2. Empirical evaluation shows promising results – *MviewGraph* can achieve a relatively high accuracy of 1 meter, even when 30% noisy measurements are presented.
- We construct and publish a benchmark dataset for the PIL task, which to our knowledge is the first public dataset that is tailored for multi-view multi-modal indoor localization and contains both magnetic and visual signals.

In the rest of this paper, we first review related work in Section 2. We then introduce the overall architecture of our system in Section 3 before presenting the location retrieval algorithms in Section 4. Afterwards, we demonstrate continuous user tracking in Section 5 and evaluate the performance of each module and the overall system in Section 6. Finally, we conclude the paper with discussions on the system and future work in Sections 7 and 8.

2 RELATED WORK

The maturity of wireless and embedded technology and the ubiquity of mobile-phone sensors have fostered the development of indoor localization techniques [35, 36]. Most existing methods however assume the existence of some form of infrastructure. For example, WiFi-based techniques require pre-deployed WiFi APs and infrared (IR) based methods need IR beacons. It is beyond the scope of this paper to deliver an exhaustive report on the research activities in this field. Instead, we provide a relatively succinct account of the most related efforts, which can be roughly classified into three categories, namely, infrastructure-based, pervasive, and fusion approaches.

Infrastructure-based approaches localize by leveraging privileged information from a dedicated infrastructure. Among them, WiFi-based approaches are the most well-studied (e.g., [4, 5, 8, 24]). WiFi triangulation uses the received signal strengths (RSS) from APs and a propagation model of WiFi radio signals to locate the user, while WiFi fingerprinting works by finding the best match among the RSS signature of the probe and those from the fingerprint gallery. Prior knowledge of AP positions is usually required, and the localization accuracy is often not high due to unstable WiFi signals. IR, ultrasound, and Bluetooth based approaches have also been proposed with meter-level localization accuracy being reported. These techniques, however, heavily rely on the existence of certain external infrastructures. For example, since an IR beacon has a very limited coverage (within a room), usually hundreds to thousands of IR beacons have to be installed to cover the entire space of a building.

Pervasive approaches aim to locate a mobile-phone user without any additional infrastructure. Among them, pedestrian dead reckoning, magnetic-field-based and vision-based methods are most related to our work. *Pedestrian dead reckoning* [6, 9, 19, 23] continuously estimates a positional displacement based on the readings of IMU sensors (gyroscope, accelerometer, compass), so as to track the user. The methods are well known to be simple and infrastructure-free, but they suffer from an inherit error-accumulation problem.

Magnetic-field-based approaches (e.g. [13, 22, 33]) utilize the locally anomalous but stable geomagnetic field for indoor localization. In [7, 13, 22], the authors study the feasibility of leveraging magnetic field signals as location fingerprints. Further, in [26, 33, 40], researchers conduct in-depth empirical evaluations and discover three favorable properties of geomagnetic field for indoor localization task: locally distributed, stable over time, and limitedly influenced by mobile objects. However, since the feature dimensionality of the geomagnetic field is low, the uniqueness of a magnetic fingerprint can not be guaranteed.

Vision-based approaches (e.g., [16, 28, 36]) construct a gallery of fingerprint images and their associated locations offline. Then in the online phase, a comparison is made between a newly captured image and the fingerprint images stored in the gallery to identify the best match [16, 28]. Traditionally, only handcrafted image features such as histograms, SIFT and texture features are utilized in these approaches. Although deep learning approaches have achieved great success in image classification due to the maturity of large annotated datasets and innovative deep neural networks [3, 10, 20, 29, 32, 38, 41, 42], few systems have taken these advantages further to address the problem of indoor localization. Meanwhile, several vision methods (e.g., [18]) attempt to reconstruct the 3D indoor structure of buildings and then perform location retrieval in the constructed 3D database with a newly captured image probe. The computational costs of this line of research efforts are often too high to be practically feasible.

Fusion approaches aim to improve performance by combining multiple modalities. In [33], a fusion approach of magnetic field and WiFi sensors is proposed to integrate the magnetic field and WiFi, thus alleviating the low dimensional issue of using a magnetic field alone, and finally reporting a three-meter-accuracy. In [30], the possibility of combining radio and camera sensors is studied. In [4], a hybrid approach utilizing WiFi and Bluetooth is proposed, where

Bluetooth hotspots are used to divide the large space into small partitions. This step is followed by WiFi fingerprinting to infer the fine-grained location of a user. Fusion approaches of [12, 14, 17, 30] re-confirm that incorporating multiple raw signals can lead to improved performance. Unfortunately, these fusion approaches either require the installation of certain infrastructure facilities, or encounter difficulties in achieving the desired meter-level accuracy.

3 OVERVIEW OF THE PROPOSED SYSTEM

Next, we provide an overview of our *MviewGraph* system, which consists of two phases: offline preparation and online localization. The overall architecture of *MviewGraph* is depicted in Fig. 1.

In the **offline preparation phase**, for each position-of-interest (PoI), four measurement sets (empirically a set contains 2-6 measurements) are collected, each for one of the four views, namely front-view, right-view, behind-view, and left-view, in a fixed (clockwise) order. Then, from the four measurement sets, *MviewGraph* constructs the multiview graph (MVG) G for each PoI l to capture the location features of l. Finally, the fingerprint gallery is assembled, where each fingerprint is stored as a tuple $\langle G, l \rangle$. All these operations are completed offline to be ready for online localization.

In the **online localization phase**, as a probe of the target location, the user collects four measurement sets (empirically a set here contains 2 measurements). Again each measurement set corresponds to one distinct view. These measurements are then transmitted to the cloud server for localization. On the server end, an MVG G of this probe is constructed for querying. Subsequently, MVG matching is performed to identify the best matched MVG $\tilde{G}^*$ of G from the fingerprint gallery. The location label of $\tilde{G}^*$ is leveraged as the user location estimation. As an extension, we also consider the practical scenario where the user would like to continuously localize by intermittently querying with the multiview multimodal probes of the current new location. For this scenario, we augment the classic particle filter method to track the user.

Measurement collection details: Each time the user takes a photo, the magnetometer will also be activated for collecting the associated magnetic field signal. This produces a multimodal measurement. For front view, the mobile-phone is held towards the walking direction (i.e., along the main direction of a path), and the collection order of the four views is fixed to clockwise.

Next, the location retrieval module is detailed in Section 4, while continuous tracking is introduced in Section 5.

4 LOCATION RETRIEVAL

Given the measurements collected over multiple views of a specific location, we propose a two-step pipeline: (1) constructing a multiview graph (MVG) that represents the geographical features of the location, then (2) employing graph matching for location retrieval. This method is illustrated in Fig. 2. It starts by collecting k measurements for each of the four views from the current location. The next two consecutive steps are: (1) from the collected measurements, an MVG G is constructed for the probe location, and (2) a collaborative-reconstruction based MVG matching is executed to identify the best match $\tilde{G}^*$ of graph G from the fingerprint gallery.

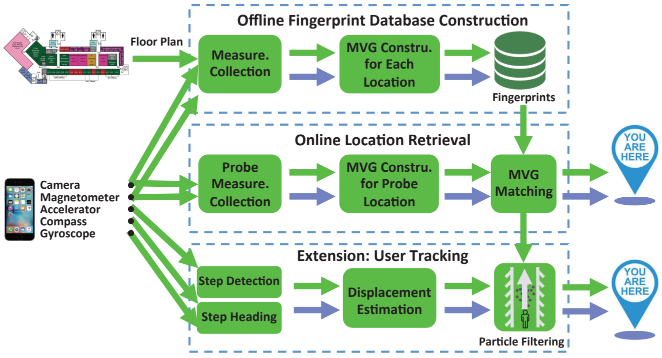


Figure 1: An overview of our system. 'Measure.' is short for 'measurement' and 'Constru.' is short for 'construction'.

4.1 Multiview Graph (MVG) Construction

From the multiview measurements collected for a specific location, we aim to construct an MVG capturing local features as well as global attributes of the location. A classical graph $\hat{G} = \{\hat{V}, \hat{E}\}$ consists of a node set $\hat{V}$ and an edge set $\hat{E}$. Likewise, an MVG $G = \{C, E\}$ consists of a set C of view clusters and an edge set E. Further, a node in the MVG is a view cluster, which is essentially the measurement set of a distinct view. The node set and edge set of an MVG are constructed as below.

Node set construction. Each node is a view cluster C_s, which refers to the measurement set $C_s = \{\mathbf{Y}_j\}_{j=1}^{K}$ collected for a particular view. Each measurement $\mathbf{Y}_j$ can be represented by the M-modality feature set $\mathbf{Y}_j = \{\mathbf{V}_{jm}\}_{m=1}^{M}$, where $\mathbf{V}_{jm}$ denotes the feature representation of $\mathbf{Y}_j$ in the m^{th} modality. Likewise, a view cluster C_s can also be represented by the M-modality feature set $C_s = \{\mathbf{F}_{sm}\}_{m=1}^{M}$, where $\mathbf{F}_{sm}$ denotes the feature representation of all the measurements in C_s in the m^{th} modality. After constructing the view cluster C_s for each view, the node set can be simply built as $C = \{C_s\}_{s=1}^{N}$. As shown in Fig. 2, we set four views ($N = 4$) for each location.

Edge set construction. As the N view clusters capture multiview and multi-modal local features, edges are further attached to pairs of view clusters, with the following motivations. (1) To our knowledge no one has studied the relationship between different views for the indoor localization task before, which we consider as both interesting and potentially useful. (2) Some specific locations may exhibit distinct edge structures, e.g., the four view clusters of a crossing location may be similar to each other, while for a location of a corridor, usually only the front and back view clusters are similar. Therefore, edges are attached as follows: an edge is established between two view clusters (i.e., two nodes) if their distance is no larger than a threshold τ. Given two view clusters C_s & $\tilde{C}_s$, we utilize the average distance between the measurements of C_s and $\tilde{C}_s$ to measure their distance. To avoid complex computations caused by a dense graph, threshold τ is set as μ, which is the mean distance between any two view clusters in the fingerprint gallery.

4.2 Our MVG Matching Scheme

Given the MVG G constructed for an unknown location, our MVG matching scheme is carried out for identifying the best match of

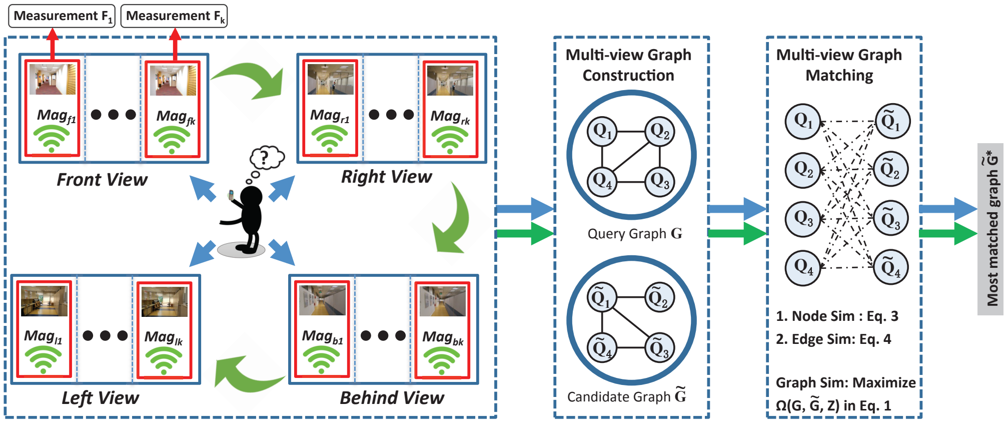


Figure 2: An illustration of the location retrieval pipeline that contains two steps: (1) MVG construction, and (2) MVG matching.

G in the fingerprint gallery. We start by defining the similarity measure between two MVGs.

Given the probe MVG G and a candidate MVG $\tilde{G}$, we define the similarity between G and $\tilde{G}$ as the optimal similarity $\Omega(G, \tilde{G}, \mathbf{Z})$ that considers both node-to-node (i.e., view-to-view) correspondence ($\mathbf{Z}$) and their graph structures ($G/\tilde{G}$). To evaluate graph structure similarity between G and $\tilde{G}$, we consider both node-wise and edge-wise similarities. Since the method to calculate node-wise and edge-wise similarities will be introduced in Subsection 4.3, here we assume they are already known. Let $\mathbf{A}^C \in R^{N\times N}$ and $\mathbf{A}^E \in R^{T_1\times T_2}$ denote the node-wise similarity and edge-wise similarity matrices, respectively. More specifically, in matrix $\mathbf{A}^C$, element $a^C_{c_1\tilde{c}_1}$ is the node-wise similarity score between node c_1 in G and node $\tilde{c}_1$ in $\tilde{G}$. In the same manner, let $a^E_{e_1\tilde{e}_1} \in \mathbf{A}^E$ be the edge-wise similarity score between edge e_1 in G and edge $\tilde{e}_1$ in $\tilde{G}$. Now the MVG matching problem can be formulated as finding the optimal correspondence between G and $\tilde{G}$ such that their similarity is maximized, which is the solution to the following constraint quadratic optimization problem

$$\max_{\mathbf{Z}} \; \Omega(G, \tilde{G}, \mathbf{Z}) = \sum_{c_1, \tilde{c}_1} z_{c_1\tilde{c}_1} a^C_{c_1\tilde{c}_1} + \sum_{\substack{e, \tilde{e} \\ e = s_1 s_2 \\ \tilde{e} = \tilde{s}_1\tilde{s}_2}} z_{s_1\tilde{s}_1} z_{s_2\tilde{s}_2} a^E_{e\tilde{e}} \qquad (1)$$

$$s.t. \quad \mathbf{Z}\cdot\mathbf{1} = \mathbf{1} \quad and \quad \mathbf{Z}^T\cdot\mathbf{1} = \mathbf{1},$$

where matrix $\mathbf{Z} \in \{0, 1\}^{N\times N}$ denotes the node correspondence, i.e., $z_{c_1\tilde{c}_1} = 1$ if node (i.e., view cluster) c_1 in G corresponds to node $\tilde{c}_1$ in $\tilde{G}$ and $z_{c_1\tilde{c}_1} = 0$ otherwise. Edge $e = s_1 s_2$ represents e which connects nodes s_1 and s_2 in G. Likewise, $\tilde{e} = \tilde{s}_1\tilde{s}_2$ represents $\tilde{e}$ that links nodes $\tilde{s}_1$ and $\tilde{s}_2$ in $\tilde{G}$.

In Eq. (1), the first term of $\Omega(G, \tilde{G}, \mathbf{Z})$ measures the node similarity between G and $\tilde{G}$, while the second term considers the edge similarity. Since one node in G is to match with only one node in $\tilde{G}$, Eq. (1) is constrained by one-to-one matching constraints, that is, $\mathbf{Z}\cdot\mathbf{1} = \mathbf{1}$ and $\mathbf{Z}^T\cdot\mathbf{1} = \mathbf{1}$, where $\mathbf{1} = \mathbf{1}_N$.

Without loss of generality, given similarity matrices $\mathbf{A}^C$ and $\mathbf{A}^E$, the optimization problem of Eq. (1) is a quadratic programming (QP) problem with equality constraints, which can be easily solved by state-of-the-art QP methods such as [11, 21, 39]. This turns out to be a simpler problem in our setting: Since in our settings there are only four nodes ($N = 4$) in an MVG and these four view clusters are collected in a fixed (i.e., clockwise) order, there actually exist only four node-to-node correspondence possibilities between G and $\tilde{G}$ (once a view in G corresponds to one of the four views in $\tilde{G}$, the correspondence is determined). Therefore, Eq. (1) can be solved by straightforward enumeration over the N possibilities.

4.3 Node-wise and Edge-wise Similarities Between Two MVGs

In Eq. (1) we assume the node-wise similarity and edge-wise similarity matrices A^C and A^E are given. Here we formally define the node-wise and edge-wise similarities between two MVGs. Different from previous methods, a collaborative-reconstruction MVG matching scheme is developed to explicitly deal with measurement outliers.

Given are a node (i.e., a measurement set for a view) C_s with its feature set $F_s = \{\mathbf{F}_{sm}\}_{m=1}^{M}$ from the probe MVG G, and the set of nodes $\{\tilde{C}_{\tilde{s}}\}_{\tilde{s}=1}^{N}$ with their feature sets $\tilde{F} = \{\tilde{\mathbf{F}}_m\}_{m=1}^{M}$ from a candidate MVG $\tilde{G}$. We model F_s in the m^{th} modality as a *hull*, i.e., $\mathbf{F}_{sm}\boldsymbol{a}$, where $\boldsymbol{a}$ is the coefficient vector and $\sum_{i=1}^{\eta} a_i = 1$, η is the number of measurements in C_s. Inspired by collaborative representation theory [37, 43], we reconstruct node C_s of G in the m^{th} modality with these multiple nodes of $\tilde{G}$ in the m^{th} modality. Formally, we reconstruct $\mathbf{F}_{sm}\boldsymbol{a}$ with $\tilde{\mathbf{F}}_m = \{\tilde{\mathbf{F}}_{\tilde{s}m}\}_{\tilde{s}=1}^{N}$ with the objective of minimizing reconstruction residual $R(\mathbf{F}_{sm}, \tilde{\mathbf{F}}_m, \boldsymbol{a}, \boldsymbol{b})$, as follows

$$\min_{\boldsymbol{a},\boldsymbol{b}} \sum_{m=1}^{M} \phi_m \|\mathbf{F}_{sm}\boldsymbol{a} - \tilde{\mathbf{F}}_m\boldsymbol{b}\|_2^2 + \gamma_1\|\boldsymbol{a}\|_{l_p} + \gamma_2\|\boldsymbol{b}\|_{l_p} \qquad (2)$$

$$s.t. \sum_{i=1}^{\eta} a_i = 1,$$

where $\boldsymbol{b} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_N]$ is the coefficient vector for reconstruction, and $\boldsymbol{b}_{\tilde{s}}$ is the sub-vector of coefficients associated with the $\tilde{s}^{th}$ node $\tilde{C}_{\tilde{s}}$ in the candidate MVG $\tilde{G}$. ϕ_m denotes the weight of the m^{th} modality. $\gamma_1\|\boldsymbol{a}\|_{l_p}$ and $\gamma_2\|\boldsymbol{b}\|_{l_p}$ are the regularization terms.

Constraint $\sum a_i = 1$ is required by the *hull* definition [43] and can eliminate the trivial solution $\boldsymbol{a} = \boldsymbol{b} = \mathbf{0}$ [25].

The *hull* $\mathbf{F}_{sm}\boldsymbol{a}$ of the probe feature set F_s in the m^{th} modality is collaboratively represented by the feature sets $\tilde{F}$ of the candidate MVG in the m^{th} modality. Elements of $\boldsymbol{a}$ are the coefficients, each being associated with one of the samples (i.e., measurements) in $\mathbf{F}_{sm}$. As such, each of these samples makes its individual contributions in the final probe representation. By minimizing the distance between $\mathbf{F}_{sm}\boldsymbol{a}$ and $\tilde{\mathbf{F}}_m\boldsymbol{b}$, outliers in both $\mathbf{F}_{sm}$ and $\tilde{\mathbf{F}}_m$ will be assigned with very small coefficients [25]. Therefore, the influence of measurement outliers (noisy measurements) could be substantially reduced.

By minimizing Eq. (2), we can obtain the optimal coefficient vectors $\boldsymbol{a}^*$ and $\boldsymbol{b}^* = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{\tilde{s}}, \cdots, \boldsymbol{b}_N]$, where $\boldsymbol{b}_{\tilde{s}}$ is the subvector of coefficients associated with the $\tilde{s}^{th}$ node $\tilde{C}_{\tilde{s}}$ in $\tilde{G}$. Then the residual of reconstructing node C_s with an individual node $\tilde{C}_{\tilde{s}}$ can be leveraged to define the node-to-node similarity [25]:

$$Sim(C_s,\ \tilde{C_{\tilde{s}}}) = A^C_{C_s\tilde{C_{\tilde{s}}}} = exp\Big(-\sum_{m=1}^{M}\phi_m\|\mathbf{F}_{sm}\boldsymbol{a}^{\boldsymbol{*}} - \tilde{\mathbf{F}}_{\tilde{s}m}\boldsymbol{b}^*_{\tilde{s}}\|_2^2\Big) \quad (3)$$

Since an edge can be considered as a pair of nodes, the edge-to-edge similarity is defined according to node-to-node similarity. Namely, for an edge $e = C_{s_1}C_{s_2}$ in G (e links nodes C_{s_1} and C_{s_2}), and an edge $\tilde{e} = \tilde{C}_{\tilde{s}_1}\tilde{C}_{\tilde{s}_2}$ in $\tilde{G}$, if C_{s_1} and C_{s_2} in G correspond to $\tilde{C}_{\tilde{s}_1}$ and $\tilde{C}_{\tilde{s}_2}$ in $\tilde{G}$, respectively, then the edge-to-edge similarity can be formulated as:

$$Sim(e, \tilde{e}) = A^E_{e,\tilde{e}} = \frac{1}{2}\Big(Sim(C_{s_1}, \tilde{C}_{\tilde{s}_1}) + Sim(C_{s_2}, \tilde{C}_{\tilde{s}_2})\Big) \quad (4)$$

However, if C_{s_1} and C_{s_2} do not correspond to $\tilde{C}_{\tilde{s}_1}$ and $\tilde{C}_{\tilde{s}_2}$, respectively, $Sim(e, \tilde{e}) = 0$. We now develop the solution of minimizing Eq. (2) in the next subsection.

4.4 Optimization

From Eq. (2), we focus on the l_2-norm setting (i.e., $p = 2$), and with a sequence of derivations as shown below, arrive at a *closed-form solution*. To achieve this, the convex optimization problem of Eq. (2) is first transformed into the Lagrangian function:

$$\begin{aligned} L(\boldsymbol{a}, \boldsymbol{b}, \lambda) = &\sum_{m=1}^{M}\phi_m\|\mathbf{F}_{sm}\boldsymbol{a} - \tilde{\mathbf{F}}_m\boldsymbol{b}\|_2^2 \\ &+\gamma_1\|\boldsymbol{a}\|_2 + \gamma_2\|\boldsymbol{b}\|_2 + \lambda(\mathbf{e}\boldsymbol{a} - \mathbf{1}) \end{aligned} \quad (5)$$

where $\mathbf{e}$ is a row vector whose elements are 1. Moreover, Eq. (5) is equivalent to

$$\begin{aligned} L(\boldsymbol{a}, \boldsymbol{b}, \lambda) = &\sum_{m=1}^{M}\phi_m\left\|[\mathbf{F}_{sm}\ \ -\tilde{\mathbf{F}}_m]\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix}\right\|_2^2 \\ &+ [\boldsymbol{a}^T\ \ \boldsymbol{b}^T]\begin{bmatrix}\gamma_1\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \gamma_2\mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} + \lambda([\mathbf{e}\ \ \mathbf{0}]\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} - 1) \end{aligned} \quad (6)$$

Denote $\mathbf{x} = \begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix}$, $\mathbf{D}_m = [\mathbf{F}_{sm}\ \ -\tilde{\mathbf{F}}_m]$, $\mathbf{B} = \begin{bmatrix}\gamma_1\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \gamma_2\mathbf{I}\end{bmatrix}$ and $\mathbf{u} = [\mathbf{e}\ \ \mathbf{0}]^T$. Now the Lagrangian function of Eq. (6) becomes

$$L(\mathbf{x}, \lambda) = \mathbf{x}^T(\sum_{m=1}^{M}\phi_m\mathbf{D}_m^T\mathbf{D}_m)\mathbf{x} + \mathbf{x}^T\mathbf{B}\mathbf{x} + \lambda(\mathbf{u}^T\mathbf{x} - 1) \quad (7)$$

Next we evaluate the gradients at its zero value:

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{x}} = 2(\displaystyle\sum_{m=1}^{M}\phi_m\mathbf{D}_m^T\mathbf{D}_m)\mathbf{x} + 2\mathbf{B}\mathbf{x} + \lambda\mathbf{u} = 0 \\ \dfrac{\partial L}{\partial \lambda} = \mathbf{u}^T\mathbf{x} - 1 = 0 \end{cases} \quad (8)$$

By Eq. (8), the closed form solution to Eq. (5) is obtained:

$$\mathbf{x} = \begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} = \frac{\mathbf{P}^{-1}\mathbf{u}}{\mathbf{u}^T\mathbf{P}^{-1}\mathbf{u}}, \qquad \lambda = -\frac{2}{\mathbf{u}^T\mathbf{P}^{-1}\mathbf{u}} \quad (9)$$

with

$$\mathbf{P} = \sum_{m=1}^{M}\phi_m\mathbf{D}_m^T\mathbf{D}_m + \mathbf{B} \quad (10)$$

In summary, we have obtained the closed-form solution of Eq. (9) to the optimization problem of Eq. (2).

5 EXTENSION TO TRACKING WITH AUGMENTED PARTICLE FILTER

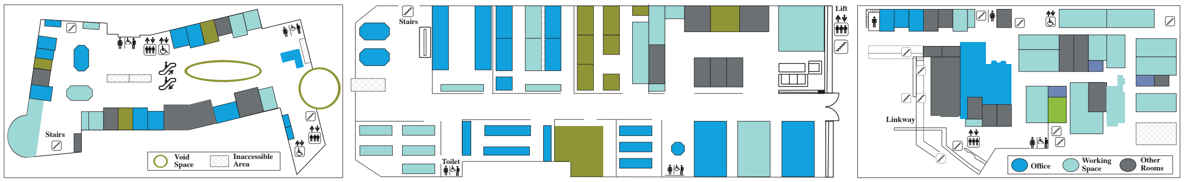
At this point we are able to locate the user via a one-time location retrieval. Next, we introduce an extension of *MviewGraph* with augmented particle filter to track the user location continuously over time. The key idea is to represent the location probability distribution with a set of particles, where each particle $O_i = \{p, w\}$ is endowed with a location estimation p as well as its weight w indicating the estimation confidence.

Particle Displacement. With the newly collected multiview measurements by the user, the new location of each particle is predicted as $p' = p + \Delta p$, where p is the last location estimation and Δp refers to the displacement computed from the IMU sensor readings. Since IMU sensor based step estimation have been well studied [6, 9, 19, 23, 34], we adopt the approach proposed in [6] to estimate the number of steps, step length and step headings. Leveraging this information and the floor plan, Δp can be effectively computed. The new location of a particle is then generated by sampling from a 2D Gaussian distribution centered at its predicted location p' to deal with sensor errors (noise).

Particle Weight Update. At the user's current location, consider G as the MVG constructed from the current multiview measurements, p as the location estimation in a particle O_i, and $\tilde{G}$ as the MVG of p in the fingerprints. The weight of particle O_i is then set to be the similarity between G & $\tilde{G}$, which is defined in Eq. (1). Particularly, if a particle hits a wall, its weight is set to 0.

Particle Resampling. We perform weight-based resampling over the entire set of particles. In brief, particles with higher weights will be sampled more often than those of low weights. This allows the elimination of those particles that are associated with very low weights, namely wrongly moved particles with very low confidence.

Location Estimation. The particle weights reflect the likelihood of the predicted location. There exist two common strategies to predict a final location: one is to directly use the location of the highest weighted particle, the other is to utilize the weighted average of the top 40% highest weighted particles. Empirically we find that the second method yields more stable and precise locations.



(a) Mall (b) Lab (c) Office

Figure 3: Floor plans of three buildings with different types.

6 EXPERIMENTS

6.1 Utilized Datasets

We carefully selected three complex and large indoor environments to acquire three real-world data sets for evaluation. These environments are of distinct types, which are a shopping mall (Mall), a laboratory (Lab), and an office building (Office). The areas of the environments are 10,731 m^2, 2,480 m^2, and 5,508 m^2, respectively.

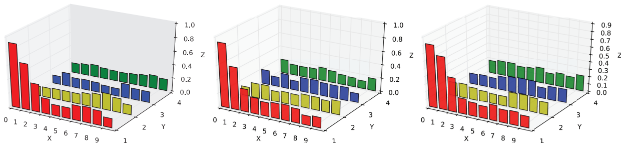
Fingerprint construction. Practically, a user is very likely to walk along the main direction of a path rather than walking towards the sides [26, 33], and thus we collect fingerprints every 1 m along the pathways inside each building. Note that there exist also pathways within rooms. In total, 1,632 PoIs (position of interests) are chosen as fingerprints where, for each PoI, four-view measurements are collected. For each view of a fingerprint PoI, the number of measurements range from 2 to 6. Overall 19, 217 multimodal measurements are taken to form the fingerprint gallery. We make one of the datasets (Lab) publicly available for the community to benchmark across different methods[1]. To our knowledge, this is the first multiview multimodal real-world indoor dataset that contains both magnetic and visual signals.

Feature extraction. Each measurement has two modalities: an image and its associated magnetic signal. For the images, the convolutional neural network (CNN) termed Places-CNN [42] is utilized to extract deep features, which gives rise to a 4,096 dimensional feature vector representing the image. The Euclidean norm termed *magnitude* is computed for each 3D magnetic signal, and then the *mean, mode* and *variance* of all the *magnitudes* are employed as the feature vector representing the magnetic signals.

Test data construction. In order to compare all the methods, 148, 120, and 272 test locations are randomly selected for Mall, Lab, and Office, respectively. These test locations were scattered all over the three buildings. Specifically, some of the comparison methods (Magic, Pedes, and MviewPF) are tracking-based and to test them we have walked many trajectories in the buildings, making random turns. The duration of the walks ranged from 6 seconds to 3 minutes, while the starting points of the walks were randomly selected.

6.2 Experimental settings

Our *MviewGraph* prototype runs on an iPhone 6 platform, where extensive experiments are conducted in the three aforementioned complex indoor environments. The floor maps of the environments are illustrated in Fig. 3.



(a) Examp. Loc A (b) Examp. Loc B (c) Examp. Loc C

Figure 4: The pairwise MVG similarities between 3 test points and fingerprint locations at different distances to them.

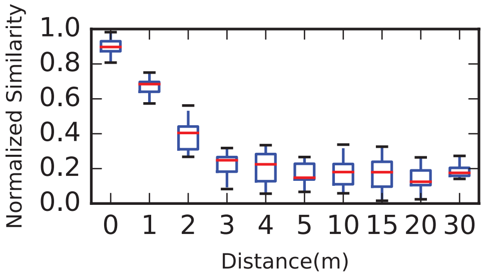


Figure 5: Pairwise MVG similarity w.r.t. distance.

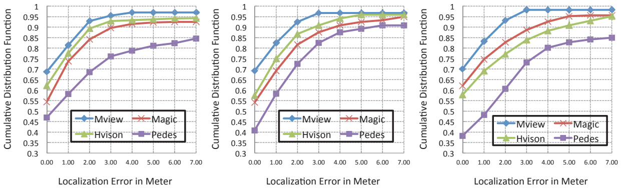
Workflow from the user's perspective. The user takes two photos for each of the four views, and then clicks 'locate me' to perform localization. The eight photos and their associated magnetic signals (i.e., eight measurements) are then sent to the *MviewGraph* server, which subsequently returns a location tag to the user. If the user selects continuous tracking, the IMU readings as well as recent trajectory data will be submitted to the server in addition to the above mentioned measurements.

Throughout our experiments, the sampling frequency is set to 30 Hz for the magnetometer, 30 Hz for the accelerometer, and 50 Hz for the gyroscope. A MacBook Pro was used as the server, equipped with an Intel Core i5 CPU at 2.6 GHz and 8GB of RAM.

6.3 Empirical Evaluation of MVG

In Section 4.1 we proposed a multiview graph (MVG) to represent a location, which is empirically evaluated in this section. The training set contains multi-view measurements collected over 1,632

[1] See the test set at https://archive.org/details/multiviewindoorimage2, and the training set at https://archive.org/details/MilabTestSetAMultiviewIndoorImageSetForALaboratory.



(a) Mall (b) Lab (c) Office

Figure 6: Performance comparison with representative existing methods in 3 different environments.

fingerprint locations. Our testset data is collected two weeks later, where 216 test locations are randomly selected. Our MVG matching scheme presented in Section 4.2 is used to compute similarity scores between a test location and all fingerprint locations. Empirically similar findings are observed over the 216 test locations. In what follows we will present three of them as examples, as well as describing the statistical analyses over all these 216 locations.

Fig. 4 shows the normalized MVG similarities between each of the three example locations and the fingerprint locations at different distances. Specifically, in Fig. 4(a), the first row plotted in red represents similarities between target location A and possible fingerprint locations with a range of 0-9m distance to A, with distance increasing from left to right in that row. The second row plotted in yellow demonstrates similarities between A and fingerprints with 10-19m distance to A, and so forth. The three subfigures of Fig. 4 suggest that MVG similarity scores between two locations drop dramatically with increasing distance, and the similarity score between two locations within a distance of 2 m is significantly higher than those where the two locations are farther apart. These two empirical properties demonstrate that the proposed MVG does possess a promising location resolution capability.

We further statistically investigate the 216 test locations. As presented in Fig. 5, a box plot is engaged to display the normalized pairwise MVG similarity score w.r.t. distance. Specifically, the red line inside a box denotes the median similarity score of a specific distance, while the bottom and top of a box show the first and third quartiles. Clearly the median normalized similarity at 0 m and 1 m distance are 0.90 and 0.67, while it is 0.39, 0.23, 0.21, 0.16, 0.18, 0.15, 0.19 for 2 m, 3 m, 4 m, 5 m, 10 m, 20 m and 30 m, respectively. This reconfirms the observation that the MVG similarity score between two locations that are within a distance of 2 m is significantly higher than those where the two locations are farther apart.

6.4 Comparison with Existing Methods

After empirically analyzing the effectiveness of MVGs, we now focus on performance comparison of our system w.r.t. existing methods. Among the possible comparison methods, we consider a pedestrian dead reckoning (Pedes) method [23], a handcrafted-vision-feature-based (Hvision) method [31], and a magnetic-field-based method (Magic) [33]. These methods are selected because they are of different types and are closely related to our work. For a fair comparison, in the Magic method [33] only the magnetic field is utilized (namely without WiFi). This is to ensure that all comparison methods are without additional infrastructure assistance.

Fig. 6 demonstrates comparison results over three complex indoor environments, where the x-axis depicts the localization error in meters and the y-axis denotes the cumulative error distribution function (CDF). Note that Mview is short for *MviewGraph*, and the location error of 0 m means that the location is tagged with the exact location label. Our first observation is that *MviewGraph* significantly and consistently outperforms the remaining methods in all three environments. We also note that over 80-percentile of errors occurs within 1 m for *MviewGraph*, which sheds light on the resulting meter level indoor localization precision of our approach.

We define *precision localization* as locations being positioned within 1 m error, and define *localization failure* as locations positioned with error over 8 m. The *precision localization ratio* (PLR) and *localization failure ratio* (LFR) for each comparison method are elaborated in the second and fifth rows of Table I, where 'w/' is shorthanded for 'with' and 'w/o' stands for 'without'. As shown in Table I, *ViewGraph* consistently outperforms the other methods in terms of higher *precision localization ratio* and lower *localization failure ratio*. For example, *ViewGraph* achieves an average ratio of 82.3% of *precision localization* (1 m accuracy) and 2.1% *localization failure* rate across all three environments of the dataset, while Pedes achieves 54.9% and 11.1%, Hvision 73.9% and 3.8%, and Magic 72.5% and 4.0%. We attribute the good performance to the capability of our MVG structure to capture more comprehensive multiview location features, especially when compared with the remaining methods.

6.5 Comparison in the Presence of Noise

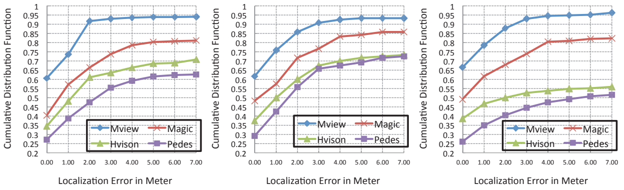
Up to this point, the measurements in the test set were collected with as little noise as possible. Here we conduct further experiments to examine whether similar performance can be achieved in the presence of measurements with large noise. There are a number of ways to introduce noise: (1) capture blurred, unclear, or human occluded photos; (2) take low resolution photos; (3) vibrate the phone when collecting magnetic signals; and (4) sway or move irregularly when engaging an IMU sensor. This leads to three new testsets, each for one of the three respective environments, where 30% of the measurements are collected with noise as mentioned above.

We examine 148, 120, and 272 locations of the Mall, Lab, and Office, respectively, with results summarized in Fig. 7. Again, *MviewGraph* clearly outperforms the remaining methods in the presence of considerable noise. More importantly, when comparing Fig. 7 with Fig. 6, we observe a significant deterioration of Magic, Hvision, and Pedes in the presence of 30% noisy measurements, while there is little influence for *MviewGraph*.

To quantify the performance deterioration rate, we also list in Table I the PLR (ratio of error ⩽ 1 m) and LFR (ratio of error > 8 m) in the presence and absence of large noise. Here the first and fourth rows display the PLR and LFR of all tested methods with (w/) the presence of 30% noisy measurements, respectively. The second and fifth rows demonstrate the PLR and LFR without (w/o) the presence of noise. The third row depicts the decline of PLR when large noise is introduced, while the sixth row (last

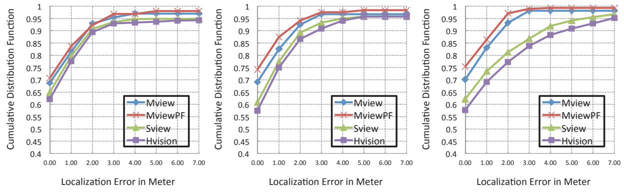
Table 1: Performance comparison with representative existing methods in terms of precision localization ratio (PLR) and localization failure ratio (LFR).

		Mall				Lab				Office			
		Mview	Magic	Hvision	Pedes	Mview	Magic	Hvision	Pedes	Mview	Magic	Hvision	Pedes
1	PLR w/ noise	**73.6%**	57.2%	48.1%	38.7%	**75.8%**	57.5%	50.0%	42.5%	**78.6%**	61.8%	46.7%	34.9%
2	PLR w/o noise	**81.4%**	73.7%	77.6%	58.2%	**82.5%**	69.2%	75.0%	58.3%	**83.1%**	74.6%	69.1%	48.2%
3	PLR Falloff	**7.8%**	16.5%	29.5%	19.5%	**6.7%**	11.7%	25.0%	15.8%	**4.5%**	12.8%	22.4%	13.3%
4	LFR w/ noise	**5.4%**	14.9%	23.4%	29.5%	**6.7%**	11.7%	20.8%	21.7%	**3.3%**	13.6%	35.3%	37.9%
5	LFR w/o noise	**2.7%**	5.9%	4.5%	12.0%	**1.7%**	4.2%	3.3%	9.2%	**1.5%**	1.8%	3.7%	12.1%
6	LFR Increase	**2.7%**	9.0%	18.9%	17.5%	**5.0%**	7.5%	17.5%	12.5%	**1.8%**	11.8%	31.6%	25.8%



(a) Mall (b) Lab (c) Office

Figure 7: Performance comparison with existing methods in the case of 30% noisy measurements.



(a) Mall (b) Lab (c) Office

Figure 8: Performance comparison with extension and reduction of MviewGraph.

row) lists the increase of LFR caused by the noise. As a result, the percentile of *precision localization* for *MviewGraph* reduces by 6.3% on average, with 25.6%, 13.7%, and 16.2% for Hvision, Magic, and Pedes, respectively. Likewise, the percentile of *localization failure* for *MviewGraph* increases by 3.2% on average, with 22.7%, 9.4%, and 18.6% for Hvision, Magic, and Pedes, respectively. These results reveal that *MviewGraph* is significantly more robust to noise than other methods, and also that large noise does have a significant negative impact on localization accuracy.

6.6 Comparison with Extension and Reduction of MviewGraph

We also compare *MviewGraph* with its extension and reduction. On the extension side, we add the augmented particle filter (introduced in Section 5) into *MviewGraph*, which is denoted as MviewPF. For reduction, the multiview settings in *MviewGraph* are removed, and only the single front-view measurements are used, which is denoted as Sview. Comparison results are illustrated in Fig. 8. The CDF curve of Hvision is also reported as a baseline. We observe that (1) MviewPF and *MviewGraph* perform significantly better than Sview and Hvision, and (2) MviewPF is slightly better than *MviewGraph*. Extending *MviewGraph* with a particle filter indeed improves the performance, since it incorporates trajectory contexts. It is worth noting that MviewPF is sensitive to noise since the particle filter requires leveraging the readings of IMU sensors.

7 DISCUSSION

- In empirical evaluation, we found that some locations exhibit unique signatures on one or more phone sensors. For example, a location may experience unusual geomagnetic signal readings, and an elevator may have a distinct pattern of influencing the phone's accelerometer. These semantic landmarks can be used to reset the error during user tracking.
- Studying the influence of each specific noise type will be of interest and may lead to novel understandings of the proposed and the existing methods.
- Although our current system does not rely on infrastructure, the system may benefit from spatial constraints or lightweight infrastructure assistance [23] where it is feasible. We will further explore these possibilities.

8 CONCLUSIONS

We have developed *MviewGraph*, an easy-to-use, accurate and robust indoor localization system for multiview and multimodal scenarios. *MviewGraph* enables a mobile-phone user to localize within one-meter accuracy without any infrastructure assistance, and even in the presence of large noise. The only requirement is to capture a few photos (e.g., 2 photos per view). We have designed a dedicated MVG structure to capture local features and global contexts of a location. We have also formulated the location retrieval problem as an MVG matching problem and derived a closed-form solution. For future work, we will investigate the incorporation of semantic landmarks and spatial constraints in localization and study the influence of each particular noise type.

REFERENCES

[1] Hamid Mohammed Ali and Alaa Hamza Omran. 2015. Floor Identification Using Smartphone Barometer Sensor for Indoor Positioning. *IJESRT* 4, 2 (2015), 384–391.

[2] Firas Alsehly, Tughrul Arslan, and Zankar Sevak. 2011. Indoor positioning with floor determination in multi story buildings. In *Proc. Int. Conf. on Indoor Positioning and Indoor Navigation*. 1–7.

[3] Yalong Bai, Kuiyuan Yang, Wei Yu, Chang Xu, Wei-Ying Ma, and Tiejun Zhao. 2015. Automatic Image Dataset Construction from Click-through Logs Using

Deep Neural Network. In *Proceedings of the 2015 Annual ACM Conference on Multimedia Conference*. 441–450.

[4] Artur Baniukevic, Christian S. Jensen, and Hua Lu. 2013. Hybrid Indoor Positioning with Wi-Fi and Bluetooth: Architecture and Performance. In *Proc. IEEE 14th Int. Conf. on Mobile Data Management*. 207–216.

[5] Igor Bisio, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone. 2013. GPS/HPS-and Wi-Fi Fingerprint-Based Location Recognition for Check-In Applications Over Smartphones in Cloud-Based LBSs. *IEEE Trans. on Multimedia* 15, 4 (2013), 858–869.

[6] Agata Brajdic and Robert Harle. 2013. Walk detection and step counting on unconstrained smartphones. In *The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13*. 225–234.

[7] Daniel Carrillo, Victoria Moreno, Benito beda, and Antonio F. Skarmeta. 2015. MagicFinger: 3D Magnetic Fingerprints for Indoor Location. *sensors* 15, 7 (2015), 17168–17194.

[8] Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N. Padmanabhan. 2010. Indoor localization without the pain. In *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking, MOBICOM 2010*. 173–184.

[9] Dae-Ki Cho, Min Y. Mun, Uichin Lee, William J. Kaiser, and Mario Gerla. 2010. AutoGait: A mobile platform that accurately estimates the distance walked. In *Eigth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2010*. 116–124.

[10] Nikolaos D. Doulamis and Anastasios D. Doulamis. 2014. Semi-supervised deep learning for object tracking and classification. In *Proc. IEEE International Conference on Image Processing*. 848–852.

[11] Steven Gold and Anand Rangarajan. 1996. A Graduated Assignment Algorithm for Graph Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* 18, 4 (1996), 377–388.

[12] Tao Guan, Yunfeng He, Juan Gao, Jianzhong Yang, and Junqing Yu. 2013. On-Device Mobile Visual Location Recognition by Integrating Vision and Inertial Sensors. *IEEE Trans. on Multimedia* 15, 7 (2013), 1688–1699.

[13] Janne Haverinen and Anssi Kemppainen. 2009. Global indoor self-localization based on the ambient magnetic field. *Robotics and Autonomous Systems* 57, 10 (2009), 1028–1035.

[14] Xiang He. 2015. Probabilistic Multi-Sensor Fusion Based Indoor Positioning System on a Mobile Device. *sensors* (2015), 31464–31481.

[15] Graham Healy and Alan F. Smeaton. 2009. Spatially Augmented Audio Delivery: Applications of Spatial Sound Awareness in Sensor-Equipped Indoor Environments. In *MDM 2009, Tenth International Conference on Mobile Data Management*. 704–708.

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proc. IEEE Int. Conf. on Computer Vision*. 2938–2946.

[17] Volkan Kilic, Mark Barnard, Wenwu Wang, and Josef Kittler. 2015. Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering. *IEEE Trans. on Multimedia* 17, 2 (2015), 186–200.

[18] John Kua, Nicholas Corso, and Avideh Zakhor. 2012. Automatic loop closure detection using multiple cameras for 3D indoor localization. In *Computational Imaging X, part of the IS&T-SPIE Electronic Imaging Symposium*. 82960V.

[19] Kun-Chan Lan and Wen-Yuah Shih. 2014. Using Smart-Phones and Floor Plans for Indoor Location Tracking. *IEEE Trans. Human-Machine Systems* 44, 2 (2014), 211–221.

[20] Howard Lee and Yi-Ping Phoebe Chen. 2015. Image based computer aided diagnosis system for cancer detection. *Expert Syst. Appl.* 42, 12 (2015), 5356–5365.

[21] Marius Leordeanu and Martial Hebert. 2005. A Spectral Technique for Correspondence Problems Using Pairwise Constraints. In *10th IEEE International Conference on Computer Vision (ICCV*. 1482–1489.

[22] Binghao Li, Thomas Gallagher, Andrew G. Dempster, and Chris Rizos. 2012. How feasible is the use of magnetic field alone for indoor positioning. In *Proc. Int. Conf. on Indoor Positioning and Indoor Navigation*. 1–9.

[23] Fan Li, Chunshui Zhao, Guanzhong Ding, Jian Gong, Chenxing Liu, and Feng Zhao. 2012. A reliable and accurate indoor localization method using phone inertial sensors. In *The 2012 ACM Conference on Ubiquitous Computing, Ubicomp '12*. 421–430.

[24] Huaiyu Li, Xiuwan Chen, Guifei Jing, Yuan Wang, Yanfeng Cao, and Fei Li. 2015. An Indoor Continuous Positioning Algorithm on the Move by Fusing Sensors and Wi-Fi on Smartphones. *sensors* 15, 12 (2015), 31244–31267.

[25] Anan Liu, Weizhi Nie, Yue Gao, and Yuting Su. 2016. Multi-Modal Clique-Graph Matching for View-Based 3D Model Retrieval. *IEEE Trans. Image Processing* 25, 5 (2016), 2103–2116.

[26] Zhenguang Liu, Luming Zhang, Qi Liu, Yifang Yin, and Li Cheng. 2016. Fusion of Magnetic and Visual Sensors for Indoor Localization: Infrastructure-free and More Effective. *IEEE Trans. Multimedia* PP, 99 (2016). DOI: https://doi.org/10.1109/TMM.2016.2636750

[27] Carlos Lopez and Yi-Ping Phoebe Chen. 2006. Using object and trajectory analysis to facilitate indexing and retrieval of video. *Knowl.-Based Syst.* 19, 8 (2006), 639–646.

[28] Van Vinh Nguyen and Jong Weon Lee. 2012. A Hybrid Positioning System for Indoor Navigation on Mobile Phones using Panoramic Images. *TIIS* 6, 3 (2012), 835–854.

[29] Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012. Harvesting visual concepts for image search with complex queries. In *ACM Multimedia*. 59–68.

[30] Savvas Papaioannou, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2014. Fusion of Radio and Camera Sensor Data for Accurate Indoor Positioning. In *Proc. IEEE Int. Conf. on Mobile Ad Hoc and Sensor Systems*. 109–117.

[31] Nishkam Ravi, Pravin Shankar, Andrew Frankel, Ahmed M. Elgammal, and Liviu Iftode. 2006. Indoor Localization Using Camera Phones. In *Proc. 7th IEEE Workshop on Mobile Computing Systems & Applications: Supplement*. 1–7.

[32] Xu Shen, Xinmei Tian, Anfeng He, Shaoyan Sun, and Dacheng Tao. 2016. Transform-Invariant Convolutional Neural Networks for Image Classification and Search. In *Proceedings of the 2016 ACM Conference on Multimedia Conference*. 1345–1354.

[33] Yuanchao Shu, Cheng Bo, Guobin Shen, Chunshui Zhao, Liqun Li, and Feng Zhao. 2015. Magicol: Indoor Localization Using Pervasive Magnetic Field and Opportunistic WiFi Sensing. *IEEE Journal on Selected Areas in Communications* 33, 7 (2015), 1443–1457.

[34] Alan F. Smeaton, James Lanagan, and Brian Caulfield. 2012. Combining wearable sensors for location-free monitoring of gait in older people. *JAISE* 4, 4 (2012), 335–346.

[35] Chenshu Wu, Zheng Yang, and Yunhao Liu. 2015. Smartphones Based Crowdsourcing for Indoor Localization. *IEEE Trans. Mob. Comput.* 14, 2 (2015), 444–457.

[36] Han Xu, Zheng Yang, Zimu Zhou, Longfei Shangguan, Ke Yi, and Yunhao Liu. 2016. Indoor localization via multi-modal sensing on smartphones. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 208–219.

[37] Meng Yang, Lei Zhang, David Zhang, and Shenlong Wang. 2012. Relaxed collaborative representation for pattern classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2224–2231.

[38] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. 2016. Multilayer and Multimodal Fusion of Deep Neural Networks for Video Classification. In *Proceedings of the 2016 ACM Conference on Multimedia Conference*. 978–987.

[39] Yongqing Yang, Jinde Cao, Xianyun Xu, Manfeng Hu, and Yun Gao. 2014. A new neural network for solving quadratic programming problems with equality and inequality constraints. *Mathematics and Computers in Simulation* PP, 101 (2014), 103–112.

[40] Chi Zhang, Kalyan Subbu, Jun Luo, and Jianxin Wu. 2015. GROPING: Geomagnetism and cROwdsensing Powered Indoor NaviGation. *IEEE Trans. Mob. Comput.* 14, 2 (2015), 387–400.

[41] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM Multimedia*. 33–42.

[42] Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *Proc. NIPS*. 487–495.

[43] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang. 2014. Image Set-Based Collaborative Representation for Face Recognition. *IEEE Trans. Information Forensics and Security* 9, 7 (2014), 1120–1132.