Fusion of Magnetic and Visual Sensors for Indoor Localization: Infrastructure-Free and More Effective

Zhenguang Liu, Luming Zhang, Qi Liu, Yifang Yin, Li Cheng, and Roger Zimmermann

***Abstract*—Accurate and infrastructure-free indoor positioning can be very useful in a variety of applications. However, most existing approaches (e.g., WiFi and infrared-based methods) for indoor localization heavily rely on infrastructure, which is neither scalable nor pervasively available. In this paper, we propose a novel indoor localization and tracking approach, termed VMag, that does not require any infrastructure assistance. The user can be localized while simply holding a smartphone. To the best of our knowledge, the proposed method is the first exploration of fusing geomagnetic and visual sensing for indoor localization. More specifically, we conduct an in-depth study on both the advantageous properties and the challenges in leveraging the geomagnetic field and visual images for indoor localization. Based on these studies, we design a context-aware particle filtering framework to track the user with the goal of maximizing the positioning accuracy. We also introduce a neural-network-based method to extract deep features for the purpose of indoor positioning. We have conducted extensive experiments on four different indoor settings including a laboratory, a garage, a canteen, and an office building. Experimental results demonstrate the superior performance of VMag over the state of the art with these four indoor settings.**

***Index Terms*—Convolutional neural network, indoor localization, magnetic field, particle filter, visual image.**

I. Introduction

ACCURATE and reliable indoor positioning can be very useful in a variety of applications [1], [2]. Examples include localizing survivors inside a building in case of fire, guiding robots in a fully automated factory where there is no human presence on-site, navigating a person to a room in an unfamiliar building, and finding a parking space for a car in an underground structure.

However, precise indoor positioning is still an open challenge. In outdoor environments, GPS is commonly used for navigation. However, GPS can not provide accurate localization for indoor environments because satellite signals are likely to be blocked by walls and ceilings. WiFi, infrared and ultrasound based approaches have shown great promise in indoor positioning. However, due to the limited coverage of a single signal transmitter/receiver, these approaches heavily depend on specific infrastructure, which is both expensive and difficult to maintain. Moreover, it might be infeasible to deploy signal transmitters/receivers in certain buildings due to safety or privacy concerns.

In this paper we propose a novel indoor localization and tracking system termed VMag for smartphone users, free from any infrastructure assistance. To the best of our knowledge, our work is the first to integrate both magnetic and visual sensing for indoor localization. The motivation for utilizing magnetic and visual sensing is twofold. 1) Both visual images and the geomagnetic field are omnipresent across the globe. They can be conveniently captured by common sensors of a smartphone. 2) Images and the geomagnetic field are complementary in indoor localization because images are usually distinguishable across distant locations while magnetic signals are known to be more locally distinctive [1], [3], [4].

Toward an infrastructure-free and accurate solution for indoor localization utilizing visual, magnetic and inertial sensors of a smartphone, three main challenges are yet to be addressed as follows.

Low resolution of magnetic sensor readings: The magnetic field signal $\boldsymbol{m}$ measured by a smartphone magnetometer consists of three components $\langle m_x, m_y, m_z \rangle$, which correspond to the magnetic intensities (in units of μT) in x, y and z directions, respectively. The low dimensional vector $\boldsymbol{m}$ is usually not reliable to form a unique location signature [1].

Noisy sensor readings: Sensor noise is unavoidable primarily due to hasty movements of a smartphone user and the inherent bias of different smartphone sensors [5]. Specifically, gyroscopes and accelerometers may exhibit noisy sensor readings due to user sway and movement irregularities.

Diverse walking patterns: It is well known that users of different physical attributes such as gender, age, and height may

Manuscript received August 30, 2016; revised November 4, 2016; accepted November 30, 2016. Date of publication December 7, 2016; date of current version March 15, 2017. This work was supported by Singapore's Ministry of Education Academic Research Fund Tier 1 under Grant T1 251RES1415, by the National Natural Science Foundation of China under Grant 61572169 and Grant 61472266, by the National University of Singapore Suzhou Research Institute, Suzhou, and by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Liang Zhou.

Z. Liu, Q. Liu, Y. Yin, and R. Zimmermann are with the School of Computing, National University of Singapore, Singapore 117417 (e-mail: zhenguangliu@zju.edu.cn; qiliu@u.nus.edu; idmyiny@nus.edu.sg; rogerz@comp.nus.edu.sg).

L. Zhang is with the Department of Computer and Information, Hefei University of Technology, Hefei 230000, China, and also with the National University of Singapore Suzhou Research Institute, Suzhou 215123, China (e-mail: zglumg@gmail.com).

L. Cheng is with the Bioinformatics Institute, the Agency for Science Technology and Research, and the School of Computing, National University of Singapore, Singapore 117417 (e-mail: chengli@bii.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2636750

possess different gait patterns [6], [7]. During user tracking, the gait pattern, especially the step length, is very informative in predicting a person's new location. Using a generic gait model to estimate the step length of a specific user may introduce non-negligible errors.

The key contributions of the introduced solution can be summarized as follows. 1) We propose to fuse geomagnetic field measurements and visual sensing for the purpose of indoor positioning. 2) We develop a novel context-aware particle filter approach as opposed to the typical first-order Markov assumption of standard particle filtering. Instead of presuming that the location X_t at time t depends only on location X_{t-1} of the immediately previous time step $t-1$, we further incorporate available contextual information such as past traces, the latest measurements, and floor plans of a building. 3) Most existing visual based methods for indoor localization leverage handcrafted image features such as color histograms and SIFT [8], [9]. In contrast, we incorporate deep learning methods to extract deep features from empirical data for the indoor localization task.

Extensive experiments have been systematically carried out to evaluate the effectiveness of our proposed VMag system. Experiments with four different indoor settings that cover a wide range of situations (a research laboratory, a garage, a canteen and an office) demonstrate that VMag achieves a promising performance where 91% of the localization errors occur within 0.85 m, 1.34 m, 1.34 m and 0.85 m, respectively, for the four typical settings.

Assumptions: VMag focuses on accurate indoor tracking of a smartphone user free from any infrastructure assistance. The only requirement is for a user to hold her smartphone in a vertical portrait position and capture images only along the main direction of her path (which should be easy for the user and maximize localization accuracy). VMag follows the commonly adopted (e.g., [10]–[12]) fingerprint-based framework of infrastructure-free indoor localization systems, thus fingerprints need to be collected before performing localization. Moreover, since VMag focuses on user tracking, we only consider all the possible paths in buildings rather than covering all possible locations. Note that there are paths even within rooms. The two main opposing directions of a path are treated as two different paths. For a crossing, we collect the fingerprints along the main directions of each of the paths that intersect at the crossing, respectively.

II. Empirical Evaluation on Strengths and Challenges of Geomagnetic Field and Visual Images

Before introducing VMag, we conduct an in-depth empirical evaluation of the strengths and challenges of visual and magnetic measurements for indoor localization.

A. Strengths of Magnetic Field

Our empirical evaluation suggests that the properties of a geomagnetic field which are advantageous for the purpose of indoor localization are as follows.

1) *Drastic Geomagnetic Field Changes Across Locations:* To study the variability of the geomagnetic field in space, we have measured the geomagnetic field $\boldsymbol{m} = \langle m_x, m_y, m_z \rangle$ intensities while walking in a corridor. Fig. 1 illustrates the measured geomagnetic field intensities. Clearly the magnetic field intensities change drastically across locations, and there exist many local anomalies in the magnetic field due to local disturbances. This re-confirms prior reports of similar observations in the literature (e.g., [1], [4]),

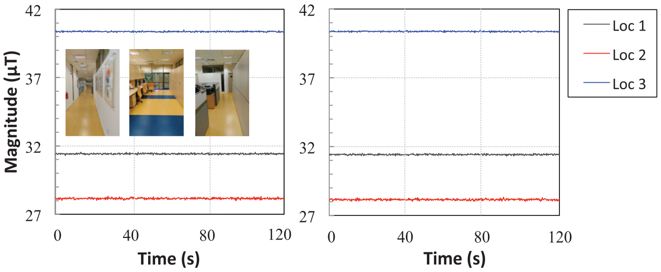


Fig. 1. Drastic changes of the magnetic field $\boldsymbol{m} = \langle m_x, m_y, m_z \rangle$ across locations. Values of $\boldsymbol{m}$ were measured while walking in a corridor.

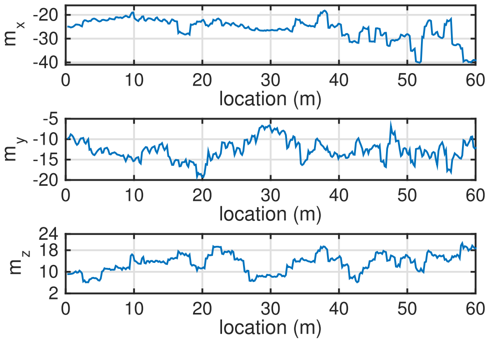


Fig. 2. Magnetic field at three different locations. The magnitudes were measured at two different times, four weeks apart.

2) *Stable Geomagnetic Field:* To evaluate the stability of the geomagnetic field across locations, we have measured the magnetic field $\boldsymbol{m}$ at 20 randomly selected locations in three different buildings of NUS (the National University of Singapore). The first round of measurements was acquired on 21 February 2016, which was followed by a second round of repeated measurements four weeks later. Since similar magnetic field properties were observed for the 20 locations, we illustrate the magnetic field signals of three locations as examples to show the findings. The left panel of Fig. 2 shows the field magnitudes $||\boldsymbol{m}||$ (the Euclidean norm of magnetic field $\boldsymbol{m}$) measured on 21 February 2016, while the right panel displays the magnitudes measured four weeks later. The figure clearly shows that the field magnitudes at the same location at two different times are almost the same. The stability of the magnetic field has been similarly reported in the literature (e.g., [1], [4]).

3) *Limited Influence of Common Objects on the Geomagnetic Field:* Existing literature [1], [3] has reported that moving objects such as cars, lifts, trolleys and people bear little influence on the geomagnetic field values a few meters away. Here we instead focus on the influence from common indoor objects such as turning on/off computers, printers, and refrigerators. We continuously measured the magnitude while turning on/off

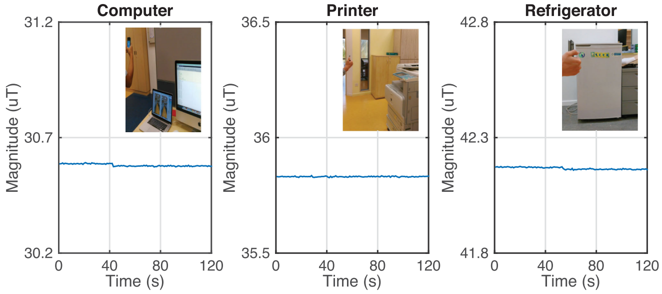


Fig. 3. Influence of turning on/off indoor equipment at a distance of 1.5 m.

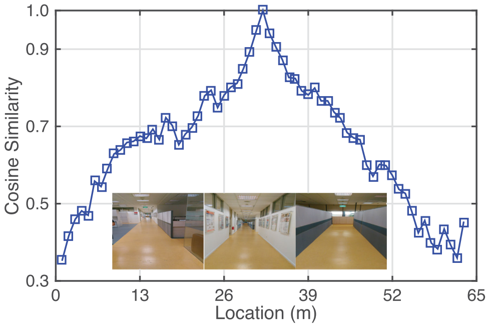


Fig. 4. Average similarity between the ith image of a path and the midpoint image of the path. The average is calculated based on 13 different paths from three different buildings.

a computer, a printer and a refrigerator, respectively. We varied the distance between the measuring location and the equipment from 0 m to 3 m with an interval of 0.5 m, and found that the influence of turning common equipment on/off is negligible when the distance goes beyond 1.5 m. Fig. 3 shows the change of magnitude at a distance of 1.5 m.

B. Strengths of Visual Images

The inspiration of considering visual data for localization purposes stems from the fact that human beings can often easily identify locations at first glance. Our empirical evaluation suggests a plethora of strengths of visual images as demonstrated below. It is noteworthy to mention that in this experimental study, each collected image is transformed into a 4096 dimensional vector using a convolutional neural network (i.e., the Places-CNN network proposed in [13]). For illustration simplicity the similarity of two images is measured by the cosine similarity between their corresponding 4096 dimensional vectors.

1) *Distinct Visual Images of Different Locations:* To study the variability of images across locations, images were collected from every meter along 13 different paths of three different buildings. Let $\langle a_1^{(j)}, a_2^{(j)}, \cdots, a_m^{(j)} \rangle$ denote the sequential images collected on the jth $(1 \leqslant j \leqslant 13)$ path, and $a_{\mathrm{mid}}^{(j)}$ be the image collected at the mid-point location of the jth path. We calculate the similarity $s(a_i^{(j)}, a_{\mathrm{mid}}^{(j)})$ between each image $a_i^{(j)}$ and the mid-point image $a_{\mathrm{mid}}^{(j)}$ of the same path. Then the average similarity $\frac{1}{13}\sum_{j=1}^{13} s(a_i^{(j)}, a_{\mathrm{mid}}^{(j)})$ is computed to reflect the similarity between the ith image and the mid-point image of a same path. Fig. 4 shows the empirical results. Consistent with human experience, Fig. 4 suggests that images at different locations are often distinct, while images of nearby locations are usually more similar than those farther away.

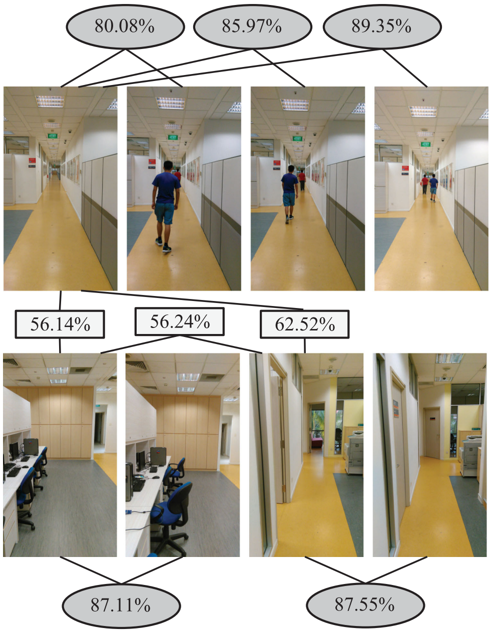


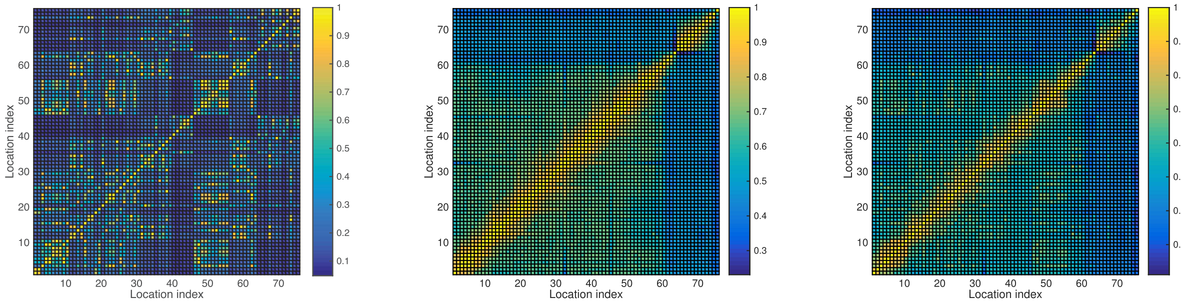
Fig. 5. Four images in the upper row were taken while people were walking in a corridor. The four images in the lower row were taken at two different times to evaluate the influence of displaced chairs and closed doors. The percentage values denote the similarity between two images.

Note that since floor identification is a well-studied topic (e.g., [14], [15]), here we only focus on studying the variability of images from the same floor.

2) *Stable Visual Properties:* In order to study the stability of the visual properties for indoor locations, we have collected images from 20 distinct locations on 22 February 2016 and 14 March 2016, respectively. Subsequently we calculated the similarity between every pair of images. Experimental results show that the similarities between two images, which were collected at the same location but at different times, are all higher than 87.05%. In contrast, the similarities between two images that were collected at different locations are all lower than 63.36%.

3) *Limited Influence of Mobile Objects on Visual Images:* It is of interest to study the visual influence of mobile objects, such as a walking person, open and closed doors, and the displacement of objects. Our experiments were carried out as follows. *First*, four images of a corridor with walking people were collected as displayed in the upper row of Fig. 5. The similarity between each occluded image and its original image is demonstrated in the three ellipses that are shown above the images. For comparison, we also computed the similarities between the images from two different locations, which are illustrated in the three rectangular labels of Fig. 5. We can see that the occluded and unoccluded images of the same location are still significantly more similar than those of different locations.

Second, we collected a pair of images for a location at the start and end of a time interval during which the chairs in the image were displaced. The two images are shown on the left of the lower row of Fig. 5. We also collected another pair of images for a location where the two doors in the image have been opened



(a) (b) (c)

Fig. 6. Geomagnetic field and image similarity between all pairs of locations. Locations 1–63 are from a corridor and locations 64–75 are from another corridor. (a) Geomagnetic field similarity. (b) Image similarity. (c) Fusion similarity.

and closed, respectively. The images are shown as the right two images of the lower row of Fig. 5. The similarities of each pair of images are shown in the two ellipses below the figures. We can see that the perturbed images of the same location are still significantly more similar than those from different locations. We have conducted analogous experiments at 19 other locations, which provided similar results.

C. Fusion of Geomagnetic Field Signals and Visual Images

The fundamental reason that geomagnetic field signals and visual images can be combined for indoor localization lies in their complementary nature. Visual images of nearby locations tend to be significantly more similar than those from remote locations, while remote locations may have a similar magnetic field signal but nearby locations may have a different one. Fig. 6 provides an exemplar illustration of this phenomenon.

The images and geomagnetic fields of two corridors were collected. The location indices of the first corridor are 1–63 and those of the second corridor are 64–75. Fig. 6(a) and 6(b) show the magnetic field and image similarities between each pair of locations, respectively. Fig. 6(c) depicts the fusion similarity between each pair of locations, where the fusion similarity is obtained as a linear combination of the magnetic field and image similarities. In Fig. 6(a), the similarity between two magnetic fields $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ is measured by $1/(1 + ||\boldsymbol{m_1} - \boldsymbol{m_2}||)$.

From Fig. 6(a), we observe that remote locations may have a similar magnetic field signal while nearby locations may have a different one. From Fig. 6(b), we can see that images of nearby locations are usually significantly more similar than those of remote locations. In Fig. 6(c), the yellow area around the diagonal indicates the high probability locations where the measurements may come from. We observe that the yellow area in Fig. 6(c) is much more narrow compared to Fig. 6(b), which implies that the high probability locations are much more concentrated after fusion.

D. Challenges

The first challenge concerns the low resolution of magnetic measurements: The measured magnetic field signal $\boldsymbol{m}$ is usually not reliable enough to form a unique location signature due to its low dimensionality [1]. In order to address this shortcoming, we first add magnitude $||\boldsymbol{m}||$ as a new dimension of the magnetic field, and then combine the magnetic field with visual images to form a more unique location signature.

The second issue is noisy sensor readings: Sensor noise is almost inevitable. In order to deal with the noise of inertial sensors, we adopt a probabilistic model, which employs particle filtering. In other words, we maintain a set of potential user locations and their associate probabilities instead of only one certain location.

The third issue relates to diverse gait patterns: It is well understood that users of different physical profiles exhibit different gait patterns [6]. Always using a generic gait model to estimate the step length of a specific user may produce notable errors in user location prediction. Thus we propose to start with a generic gait model initially and then train personalized gait models dedicated to individuals as more data becomes available.

III. VMag System

After the empirical evaluation of the strengths and challenges of both visual images and the geomagnetic field, we are now ready to present our VMag system. Throughout this paper, a measurement for a location is defined as the image and the magnetic field collected by a smartphone (using the smartphone camera and magnetometer) at that location.

A. Overview of VMag System

VMag consists of two phases, i.e., offline preparation and online positioning. In the offline preparation, the measurement fingerprints for each location are collected manually using smartphones and then stored at the server end. The floor plan of the building is also collected and stored, and the positioning models (on the server side) are built and trained. All these operations of the offline phase can be finished in advance, so they do not consume any time during the online positioning.

In online positioning, a user holds her phone, in which the VMag application is installed, in a portrait orientation and points it in the direction she is walking (i.e., along a path). After each step (step detection has been well studied and we adopt the step detection approach proposed in [6]), the VMag application takes an image and measures the magnetic field at the current location. The image and magnetic field constitute a measurement and are sent to the server for localization. The motion data (e.g., the

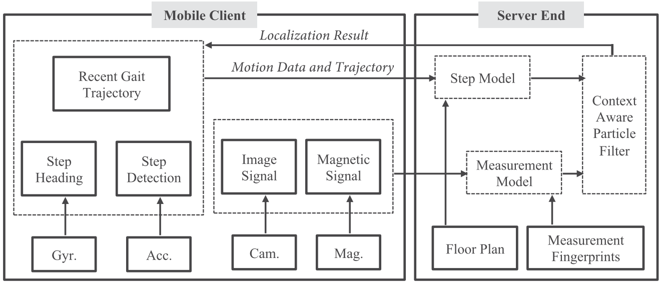


Fig. 7. Architecture of VMag system.

detected step heading) and the trajectory data are also sent to the server to support tracking of the user.

Fingerprint collection in the offline preparation: A fingerprint is a pair of a measurement and its corresponding location. To collect the fingerprints for each location, VMag divides all spaces of a building into a grid of 60 cm × 60 cm cells. Each grid cell is regarded as a unique location and assigned a location label. Recall that there are paths even in the rooms. For each location, we randomly collect measurements 10 times inside its corresponding grid cell while holding the smartphone vertically and pointing into the main direction of the path. The 10 measurements are termed the measurements of the location. All the collected measurements and their associated location labels form the fingerprint map.

The overall architecture of VMag is depicted in Fig. 7. VMag consists of mobile clients and a server. Each mobile client collects the measurements and the motion data and performs trajectory data logging. Motion data mainly includes step headings detected by using the readings of the smartphone gyroscope and accelerometer. The trajectory data logging module keeps a sliding window of the most recent gait information to facilitate user tracking.

The *server end* consists of three major components, namely the step model, the measurement model and the context aware particle filter (also referred in short as CPF). The step model estimates the step length and heading of the user. The measurement model transforms a measurement into a deep feature vector and defines the *measurement similarity* between measurements. The CPF tracks the user by utilizing a set of particles to simulate the probability distribution of the potential user locations. Once a user step is detected, CPF invokes the step model to predict the new location of the user and then revises the prediction in light of the new measurement obtained at the current location by processing it through the measurement model. The details of the three components are introduced below.

B. Step Model

The step model estimates the step length and heading utilizing the motion data, the user trajectory and the floor map of the building. The step model is primarily responsible for estimating the location displacement after one user step in CPF user tracking.

Step detection and step heading detection have been well studied and we adopt the step detection approach proposed in [6]. Since observations show that a user is very likely to walk along the main direction of a path rather than perpendicular to it [1], we select the direction among all the main directions of the paths that are closest to the detected step heading as the final step heading.

For the *step length estimation*, VMag uses an initial generic step length (which is 60 cm) at the very beginning and then trains a personalized step length when more personal data has been collected. Formally, let $\langle P_{t-k}, P_{t-k+1}, \cdots, P_{t-1} \rangle$ be the most recent gait trajectory, where P_i is the estimated user location after the ith step, and let $||P_i - P_{i-1}||$ be the Euclidean distance between locations P_i and P_{i-1}. First, VMag calculates the continuity of the recent gait trajectory. The trajectory is continuous iff $||P_i - P_{i-1}|| \leqslant \tau$ for all $t-k+1 \leqslant i \leqslant t-1$. Here τ is the maximum possible step length of a user, which is set to 2 m in our settings. Second, step length ℓ is estimated as follows.

1) If the trajectory is continuous, ℓ is estimated as the weighted average of distances between P_i and P_{i-1}. Precisely

$$\ell = \sum_{i=t-k+1}^{t-1} w_i ||P_i - P_{i-1}|| \tag{1}$$

where $w_i = \frac{m_i}{\sum_{i=t-k+1}^{t-1} m_i}$ and $m_i = \exp\left(-\frac{(t-i)^2}{8}\right)$. The setting of w_i is based on the idea that a more recent step length $||P_i - P_{i-1}||$ should have a higher weight w_i.

2) If the trajectory is not continuous, VMag organizes the locations in the trajectory into location communities and then tries to find which one is the right community. To this aim, VMag first constructs an undirected and unweighted graph G. The vertexes of G are the locations of the trajectory. The edges of G are determined in the following manner. If the distance $||P_i - P_j|| \leqslant \tau \, (t-k+1 \leqslant i, j \leqslant t-1)$, then we add an edge between P_i and P_j, otherwise the edge is omitted. Subsequently VMag conducts a connected component analysis in graph G. The connected component C that has the maximum number of locations is regarded as the right location community, and the weighted average distance between each pair of continuous locations in C is defined as the estimation of step length ℓ.

C. Measurement Model

The measurement model transforms a measurement into a deep feature vector and defines the similarity between measurements. Both the deep feature vector extraction for a measurement and the measurement similarity calculation will be invoked by CPF during user tracking.

For the measurement feature extraction, we adopt a neural network to extract deep features. Given a measurement Z, the extraction process contains two steps. Step 1 transforms the image of Z into deep image features. Step 2 leverages a fully connected neural network, which fuses the obtained deep image features with the magnetic field signal of Z to extract the final deep feature vector.

Step 1: Deep image features extraction: We employ the convolutional neural network (CNN) termed Places-CNN to extract deep image features from the images. In order to make Places-CNN more suitable for our settings, we re-trained the

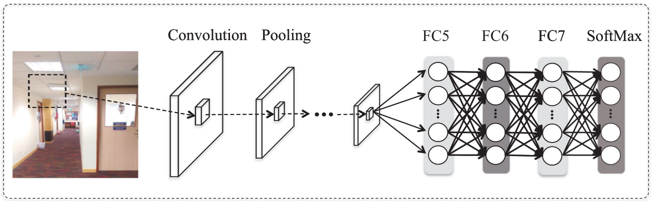


Fig. 8. Architecture of deep image features extraction. "FC" is short for "fully connected layer."

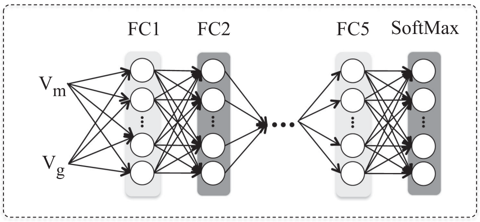


Fig. 9. Architecture of deep feature extraction. "FC" is short for "fully connected layer."

Places-CNN network on a new dataset that combines both the original Places dataset [13] and the images of our four collected indoor environments. The neural network architecture and training method are the same as those proposed in [13]. The reasons of employing the Places-CNN network in [13] are that Places-CNN is specially designed for scene recognition and achieves an astonishing performance on place classification tasks. The network architecture of Places-CNN is illustrated in Fig. 8 [13], [16]. In the Places-CNN model, the convolution and pooling layers capture image local connectivity. After the convolution and pooling layers, the output is then passed to three fully connected layers. We take the outputs of FC7 (fully connected layer 7) as the 4096-dimensional descriptor of the image. The 4096-dimensional descriptor is termed the *deep image features* and denoted as vector V_g.

Step 2: Fusing image and magnetic signals to extract deep features for the measurement: After extracting deep image features V_g, VMag further combines V_g with the magnetic field signal to obtain the final deep feature vector for the measurement. As illustrated in Fig. 9, VMag uses a neural network with five fully connected layers to extract the deep feature vector. The inputs are V_g and V_m, which are the deep image features and the magnetic vector of a measurement, respectively. The magnetic vector V_m is defined as $\langle m_x, m_y, m_z, ||m||\rangle$, where $||m||$ is the Euclidean norm of $\langle m_x, m_y, m_z\rangle$. The outputs are the 4,196-dimensional descriptor, which is the outputs of FC5 and termed the *deep feature vector* of the measurement.

The reasons that we use the outputs of FC5 rather than the outputs of the network (i.e., the outputs of the SoftMax layer) are as follows. As mentioned earlier, VMag divides all paths of a building into grid cells. Each grid cell is regarded as a unique location and assigned a location label. The outputs of the network are just the probabilities of all the locations where the input features may come from. In contrast, the outputs of the FC5 layer are the combined image and magnetic signals.

The details of training the neural network in Fig. 9 are as follows. We utilize all the fingerprints (i.e., pairs of a measurement and its corresponding location) to train the neural network. The sizes of the five layers of the network, from FC1 to FC5, are 8192, 7168, 6144, 5120, and 4196, respectively. The size of FC5 is set to 4,196 because this selection does not compress input features while capturing location information after the non-linear transformations of the hidden layers. At the same time, the experimental results further verify the effectiveness of this selection. Since the output layer emits the probability distribution of location labels, the size of the output layer equals the number of the location labels. We utilize standard cross entropy as the loss function. For parameter initialization, we adopt the Glorot-normal initialization in [17]. In the training, we adopt a standard back-propagation algorithm with stochastic gradient descent. The starting learning rate α is set to 0.1, and when no significant accuracy increase is observed after several epochs of training, we stop the training process.

After introducing the deep feature vector extraction for a measurement, we formally define the measurement similarity below.

Definition 1: measurement similarity: Let $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ be the deep feature vectors of measurements A and B, respectively. The measurement similarity between A and B is

$$s(A,B) = \exp\Big(-\frac{||\boldsymbol{v_1} - \boldsymbol{v_2}||}{2\sigma^2}\Big) \tag{2}$$

where σ is a parameter to adjust the impact of the distance.

IV. Tracking With Context-Aware Particle Filter

Our proposed CPF (context-aware particle filter) tracks the user on a per-step basis. After each user step, CPF updates the location estimation. Instead of using only one certain position to estimate the true location, CPF utilizes a probabilistic distribution to depict the potential locations of the user.

The objective of CPF is to estimate the posterior probability distribution of the user location given all the measurements up to the current time. Towards this, CPF implements a recursive Bayesian filter using Monte Carlo simulations. The core idea is to represent the posterior probability distribution with a set of particles and their associated weights [18]. The term recursive stems from the fact that the filter recursively estimates the posterior probability distribution at time t from the one at time $t-1$ [19].

Formally, let time t be the time right after the user's tth step, and X_t be the user location at time t, namely the user location after the tth step. Let $X_{t_1:t_2}$ be the sequential user locations from time t_1 to t_2, i.e., $X_{t_1:t_2} = \langle X_{t_1}, X_{t_1+1}, \cdots, X_{t_2}\rangle$. Let Z_t be the measurement at time t, and $Z_{t_1:t_2}$ denote the sequential measurements from time t_1 to t_2. Let w_t^i be the weight of the ith particle at time t.

A particle in CPF is defined as $X^i_{(t-k+1):t} = \langle X^i_{t-k+1}, X^i_{t-k+2}, \cdots, X^i_t\rangle$, which corresponds to a potential trajectory of the user. t is the current time, k is the sliding window width of the trajectory, and superscript i denotes that the particle is the ith potential trajectory. CPF maintains many particles and each particle has an associated weight signaling the probability of this trajectory. Different from traditional particle filtering where

Algorithm 1: Tracking with Context-aware Particle Filtering

```
1: Particle initialization: Sample $X_0^i$ according to $Z_0$.
2: for $t = 1 : t_{now}$ do
3:     Sampling: Sample $X_t^i \sim q(X_t^i | X_{(t-k):(t-1)}^i, Z_t)$.
4:     Weight updating:
       $w_t^i \propto w_{t-1}^i \frac{p(X_t^i | X_{(t-k):(t-1)}^i, Z_{t-1}) p(Z_t | X_t^i)}{q(X_t^i | X_{(t-k):(t-1)}^i, Z_t)}$
5:     Normalization: Normalize $w_t^i$ such that $\sum_{i=1}^n w_t^i = 1$
6:     Location estimation: $\hat{X}_t$ is defined as weighted
       average of $X_t^i$ from the top $40\%$ highest weighted
       particles.
7:     Resampling: Resample particles according to $w_t^i$,
       then set each weight $w_t^i = 1/n$ after resampling.
8: end for
```

only a single location X_t^i is selected as a particle, our setting preserves trajectory contexts.

Now we are ready to present the skeleton for user tracking using CPF in Algorithm 1. Line 1 of Algorithm 1 is the particle initialization step. Then after each user step (line 1), the new potential user locations and their associated probabilities are updated (lines 1–1). In the update, the new potential user locations are predicted by leveraging the current locations and the step length and heading estimated by the step model (line 3). Afterwards, the measurement obtained at the new location is utilized to revise the probability of each predicted location (line 1). Lines 1 and 1 provide a user location estimation by integrating all the particles and their weights. Line 7 is the resampling step to eliminate the wrongly moved particles. The details of the algorithm are described next.

A. Particle Initialization (Line 1 in Algorithm 1)

Sampling from measurement probability: Given the measurement Z_0 at time 0, initial particles should be sampled according to $p(X_0^i | Z_0)$, which is the probability distribution of all possible locations X_0^i given measurement Z_0. However, $p(X_0^i | Z_0)$ is unknown and thus can not be sampled. Instead, CPF performs sampling according to the measurement probability $p(Z_0 | X_0^i)$, which is the probability of obtaining measurement Z_0 at location X_0^i. This is because $p(X_0^i | Z_0) \propto p(Z_0 | X_0^i)$, which is proved in the following theorem.

Theorem 1: At time 0, probability $p(X_0^i | Z_0) \propto p(Z_0 | X_0^i)$.

Proof: Since at time 0, probability $p(X_0^i)$ can be regarded as having the same value for all locations, we have

$$p(X_0^i | Z_0) = \frac{p(Z_0 | X_0^i) p(X_0^i)}{p(Z_0)} \propto p(Z_0 | X_0^i). \tag{3}$$

■

Calculation of $p(Z_0 | X_0^i)$: CPF performs sampling according to the measurement probability $p(Z_0 | X_0^i)$, and a location X_0^i with higher measurement probability has a higher probability to be sampled. Recall that in the offline phase, we have divided the user interested areas of a building into grid cells and each cell is regarded as a unique location. The measurements of each cell are also collected. To calculate the measurement probability $p(Z_0 | X_0^i)$, CPF first finds the grid cell β to which location X_0^i belongs, and then calculates the *measurement similarities* between Z_0 and each of the measurements of β. The average measurement similarity is utilized as $p(Z_0 | X_0^i)$.

B. Particle Propagation (Line 3 in Algorithm 1)

After each user step, CPF updates the particles and their associated weights to depict the probability distribution of the new potential locations of the user. The update of particles is termed particle propagation.

Particle propagation is performed by displacing each old particle $X_{(t-k):(t-1)}^i = \langle X_{t-k}^i, X_{t-k+1}^i, \cdots, X_{t-1}^i \rangle$ with a new particle $X_{(t-k+1):t}^i = \langle X_{t-k+1}^i, X_{t-k+2}^i, \cdots, X_t^i \rangle$. This moves the sliding window of each potential trajectory (i.e., each particle) forward from time $t-1$ to t. Towards this, for each particle, its current location X_t^i needs to be predicted. CPF performs X_t^i prediction as follows.

1) If the trajectory $X_{(t-k):(t-1)}^i$ is continuous, then the new location X_t^i can be directly predicted as $\hat{X}_t^i = X_{t-1}^i + \boldsymbol{u}$. Here $\boldsymbol{u}$ is the location displacement vector where $\vec{\boldsymbol{u}}$ is the step heading and $||\boldsymbol{u}||$ is the step length. Note that the step heading and length can be obtained by simply invoking the step model introduced in Section III-B.

2) If the trajectory is not continuous, we can not use $X_{t-1}^i + \boldsymbol{u}$ to directly predict $\hat{X}_t^i$ because X_{t-1}^i may be a wrong prediction. Instead, X_t^i is predicted as $\hat{X}_t^i = \mu + \boldsymbol{u}$, where μ is the *reference location* of the trajectory $X_{(t-k):(t-1)}^i$ given the current measurement Z_t. The concept of *reference location* is defined as follows.

Definition 2: reference location: Given a measurement Z, the reference location of a trajectory $X_{t_1:t_2}$ is defined as the location which is part of trajectory $X_{t_1:t_2}$ and has measurements with the highest average measurement similarity to Z.

After predicting the new user location $\hat{X}_t^i$ for a particle, the new user location X_t^i of this particle is actually generated by sampling from a 2D Gaussian distribution centered at the predicted location $\hat{X}_t^i$, i.e., $X_t^i \sim N(\hat{X}_t^i, V)$. Sampling is used to deal with sensor errors. The Gaussian distribution is centered at the predicted location $\hat{X}_t^i$, and the covariance matrix V is defined as a diagonal matrix $diag(\sigma_1^2, \sigma_2^2)$, where σ_1 controls the variance of locations along the main direction of the path and σ_2 controls the variance of locations along the direction perpendicular to the main direction. In our settings, σ_1 is set to 60 cm, which is approximately one-user-step length. σ_2 is set to 30 cm since a user is very likely to walk along the main direction of a path rather than walking towards the sides [1].

If the new sampled location X_t^i is invalid in the floor plan, e.g., X_t^i hits a wall, its corresponding new particle $X_{(t-k+1):t}^i$ is abandoned.

C. Particle Weight Update (Line 4 in Algorithm 1)

After particle propagation, the weight w_t^i of each new particle should be updated. The weight updating works recursively where w_t^i is updated by using w_{t-1}^i. Formally

$$w_t^i \propto w_{t-1}^i \frac{p(X_t^i|X_{(t-k):(t-1)}^i, Z_{t-1})p(Z_t|X_t^i)}{q(X_t^i|X_{(t-k):(t-1)}^i, Z_t)}. \quad (4)$$

The derivation of (4) is proven below.

Proof: Let $p(X_{0:t}|Z_{0:t})$ denote the true posterior, and let $q(X_{0:t}|Z_{0:t})$ be the proposal distribution from which particles $X_{0:t}^i$ are sampled, i.e., $X_{0:t}^i \sim q(X_{0:t}|Z_{0:t})$. To compensate for the discrepancy between the true posterior and the proposal distribution [20], the weight w_t^i of particle $X_{0:t}^i$ is set to [18]

$$w_t^i \propto \frac{p(X_{0:t}^i|Z_{0:t})}{q(X_{0:t}^i|Z_{0:t})}. \quad (5)$$

1) For the numerator of (5), $p(X_{0:t}^i|Z_{0:t})$ can be calculated as follows:

$$\begin{aligned} &p(X_{0:t}^i|Z_{0:t}) \\ &= \frac{p(X_{0:t}^i|Z_{0:t-1})p(Z_t|X_{0:t}^i, Z_{0:t-1})}{p(Z_t|Z_{0:t-1})} \\ &= \frac{p(X_{0:t-1}^i|Z_{0:t-1})p(X_t^i|X_{0:t-1}^i, Z_{0:t-1})p(Z_t|X_{0:t}^i, Z_{0:t-1})}{p(Z_t|Z_{0:t-1})}. \end{aligned}$$

2) For the denominator of (5), $q(X_{0:t}^i|Z_{0:t})$ can be chosen to be factorized as follows [18]:

$$q(X_{0:t}^i|Z_{0:t}) = q(X_{0:t-1}^i|Z_{0:t-1})q(X_t^i|X_{0:t-1}^i, Z_{0:t}).$$

Therefore, we have the following weight updating equation:

$$w_t^i \propto w_{t-1}^i \frac{p(X_t^i|X_{0:t-1}^i, Z_{0:t-1})p(Z_t|X_{0:t}^i, Z_{0:t-1})}{p(Z_t|Z_{0:t-1})q(X_t^i|X_{0:t-1}^i, Z_{0:t})} \quad (6)$$

where

$$w_{t-1}^i \propto \frac{p(X_{0:t-1}^i|Z_{0:t-1})}{q(X_{0:t-1}^i|Z_{0:t-1})}.$$

All the weights will be normalized in each round of weight updating. Thus, (6) can be simplified to

$$w_t^i \propto w_{t-1}^i \frac{p(X_t^i|X_{0:t-1}^i, Z_{0:t-1})p(Z_t|X_{0:t}^i, Z_{0:t-1})}{q(X_t^i|X_{0:t-1}^i, Z_{0:t})}. \quad (7)$$

In our settings, only the most recent trajectory and the most recent measurement are taken into consideration rather than all the contexts from the very beginning. Let k be the sliding window width of the trajectory, then (7) can be further simplified as

$$w_t^i \propto w_{t-1}^i \frac{p(X_t^i|X_{(t-k):(t-1)}^i, Z_{t-1})p(Z_t|X_{(t-k+1):t}^i, Z_{t-1})}{q(X_t^i|X_{(t-k):(t-1)}^i, Z_t)}. \quad (8)$$

For computational simplicity, measurement Z_t can be regarded as depending on the current location only, so (8) can be finally simplified to

$$w_t^i \propto w_{t-1}^i \frac{p(X_t^i|X_{(t-k):(t-1)}^i, Z_{t-1})p(Z_t|X_t^i)}{q(X_t^i|X_{(t-k):(t-1)}^i, Z_t)}$$

which is the same as (4). ■

Next, we introduce how to calculate the three probabilities in (4). *First, the proposal distribution $q(X_t^i|X_{(t-k):(t-1)}^i, Z_t)$ calculation.* According to the particle propagation process introduced in the particle propagation step (step 2), the proposal distribution $q(X_t^i|X_{(t-k):(t-1)}^i, Z_t)$ in (4) is

$$\frac{1}{2\pi|V|^{1/2}} \exp\left(-\frac{1}{2}(X_t^i - \hat{X}_t^i)^T V^{-1}(X_t^i - \hat{X}_t^i)\right) \quad (9)$$

where $\hat{X}_t^i$ is the predicted current user location (the location prediction is introduced in the particle propagation step), and the covariance matrix $V = diag(\sigma_1^2, \sigma_2^2)$, where σ_1 controls the variance of locations along the main direction of the path and σ_2 controls the variance of locations along the direction perpendicular to the main direction.

Second, the transition probability $p(X_t^i|X_{(t-k):(t-1)}^i, Z_{t-1})$ calculation. Since the real transition probability is unknown, we can estimate it with a 2D Gaussian distribution around the reference location in the past trajectory $X_{(t-k):(t-1)}^i$. Formally, the transition probability $p(X_t^i|X_{(t-k):(t-1)}^i, Z_{t-1})$ in (4) can be estimated as

$$\frac{1}{2\pi|Q|^{1/2}} \exp\left(-\frac{1}{2}(X_t^i - \mu)^T Q^{-1}(X_t^i - \mu)\right) \quad (10)$$

where μ is the reference location of trajectory $X_{(t-k):(t-1)}^i$ given measurement Z_{t-1}, and $Q = diag(\sigma_3^2, \sigma_4^2)$, where σ_3 controls the variance of locations along the main direction of the path and σ_4 controls the variance of locations along the direction perpendicular to the main direction. In our settings, σ_3 is set to 60 cm, which is around the length of one user step, and σ_4 is set to 30 cm.

Third, the measurement probability $p(Z_t|X_t^i)$ calculation: In (4), $p(Z_t|X_t^i)$ is estimated as follows. CPF first finds the grid cell β to which location X_t^i belongs (recall that we have already divided the whole indoor space into grid cells and collected measurements for each cell in the offline phase), and then it calculates the measurement similarities between Z_t and the measurements of β. The average similarity is utilized as $p(Z_t|X_t^i)$.

D. Location Estimation (Lines 5 and 6 in Algorithm 1)

After particle weights have been updated, the weights of all particles are normalized such that $\sum_{i=1}^n w_t^i = 1$. There are two alternative methods to estimate the real location [1]. The first is to use the location X_t^i in the highest weighted particle $X_{(t-k):t}^i$, and the second is to use the weighted average of X_t^i from the top 40% highest weighted particles. Through experiments, we found that the second method yields more steady and accurate locations.

E. Particle Resampling (Line 7 in Algorithm 1)

After the four aforementioned steps, resampling is performed. In resampling, the weight of a particle is regarded as the probability that it will be sampled. Particles with higher weight will be sampled more often than others [19]. In this way, resampling is able to eliminate the wrongly moved particles, which have weights approximately equal to zero [19]. The weight of each resampled particle is set to $1/n$.

F. Discussion

1) *Computational Complexity Analysis:* The most time consuming steps of the VMag system are the fingerprint collection and the neural network training. However, both of these two steps can be finished in the offline phase, which means they do not consume any time during the online positioning phase. More importantly, we would like to point out that the time-consuming fingerprint collection and model building processes are inevitable in most existing infrastructure-free indoor localization approaches.

For the online tracking phase, the overall computational complexity is linear in n, where n is the number of particles. The detailed computational complexity analysis is as follows. Algorithm 1 illustrates the steps in online positioning. The sampling (line 3 of Algorithm 1) can be done in a few operations using the Box-Muller method. The weight updating (line 4) is the most time consuming step due to the deep features extraction and the three probabilities calculation. However, since the feature extraction network is already trained in the offline phase, this step exhibits constant-time complexity for each particle. The steps of normalization (line 5), location estimation (6) and resampling (line 7) can also be done in $O(n)$ operations. As such, the overall computational complexity is $O(n)$.

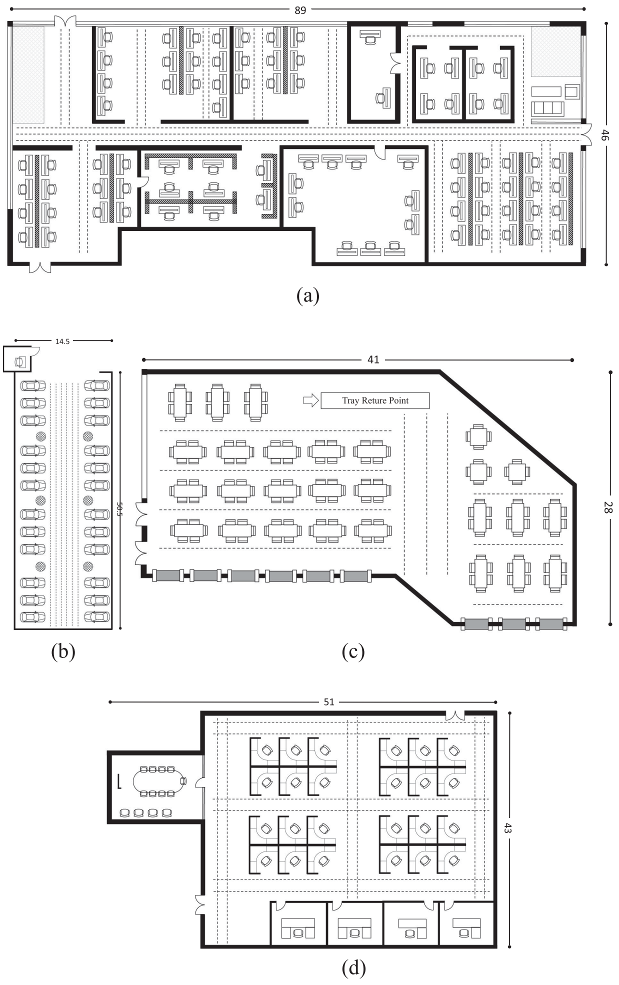
2) *Self-Adaptive Indoor Localization System:* The proposed VMag system is an infrastructure-free and fingerprint based approach. Analogous to existing infrastructure-free indoor localization approaches, when the whole indoor environment is drastically reshaped, it becomes necessary to recollect all the measurement fingerprints and retrain the model in a new offline phase. We believe it will be very interesting and challenging to design a self-adaptive retraining neural network [21]–[23] for indoor localization, which could automatically retain the network when its performance is not satisfactory. To the best of our knowledge, such a design would be the first to formulate and address this problem for infrastructure-free indoor localization. Two of the key challenges would be to automatically collect the images and magnetic field signals and to automatically detect drastic environment changes without any infrastructure assistance. Thus, we believe that this task is very novel and requires further in-depth exploration. Hence, we leave it as future work.

V. Experiments

In this section, we evaluate the VMag system with a variety of representative indoor environments to understand its effectiveness and limitations.

A. Experimental Setup

The experiments are conducted in four different indoor environments, i.e., a laboratory, a garage, a canteen and an office. The testing areas of the four indoor environments are 4094 m^2, 732 m^2, 1148 m^2, and 2193 m^2, respectively. The floor plans are illustrated in Fig. 10.



(a)

(b)

(c)

(d)

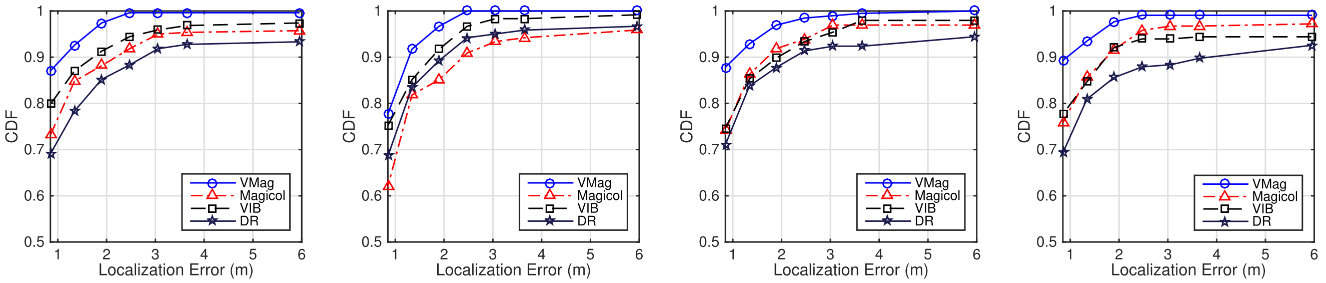
Fig. 10. Floor plans of four different types of indoor environments: (a) laboratory, (b) garage, (c) canteen, and (d) office.

1) *Hardware:* We installed our VMag client software on several iPhone 6 smartphones, so that they were able to collect magnetic field signals and images for localization purposes. The collected data is then sent to the server over a network to obtain the localization results. We utilized a MacBook Pro as the server, with an Intel Core i5 CPU at 2.6 GHz and 8 GB of RAM. The number of particles in the particle filtering was set to 2000 in all the experiments.

2) *Fingerprint Data Collection:* As mentioned before, we divided the path areas of a building into 60 cm × 60 cm grid cells. Each cell was regarded as a unique location and assigned a location label. For each location, we randomly collected 10 measurements inside its corresponding grid cell. Each measurement consists of an image and a magnetic field signal. The survey paths, along which the fingerprints for different locations were collected, are demonstrated using dashed lines in Fig. 10.

B. Comparison With Representative Existing Methods

We first compare VMag with the representative existing methods in the four different indoor environments. We selected a Dead Reckoning (DR) method [6], a visual image based (VIB)



(a) (b) (c) (d)

Fig. 11. Performance comparison with representative existing methods on the four different indoor environments: (a) laboratory, (b) garage, (c) canteen, and (d) office.

method [24] and a geomagnetic field based method (Magicol) [1] as comparison approaches. These methods were chosen because they represent different types and are closely related to our work. The DR and VIB methods are two infrastructure-free approaches, while the Magicol method is a hybrid approach that can work with or without infrastructure. The DR method [6] utilizes inertial sensors to estimate the position displacement and thus track users. VIB [24] is a representative method for most existing visual-based indoor localization methods that utilize traditional handcrafted image features for indoor localization. The Magicol method [1] utilizes the magnetic field for indoor localization, and it fuses WiFi and the magnetic field to obtain better performance.

Test data collection: In order to test all the methods, we walked many trajectories in the whole area of each building and made random turns. The duration of the walks ranged from 6 seconds to 3 minutes and the starting points of the walks were randomly selected. The number of tested locations for the environments are 540 for the laboratory, 121 for the garage, 197 for the canteen, and 215 for the office, respectively.

Fig. 11 illustrates the comparison results of the four indoor environments, where the x-axis represents the localization error and the y-axis is the cumulative distribution function (CDF). The cumulative distribution function (CDF) of the localization errors is reported starting from 0.85 m. This is because we divided the whole space of each environment into 60 cm × 60 cm grid cells and regarded each cell as a location. Thus the upper bound distance between two points inside the same cell is $0.6 \times \sqrt{2} \approx 0.85$ m. In the garage environment, WiFi is not available, so we only utilized the magnetic field for the Magicol method.

From Fig. 11, we can see that VMag significantly and consistently outperforms the other tested methods in all the four environments. From Fig. 11 we can also observe that VMag can achieve a 91% probability of 1.34 m accuracy for all the four indoor environments. In contrast, Magicol can achieve a 91% probability of 2.47 m for all the four indoor environments, VIB of 2.16 m and DR of 3.06 m.

If we regard that a location is *precisely determined* when the localization error is within 1 m, then the probability of precise determination (i.e., probability of a localization error within 1 m) for each tested method is listed in Table I. We can see that VMag consistently outperforms the other tested methods in all the four environments with respect to the precise determination probability.

TABLE I
PROBABILITY OF A LOCALIZATION ERROR $\leqslant$1 M. COMPARISON WITH EXISTING METHODS

	VMag	Magicol	VIB	DR
Laboratory	87%	73%	79%	69%
Garage	78%	62%	75%	68%
Canteen	88%	74%	75%	71%
Office	89%	76%	78%	69%

In summary, experimental results demonstrate that VMag significantly and consistently outperforms the other tested methods in all the four indoor environments, and can achieve a meter-level accuracy with a 91% probability.

C. Comparison With Localization Using Traditional Features and Using Neural Networks With Different Layers

In Fig. 9, we have designed a 5-layer neural network to extract deep features for each measurement. In this subsection, we conducted experiments to compare our deep features with traditional features, and then compare the 5-layer network with 1-layer and 3-layer networks. For traditional magnetic features, we utilize the three dimensional $\langle m_x, m_y, m_z \rangle$ magnetic signal. For handcrafted image features in indoor localization, we refer to the work of [24] and [25], and select color histogram and texture features for the images. For the 1-layer network, the size of the hidden layer is 4196. For the 3-layer network, the sizes of the hidden layers are 6144, 5120, and 4196, respectively. All other settings (such as the utilized loss function and parameter initialization) of the 1-layer and 3-layer networks are the same as those of the 5-layer network for fairness.

We demonstrate the results in Fig. 12. The experimental results show that our extracted deep features lead to a significant improvements in performance over traditional features for all the four indoor environments. This reconfirms the effectiveness of our deep feature extraction process. The results also show that the 5-layer network of deep feature extraction is slightly better than those of the 1-layer and 3-layer networks.

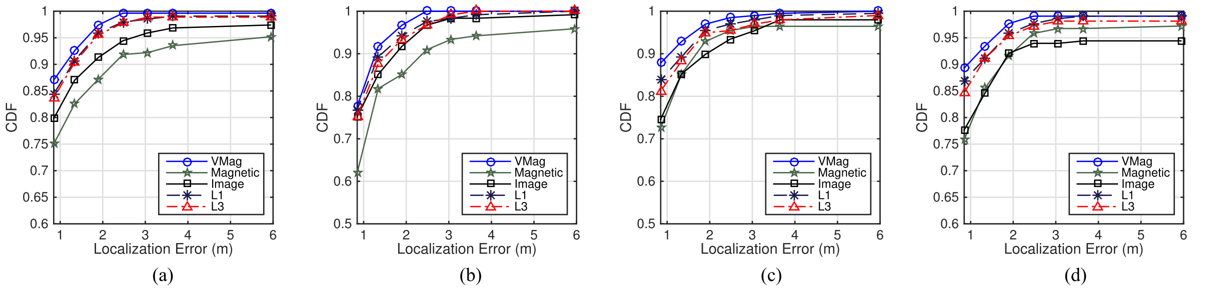


Fig. 12. Performance comparison with traditional features, 1-layer, and 3-layer neural networks. “Magnetic” denotes localization using only the magnetic field feature, “Image” denotes localization using handcrafted image features, and “L1” and “L3” denote localization using a 1-layer or 3-layer deep feature extraction network, respectively. (a) Laboratory. (b) Garage. (c) Canteen. (d) Office.

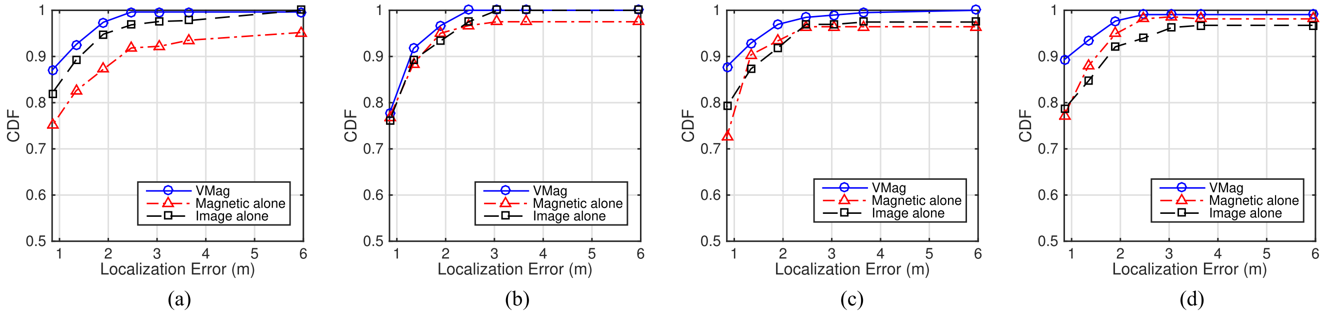


Fig. 13. Performance comparison with localization using a geomagnetic field or our extracted deep image features alone: (a) laboratory, (b) garage, (c) canteen, and (d) office.

TABLE II
PROBABILITY OF A LOCALIZATION ERROR $\leqslant$1 M. COMPARISON WITH USING MAGNETIC OR DEEP IMAGE FEATURES ALONE

	VMag	Magnetic alone	Image alone
Laboratory	87%	75%	82%
Garage	78%	73%	76%
Canteen	88%	73%	79%
Office	89%	77%	79%

D. Comparison With Localization Using a Geomagnetic Field or Deep Image Features Alone

Subsequently, we also studied the performance of localization using magnetic features or our extracted deep image features (for the deep image features extraction details see Step 1 in Section III-C) alone. To this aim, we implemented two other versions of the VMag system: one using only the magnetic field and the other using the extracted deep image features.

Fig. 13 illustrates the comparison results. We can see that combining the magnetic field and deep image features indeed leads to an improvement in performance compared to using either one of them alone.

We also list the probability of precise determination (i.e., the probability of a localization error within 1 m) for each tested method in Table II. We can see that VMag consistently outperforms the other tested methods in terms of the precise determination probability.

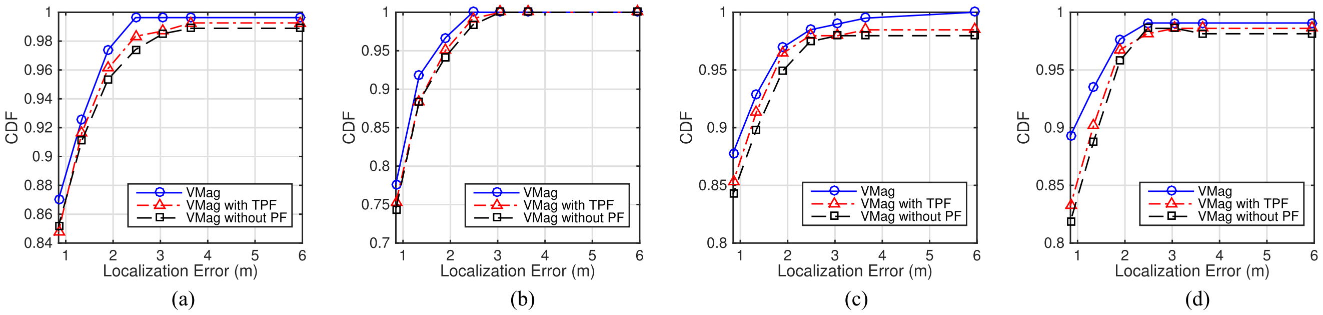
In summary, experimental results demonstrate that fusing magnetic and deep image features can achieve higher localization performance than using either one of them alone. This further verifies that the magnetic field and visual images have complementary location resolution capabilities.

E. Comparison With Localization Using Traditional Particle Filter

After comparing VMag with other approaches and evaluating the features, we compared our method with localization using a traditional particle filter (TPF). The result of localization without any particle filter is also reported to serve as a baseline. VMag with a traditional particle filter is implemented by replacing the context-aware particle filter in VMag with a traditional one. VMag without particle filter is implemented by removing the particle filter altogether.

Fig. 14 illustrates the comparison results and we can observe the following. 1) Both, localizations using the traditional and the context-aware particle filters, outperform localization without particle filter. 2) The localization using the context-aware particle filter consistently outperforms the one using the traditional particle filter for all the four environments. The detailed probability of a localization error $\leqslant$ 1 m is illustrated in Table III.

Moreover, we also studied the convergence process of the two approaches along long trajectories. We selected the trajectories with a duration of around 1.5 minutes from all the traces collected and calculated their localization errors over time. Fig. 15

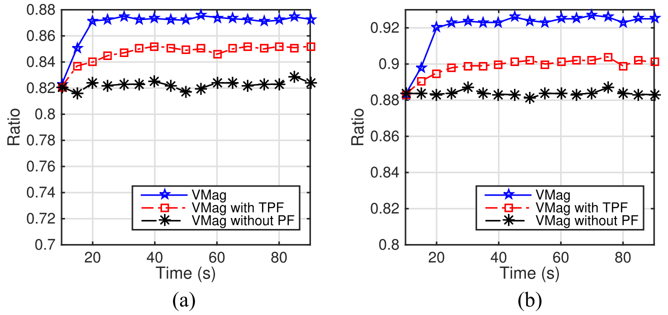


(a) (b) (c) (d)

Fig. 14. Performance comparison with localization without particle filter (PF) and localization using traditional particle filter (TPF): (a) laboratory, (b) garage, (c) canteen, and (d) office.

TABLE III
PROBABILITY OF A LOCALIZATION ERROR ⩽ 1 M. COMPARISON WITH LOCALIZATION USING TRADITIONAL PARTICLE FILTER

	VMag	VMag with traditional particle filter
Laboratory	87%	85%
Garage	78%	75%
Canteen	88%	85%
Office	89%	83%



(a) (b)

Fig. 15. Comparison with traditional particle filter on long traces. (a) shows the ratio of a localization error within 1.34 m, while (b) shows the ratio of a localization error within 0.85 m.

shows the comparison results, with the x-axis indicating time in seconds, and the y-axis representing the ratio of correctly localized locations within a certain localization error threshold. Fig. 15(a) shows the ratio of correctly localized locations within 0.85 m, and Fig. 15(b) shows the ratio of correctly localized locations within 1.34 m. Note that 0.85 m $= \sqrt{2} \times 0.6$ m is the upper bound distance between two locations in the same grid cell, and 1.34 m $= \sqrt{5} \times 0.6$ m is the upper bound distance between two locations that are in two adjacent grid cells. From Fig. 15, we can see that the context-aware particle filter converges more quickly than the traditional particle filter and yields a higher accuracy after convergence. We can also see that without particle filter, the localization accuracy will not increase over time due to the ignorance of the available contexts.

In summary, the experimental results show that our proposed CPF outperforms a traditional particle filter in terms of accuracy and convergence speed. Note that all the other settings of the two compared approaches were the same except for the particle filters utilized.

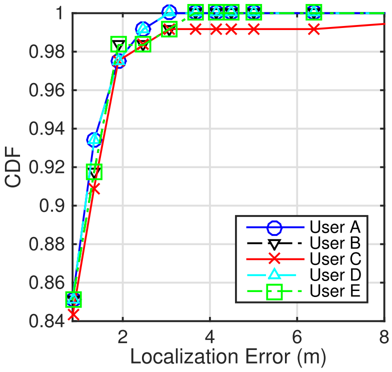


Fig. 16. Robustness of VMag with different users.

F. Robustness With Different Users

To evaluate the robustness of VMag with respect to different users we asked five students to walk around in the four indoor environments. The five students were of different heights in the range from 1.56 m to 1.85 m. Two students were female and the others male. Naturally their gait patterns are different. We recorded their localization errors, which are presented in Fig. 16. We can see that it is difficult to distinguish the localization accuracy curves of different users, which verifies the robustness of VMag. The CDF curves of the five users are also consistent with those of the earlier experiments, which were conducted by three of the authors.

G. Discussion of Grid Cell Size

In the fingerprint data collection we divided the user interesting areas into 60 cm × 60 cm grid cells. We also conducted experiments to evaluate the effect of the grid cell size on the localization accuracy of VMag. Through experiments, we found that 1) a smaller grid cell size usually yields a higher localization accuracy, and 2) VMag with relatively big grid cell sizes is still competitive in localization accuracy, i.e., it achieves a more than 85% localization accuracy of 1.70 m in all the four indoor environments for a relatively big grid cell size of 1.20 m. Due to space limitations, we omit the details of the experiments. The main reasons that a smaller grid cell size yields a better accuracy may be twofold. (a) Let ℓ be the grid cell width. Though many measurements are located at the correct grid cell, since the upper bound distance between two locations inside the same cell is $\sqrt{2}\ell$, the localization error can only be reported as no larger than $\sqrt{2}\ell$. (b) The smaller the grid cell size, the more similar the measurements, and thus the better the localization accuracy.

VI. Related Work

Over the years, an increasing number of approaches have been developed to address the critical problem of indoor localization. Most existing methods however assume the existence of certain types of infrastructures, such as Infrared (IR) beacons (for IR based methods), Bluetooth beacons (for Bluetooth based methods), ultrasound transmitters/receivers (for ultrasound based methods), and wireless access points (for WiFi based methods). It is beyond the scope of this paper to comprehensively report on all existing research efforts in this field. Instead, we focus on a relative succinct account of the most related work, which can be roughly grouped into three categories, namely infrastructure-based, infrastructure-free, and hybrid approaches, as follows.

1) *Infrastructure-based approaches* assume the presence of certain infrastructure. Among them, WiFi based approaches (e.g., [11], [26]) have been the most well-studied, and they often consist of two phases, the offline preparation and online positioning phases. In the offline phase, the WiFi signal strength of each location is measured to build a WiFi fingerprint map for all the users' locations of interest. Then during the online phase, a newly measured WiFi signal is incorporated to search for the best matching location in the map. The localization error is typically around three meters due to unstable WiFi signals. IR and Bluetooth based approaches have also been proposed with meter-level accuracy reported, which however heavily rely on the existence of certain external infrastructure. For example, since IR signals only travel within line of sight (i.e., within a room), usually hundreds to thousands of IR beacons have to be deployed to cover the whole space of a single building.

2) *Infrastructure-free approaches* can work without additional infrastructure, thus allowing for pervasive indoor localization. The most closely related work includes magnetic field based, visual image based and dead reckoning based approaches.

Magnetic field based approaches utilize the locally anomalous but stable geomagnetic field for indoor localization. In [3], [27], the authors studied the feasibility of using the magnetic field alone for indoor localization. In [4] and [10], the magnetic anomalies serve as unique magnetic fingerprints for locations. Magnetic field based approaches can significantly enhance infrastructure-free approaches as geomagnetism is natural and ubiquitous. Meanwhile the localization accuracy is usually on the order of 3.5 meters, since the feature dimensionality of the magnetic field is low, and thus the uniqueness of a fingerprint can not be guaranteed.

In visual image based approaches, a database of fingerprint images and their associated locations is constructed in advance. At runtime, a comparison is made between a newly captured image and the fingerprint images to identify the best match [28], [29]. These approaches usually only utilize traditional handcrafted image features (e.g., [25], [30], [31]) such as SIFT, color histograms, and texture related features, and the localization error is generally around two meters. Due to the availability of large datasets like ImageNet and the rise of neural networks, deep learning approaches have achieved great success for image classification tasks. [13] proposed a convolutional neural network to learn deep features from images for scene recognition tasks. Doulamis *et al.* proposed online retrainable neural networks to automatically test the performance of a network and then automatically retrain it [21], [22], [31], [33]. However, very few indoor localization systems have leveraged the advantages of deep learning methods.

Dead reckoning approaches estimate the position displacement based on readings of the inertial sensors, which is subsequently used to track the user. The dead reckoning approaches [6] are simple, but they suffer from an inherit error-accumulation problem.

3) *Hybrid approaches* refer to those combining different methods for improved performance. In [1], a fusion approach of magnetic field and WiFi sensors has been proposed, which integrates magnetic field and WiFi to deal with the low discernibility of the magnetic field and achieves an accuracy of around three meters. In [34], a hybrid approach based on WiFi and Bluetooth is proposed. It utilizes Bluetooth hotspots to divide the large space into small partitions, and then engages WiFi fingerprints to infer the location of the user. The hybrid approaches of [19], [35]–[38] further confirm that fusion of multiple raw signals can lead to improved performance, including for example the fusion of motion and WiFi sensors in [19] (with meter-level accuracy reported), and the combination of radio and camera sensors in [35] (meter-level accuracy reported). However, these hybrid approaches still require the deployment of certain infrastructure to achieve the reported accuracies.

VII. Conclusion

In this paper we have proposed a novel indoor localization and tracking approach termed VMag for smartphone users without any infrastructure assistance. VMag fuses both magnetic and visual sensing for indoor localization and achieves a more than 91 percent localization accuracy of 1.34 m. We have conducted an in-depth experimental study of the properties of the geomagnetic field and visual images for the purpose of indoor localization. After that, the complementary location recognition capabilities of the magnetic field and visual images have been studied. Based on the results, we designed a context-aware particle filtering framework to track users. Extensive experiments were conducted in four different indoor environments including a laboratory, a garage, a canteen and an office. The experimental results show that VMag achieves a 91 percent localization accuracy of 0.85 m, 1.34 m, 1.34 m and 0.85 m, respectively, in the four different indoor environments.

References

[1] Y. Shu *et al.*, "Magicol: Indoor localization using pervasive magnetic field and opportunistic WiFi sensing," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1443–1457, Jul. 2015.

[2] I. Bisio, F. Lavagetto, M. Marchese, and A. Sciarrone, "GPS/HPS-and Wi-Fi fingerprint-based location recognition for check-in applications over smartphones in cloud-based LBSs," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 858–869, Jun. 2013.

[3] B. Li, T. Gallagher, A. G. Dempster, and C. Rizos, "How feasible is the use of magnetic field alone for indoor positioning," in *Proc. Int. Conf. Indoor Position. Indoor Navigat.*, 2012, pp. 1–9.

[4] J. Haverinen and A. Kemppainen, "Global indoor self-localization based on the ambient magnetic field," *Robot. Auton. Syst.*, vol. 57, no. 10, pp. 1028–1035, 2009.

[5] M. Liu, "A study of mobile sensing using smartphones," *Int. J. Distrib. Sens. Netw.*, vol. 9, 2013, Art. no. 272916.

[6] F. Li *et al.*, "A reliable and accurate indoor localization method using phone inertial sensors," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 421–430.

[7] W. Wang, Y. Yan, L. Zhang, R. Hong, and N. Sebe, "Collaborative sparse coding for multi-view action recognition," *IEEE Multimedia Mag.*, vol. 23, no. 4, pp. 80–87, Oct.–Dec. 2016.

[8] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, vol. 2, pp. 1150–1157.

[9] L. Zhang, X. Li, L. Nie, Y. Yan, and R. Zimmermann, "Semantic photo retargeting under noisy image labels," *TOMCCAP*, vol. 12, no. 3, p. 37, 2016.

[10] D. Carrillo, V. Moreno, B. beda, and A. F. Skarmeta, "MagicFinger: 3D magnetic fingerprints for indoor location," *Sensors*, vol. 15, no. 7, pp. 17168–17194, 2015.

[11] H. Li *et al.*, "An indoor continuous positioning algorithm on the move by fusing sensors and Wi-Fi on smartphones," *Sensors*, vol. 15, no. 12, pp. 31 244–31 267, 2015.

[12] K. P. Subbu, B. Gozick, and R. Dantu, "Locateme: Magnetic-fields-based indoor localization using smartphones," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, 2013, Art. no. 73.

[13] B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[14] H. M. Ali and A. H. Omran, "Floor identification using smartphone barometer sensor for indoor positioning," *Int. J. Eng. Sci. Res. Technol.*, vol. 4, no. 2, pp. 384–391, 2015.

[15] F. Alsehly, T. Arslan, and Z. Sevak, "Indoor positioning with floor determination in multi story buildings," in *Proc. Int. Conf. Indoor Position. Indoor Navigat.*, 2011, pp. 1–7.

[16] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678 .

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[18] M. S. Arulampalam, S. Maskell, N. J. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[19] X. He, "Probabilistic multi-sensor fusion based indoor positioning system on a mobile device," *Sensors*, vol. 15, pp. 31464–31481, 2015.

[20] E. Orhan, *Particle Filtering*. New York, NY, USA: Univ. Rochester, 2012. [Online]. Available: http://www.cns.nyu.edu/ eorhan/notes/particle-filtering.pdf.

[21] N. D. Doulamis and A. D. Doulamis, "Semi-supervised deep learning for object tracking and classification," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 848–852.

[22] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "On-line retrainable neural networks: Improving the performance of neural networks in image analysis problems," *IEEE Trans. Neural Netw.*, vol. 11, no. 1, pp. 137–155, Jan. 2000.

[23] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, "Improving the performance of MPEG compatible encoding at low bit rates using adaptive neural networks," *Real-Time Imag.*, vol. 6, no. 5, pp. 327–345, Jan. 2000.

[24] N. Ravi, P. Shankar, A. Frankel, A. M. Elgammal, and L. Iftode, "Indoor localization using camera phones," in *Proc. 7th IEEE Workshop Mobile Comput. Syst. Appl.: Supplement*, Aug. 2006, pp. 1–7.

[25] J. Kim and H. Jun, "Vision-based location positioning using augmented reality for indoor navigation," *IEEE Trans. Consum. Electron.*, vol. 54, no. 3, pp. 954–962, Aug. 2008.

[26] Z. Yang, C. Wu, and Y. Liu, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[27] C. Zhang, K. Subbu, J. Luo, and J. Wu, "GROPING: Geomagnetism and crowdsensing powered indoor navigation," *IEEE Trans. Mobile Comput.*, vol. 14, no. 2, pp. 387–400, Feb. 2015.

[28] V. V. Nguyen and J. W. Lee, "A hybrid positioning system for indoor navigation on mobile phones using panoramic images," *TIIS*, vol. 6, no. 3, pp. 835–854, 2012.

[29] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2938–2946.

[30] L. Zhang *et al.*, "Detecting densely distributed graph patterns for fine-grained image categorization," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 553–565, Feb. 2016.

[31] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, "Adaptive trained neural networks for traffic prediction of VBR MPEG-2 video sources," in *Proc. Eur. Signal Process. Conf.*, 2000, pp. 1–4.

[32] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, and Q. Tian, "Discovering discriminative graphlets for aerial image categories recognition," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5071–5084, Aug. 2013.

[33] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "An adaptable neural-network model for recursive nonlinear traffic prediction and modeling of MPEG video sources," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 150–166, Jan. 2003.

[34] A. Baniukevic, C. S. Jensen, and H. Lu, "Hybrid indoor positioning with Wi-Fi and bluetooth: Architecture and performance," in *Proc. IEEE 14th Int. Conf. Mobile Data Manag.*, Jun. 2013, vol. 1, pp. 207–216.

[35] S. Papaioannou, H. Wen, A. Markham, and N. Trigoni, "Fusion of radio and camera sensor data for accurate indoor positioning," in *Proc. IEEE Int. Conf. Mobile Ad Hoc. Sens. Syst.*, Oct. 2014, pp. 109–117.

[36] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 186–200, Feb. 2015.

[37] T. Guan, Y. He, J. Gao, J. Yang, and J. Yu, "On-device mobile visual location recognition by integrating vision and inertial sensors," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1688–1699, Nov. 2013.

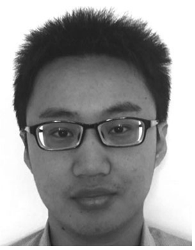
[38] V. P. Minotto, C. R. Jung, and B. Lee, "Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1032–1044, Jun. 2014.

Zhenguang Liu received the the B.S. degree from Shandong University, Jinan, China, and the Ph.D. degree from Zhejiang University, Hangzhou, China, both in computer science.

He is a Postdoctoral Research Fellow with the Department of Computer Science, National University of Singapore, Singapore. His research interests include indoor localization, data mining, and intelligent systems.

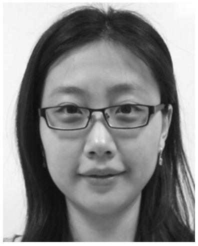
Luming Zhang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China.

He is currently with the Department of Computer Science and Information Engineering, Hefei Unviersity of Technology, Hefei, China. He is also with the National University of Singapore Suzhou Research Institute, Suzhou, China. His research interests include computer systems, visual perception analysis, image enhancement, and pattern recognition.



Qi Liu received the B.S. degree from Shandong University, Jinan, China, the M.S. degree from the National University of Singapore, Singapore, both in computer science, and is currently working toward the Ph.D. degree in computer science at the School of Computing, National University of Singapore.

His research interests include data mining, deep learning, and multimedia mining.



Yifang Yin received the B.E. degree in computer science and technology from Northeastern University, Shenyang, China, in 2011, and the Ph.D. degree from the School of Computing, National University of Singapore, Singapore, in 2016.

She was a Research Intern with the Incubation Center, Research and Technology Group, Fuji Xerox Co., Ltd., Tokyo, Japan, from 2014 to 2015. Her research interests include geotagged video annotation and retrieval, geometadata correction, and video summarization.

Li Cheng received the Ph.D. degree in computer science from the University of Alberta, Edmonton, AB, Canada.

Prior to joining the Bioinformatics Institute (BII), Agency for Science Technology and Research, Singapore, in 2010, he was with the Statistical Machine Learning Group, NICTA, Australia; TTI-Chicago, USA; and the University of Alberta, Edmonton, AB, Canada. He is currently a Research Scientist with the BII. His research interests include machine learning and computer vision.

Roger Zimmermann received the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, CA, USA, in 1994 and 1998, respectively.

He is an Associate Professor with the Department of Computer Science, National University of Singapore (NUS), Singapore, where he is also a Deputy Director with the Interactive and Digital Media Institute and a Co-Director of the Centre of Social Media Innovations for Communities. He has coauthored a book, six patents, and more than 150 conference publications, journal articles, and book chapters. His research interests include streaming media architectures, distributed and peer-to-peer systems, mobile and georeferenced video management, collaborative environments, spatiotemporal information management, and mobile location-based services.