2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 29 Nov. – 2 Dec. 2021, Lloret de Mar, Spain

MM-Loc: Cross-sensor Indoor Smartphone Location Tracking using Multimodal Deep Neural Networks

Xijia Wei
School of Informatics
University of Edinburgh
weixijia@outlook.com

Zhiqiang Wei
School of Informatics
University of Edinburgh
weizhiqiang@yahoo.com

Valentin Radu
Department of Computer Science
University of Sheffield
valentin.radu@sheffield.ac.uk

***Abstract*—Indoor positioning systems have been explored for decades to facilitate universal location-based services. However, complex environment conditions and sensing imperfections continue to be limiting factors to their large scale adoption. Rather than customising more ingenious solutions to handle corner cases in complex environments, we believe that a more efficient solution is to learn entirely from data with minimal engineering effort. We develop neural network based solutions for two positioning approaches, modelling Dead Reckoning with recurrent neural networks and WiFi Fingerprinting with deep neural networks. We propose a multimodal deep neural network architecture (MM-Loc) that bridges the features extracted by the modality-specific complements (sensor based and WiFi based) to join the two perspectives. We observe that this multimodal approach is better than single-modality models, and elegantly trains directly from raw data with minimal intervention.**

***Index Terms*—Indoor Localization, Multimodal Sensing, Sensor Fusion, WiFi Fingerprinting, Location Tracking, Dead Reckoning, Recurrent Neural Networks**

I. Introduction

A growing number of mobile applications offering intelligent and adaptive services require contextual information, with location being an important part of context [1]. However, to enable these services in complex environments, heavily-engineered systems [2, 3, 4] have been designed to be tuned to each deployment environment. Although effective in their evaluation conditions, it is uncertain if these can adapt to new observations or changing environments, being rigid with their design – requiring engineering effort to calibrate.

Complex conditions inside buildings such as radio frequency interference, dynamic environment conditions, and noisy sensor data make it hard to model WiFi fingerprints through simple modelling techniques (Bayesian models). Similarly, imperfect inertial sensors (accelerometer, gyroscope and magnetometer), drift and noise make it perform estimations on sensor data with simple techniques. These limitations have encouraged different ingenious engineering solutions to address their challenges, such as particle filters [3, 5], Kalman filters, graph-conditions [4] and constraint modelling [6]. However, these engineered systems make strong assumptions from isolated data observations by proposing rigid mathematical formulation. These formulations model the conditions available in those evaluation scenarios, but not necessarily proposing generalisable observations. More prevalent data is exposing the limits of these models with rigid assumptions, which fail when run-time data is imperfect.

978-1-6654-0402-0/21/$31.00 © 2021 IEEE

In the age of big data, we propose a data-driven approach for modelling indoor localization. Our approach adopts deep learning techniques, which have proven efficient in other domains (natural language processing, computer vision, machine translation), where large datasets are already available for training deep neural networks. We transform traditional methods used for indoor localization by modelling their underlying process with deep neural networks in end-to-end machine learning solutions. This replaces the hand-tuned models that just approximate limited data observations. Deep learning models move the focus onto automatic feature extraction that capture the complexity of large datasets. Using deep learning, we model two main localization methods: Pedestrian Dead Reckoning and WiFi Fingerprinting, separately and in conjunction. Dead Reckoning produces a sequence of estimations starting from a known location, followed by sequential position estimations using the direction of movement and displacement distance. We adopt a recurrent neural network solution to model the sequential chain of estimations based on our previous work [7]. We model WiFi Fingerprinting with a deep neural network. It reads the received signal strength (RSS) sampled from all visible access points (AP), the WiFi fingerprints, to produce latitude-longitude coordinate estimations as regression outputs.

Although we find each of these models to produce good estimations independently, our hybrid proposal built on a multimodal deep learning architecture, improves the overall localization accuracy. This bridges information from each modality-specific architecture to compensate the unbalanced sensing data issues. This is facilitated by integrating neural network structures from each of the sensing modalities with joint neural network structures. Model training process is efficient, due to uniform optimisation across the two branches of the network. Since WiFi scans are available on mobile phones at a lower rate than inertial sensors, our model relies continuously on inertial sensors, and opportunistically on WiFi data, when a fresh sample is available in the system.

We evaluate our proposed multimodal deep neural network architecture to find that it improves the performance by about 50% over the independent sensing modalities.

The contributions of this work are as follows:

- We model Dead Reckoning with recurrent neural networks operating on smartphone inertial sensor data. Estimations of displacement and direction of movement are directly learned from data within the structure of a recurrent neural network.
- To model WiFi Fingerprinting, we develop a regression approach which offers several benefits, shallow models (ideal for mobile devices with limited energy), and does not require exhaustive sampling from each space of the target building. This achieves a median estimation accuracy of about 2.6 metres.
- We introduce an efficient sensor fusion solution relying on a multimodal deep neural network architecture. This extracts modality-specific features independently from inertial sensors and WiFi fingerprints data, followed by cross-modality features to boost the performance of the two independent solutions with the ultimate median prediction accuracy within 2 metres.

II. Motivation and Related Work

Position estimation of smartphones inside buildings is not easy due to the GPS being unreliable in environments shielded by walls and ceilings. At the same time, other radio signals with longer penetration (cellular and FM) are limited to the granularity of position estimation they can offer [8]. Alternative methods have been proposed to take advantage of a broader range of sensors available on smartphones [9]. However, none have managed to produce a robust and scalable system for efficient indoor position estimation. We believe the reasons are: i) indoor spaces are too complex to model with limited and fragmented observations from the environment (limited data), and ii) current systems rely on human interpreted features extracted from data (e.g., engineered solutions to estimate the number of steps and direction of movement). This complexity makes it extremely hard to model their propagation from scarce observations and with simplistic modelling techniques.

The best approach, in our opinion, is to rely on models with high generalisation to learn directly from data, taking advantage of growing data volumes. Deep neural networks have proven successful in other fields with increasing access to data (computer vision, text processing, speech, etc.). We believe that deep neural networks can tackle the aforementioned long-standing problems that limiting indoor localization.

A. Dead Reckoning on Inertial Sensors

Dead Reckoning builds on inertial sensors to calculate current location by estimating displacement distance and direction of movement. However, the sensor drift problem limits its applications, making it hard to double integrate acceleration to estimate displacement [10]. The same issue is experienced when estimating the direction of movement. To avoid stiff and rigid engineered solutions, other works use machine learning to identify characteristics in inertial sensors, such as for step size estimation based on neural networks [11].

B. WiFi Fingerprinting on Received Signal Strength

The WiFi Fingerprinting localization approach consists of two phases: i) training phase or commonly called offline phase that collects samples prior in a training set, and ii) the runtime phase or so-called online phase that produces estimations based on incoming observations [3]. However, indoor spaces experience a challenging radio propagation environment with multi-path effect, shadowing, signal fading and other forms of signal degradation and distortion. Any slight change in the environment affects the estimation, so a model should be able to assimilate information from new data easily and capture more of the unexpected variations. Others use deep neural networks in WiFi fingerprints signal strength based indoor localization [12], and also on WiFi signals with a formulation of the propagation model for EZ [13, 14], while more recent work has been using neural networks on Channel State Information (CSI) [15].

C. Multimodal Approaches

Multimodal approaches make estimations from multiple perspectives of cross-modality data. Filtering methods like particle filters and Kalman filters have been proposed to address multimodalities of data. Specifically, HiMLoc uses particle filters to integrate inertial sensors with WiFi fingerprints based on prior observations of Gaussian processes for direction estimation, distance estimation and correlation between samples and location in buildings, and admissible human activity [3]. Similarly, WiFi-SLAM and Zee build on particle filters emphasising their importance for random system initialisation [5], while Kalman filters are used to integrate inertial sensing modalities [6]. Other engineered approaches, such as UnLoc, combining sensing modalities based on empirical observations of how some locations are unique across one or more sensors [2], MapCraft uses conditional random fields [16], LiFS uses graph constraints to map and position estimations on the trajectory [4]. Similarly, WILL builds a connected graph to estimate location at room level [17].

However, when samples are formed in multimodalities, solutions building on machine learning show their advantages of understanding correspondences between communicative multimodalities and capturing in-depth features from natural representations instead of focusing on a single modality without alternative feature inputs.

Neural networks across sensing modalities have not been used for indoor localization before, although it has been used for other context recognition tasks, like for human activity recognition [18]. We aim to customise an end-to-end multimodal deep neural network for the indoor localization task to produce location estimation based on inertial sensors and WiFi fingerprints data. Training directly on data has its drawbacks, that of moving the challenges to the quality of training dataset and cross-sensor modality alignment, although this can be eventually automated by other systems such as vision-based systems [19].

III. Methods

This section illustrates the algorithms we deploy on the single-modality models for inertial sensors and WiFi fingerprints data, following by the description of our proposed MM-Loc architecture: The cross-sensor multimodal neural network.

A. Dead Reckoning with Recurrent Neural Network

Dead Reckoning is the process of estimating continuous location by starting from a known location and estimating consecutive locations based on a stream of observations coming into the system (direction of movement and displacement). This resembles the process performed by recurrent neural networks, building on previous estimations (or features within previous estimations) and new observations to produce a sequence of predictions. Our previous work has shown the fundamental concepts of one class of the recurrent neural networks called the Long-Short Term Memory (LSTM) network that performs similarly as Dead Reckoning to estimate positions from streaming inertial sensor data [7].

Figure 1 illustrates the unrolled chain of the LSTM network, where C_t is the long-term memory at time t and h_t is the block output at time t, or short-term memory, both transmitted to the following LSTM block in the chain. The sample size

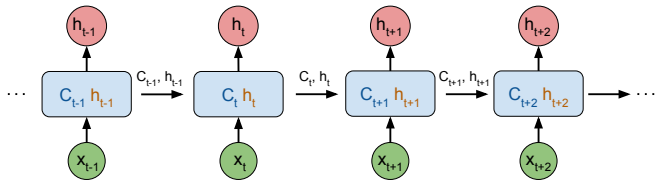


Fig. 1. Unrolled Chain of LSTM Neural Network

of the time-sequential sensor data is $(Timestep * Features)$. The feature of each data point is the magnitude value of acceleration, gyroscope and magnetic field data. The number of data points in each sample is determined by the chosen time window (here, we set the one-second time window sampled with 10Hz frequency). Each sample is offered a target position in coordinate $(X_i,\ Y_i)$. The regression output of the LSTM model is the estimated position in coordinates $(X_{est},\ Y_{est})$ based on inertial sensor data.

B. WiFi Fingerprinting with Deep Neural Networks

For periodic recalibration, the WiFi is a reliable anchoring mechanism, used extensively in previous research [2, 3, 5, 17] due to this relying on instantaneous observations to match on a database or with a pre-trained model for position estimation. The configuration of modelling WiFi Fingerprinting with Deep Neural Network (DNN) is shown in figure 2. The end-to-end DNN takes WiFi scans from sensed access points at each sampling timestep as input features and $(X_i,\ Y_i)$ coordinate of where the fingerprint is collected as target variables during the training phase while producing two numerical outputs of $(X_{est},\ Y_{est})$ for each coordinate as geographical location estimations.

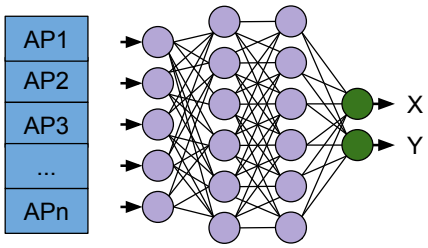


Fig. 2. WiFi Fingerprinting Deep Neural Network

C. Sensor Fusion by Multimodal Deep Neural Networks

By joining both inertial sensors and WiFi fingerprints modalities, their unique perspectives can contribute to more robust estimations. Similar to our previous work in multimodal deep learning for context recognition [18], here we explore the capacity of such a construction to combine the two aforementioned neural networks operating on each sensing modality.

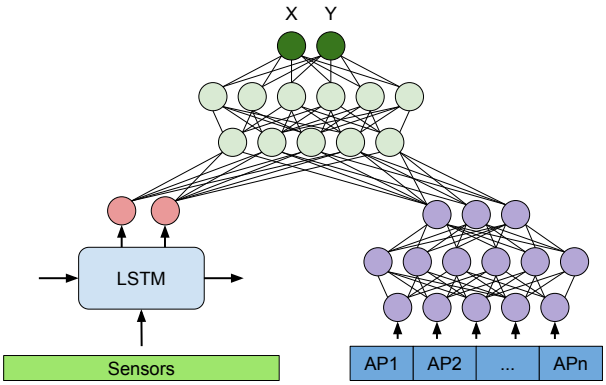


Fig. 3. MM-Loc: our proposed multimodal deep neural network architecture for indoor localization with two parallel single-modality extractors at the bottom and a joint deep neural network as a sensor-fusion regression network on the top.

Figure 3 presents our proposed MM-Loc architecture, an end-to-end multimodal deep neural network for indoor localization, which brings in the perspectives of inertial sensors and WiFi fingerprints modalities.

The MM-Loc neural network first reads time-sequential inertial sensor data by LSTM sub-network and WiFi RSS data by DNN sub-network synchronously from the beginning, where the sample size of the sensor modality at the LSTM side is $(Timestep * Sensor\_num)$ and for the DNN side is $(Timestep * AP\_num)$. Both modalities are reshaped to 128-dimensional hidden outputs from each branch. These two parallel 128 units hidden outputs are then integrated by concatenation to 256 units. The 256-dimensional joint vectors are fed into three fully-connected layers (FC) with the input size of 256, 128 and 64, then transferred to the top prediction layer. This operates as a regression, producing continuous values in two-dimensional outputs of $(X_{est},\ Y_{est})$ from the joint cross-modalities hidden layers.

The contribution of LSTM to the cross-modality component is in the form of its multiple LSTM layers worked as feature extractor, which passes sensor hidden feature towards the higher layer. The WiFi component contributes as WiFi feature extractor. We replace the two units regression outputs, originally from single modality based neural networks, with additional fully-connected layers to transfer extracted hyper-dimensional vectors to joint layers. Both neural network components pass the individual sensor modality hidden features to the cross-modality layers with element-wise calculation and then passed to the final layer for location prediction regression. The MM-Loc containing three sub-components of LSTM, DNN and cross-modality neural networks are trained synchronously with computing the loss in the forward pass and splitting the gradients on each branch on the backward pass.

What is unique about this construction is that it can handle missing inputs from the WiFi modality. This is because WiFi scans are produced at a much lower rate than inertial sensors, so when there is not WiFi scan available, the input is a vector with all components value of 0 (by normalising -100 dBm), so the WiFi branch of the network contributes negligibly to the cross-modality component, in which case the majority of the contribution comes from the inertial sensor modality branch. When both modalities have inputs, the same multimodal architecture automatically adjusts the weights of cross modalities to produces estimations based on both modality contributions.

IV. Data

In this experiment, we build the cross-sensor datasets of two scenarios by collecting multimodalities of data with ground truth locations from scratch and eventually processing them to a uniform machine learning dataset with strategies of interpolation, normalisation, overlapping, down-sampling and data alignment.

A. Data Collection

Cross-sensor data are collected by an Android application designed specially for data gathering task. It uses Android APIs, which provides sensor samples on event base, updating on value change, continuously logs inertial sensor data of accelerometer, magnetometer, gyroscope and WiFi received signal strength with sampling timestep in background. Meanwhile, the interactive map interface (aligned to Google Map API) allows user to click on the screen to mark down the ground truth location as latitude and longitude coordinates.

In order to collect the machine learning dataset of sensor features with labels, we activate the application logging on two smartphones synchronously, with one in the pocket to resemble the perspective of sensors in natural motions and another one used to record ground truth locations when passing special locations such as corners, elevators, etc. Specifically, during one round of data collection, we build two datasets synchronously: Dataset.1 that walking along the corridor to collect samples from inertial sensors and WiFi scan; Dataset.2 that collecting ground truth geographical location labels synchronously by clicking the screen to log latitude and longitude information when passing key locations such as corners.

To evaluate the generalisation of the model, we decide to collect data from two representative office buildings. Both scenarios contain variational factors which impact indoor complexity. Factors include people walking through frequently that affects how we walk along the corridors during data gathering; Various indoor working electronic equipment such as elevators, computers, printers and portable devices, which generates electromagnetic radiation; Grouped WiFi access points mixed with personal hotspot; Building materials contain reinforced concrete, metal and glass, which impact on signal propagation. Furthermore, both datasets are collected from multiple mobile devices with different hardware sensitivity and sampling rate. During the data gathering process, different walking speeds and gestures are considered to add the variety of the samplings. Table I illustrates our own-collected cross-sensor datasets for both scenarios.

TABLE I
Dataset Description

Datasets	Inertial Sensor	RSS	APs	Rounds	Time
Scenario A	24450	25541	102	14	407 Mins
Scenario B	29836	8390	750	14	497 Mins

B. Data Pre-processing

1) Inertial Sensors Dataset: Similar to our previous work in inertial sensor based LSTM model [7], to make the time-series sensor dataset fits the model, we take the same data processing strategy to generate the machine learning dataset. Specifically, it contains three parameters settings of the time window, overlapping rate and down-sampling rate. A proper time window setting is considered to balance between location estimation refreshing frequency on time window and on-device inference computational cost. An appropriate overlapping rate setting could emphasise the information between inter-samples since the information from previous time windows are reinforced on overlapping for better inferences. To improve forward-pass speed, the down-sampling operation shrinks the time-series input by discrete samples while maintaining the sampling features within the same time window interval of each second.

We first assure position invariant condition by working with the magnitude value on the three orthogonal axes from raw samplings in 3-dimension, where sensor$\{x,\ y,\ z\}$ are the values measured on each of the three Cartesian axes.

$$sensor_{magnitude} = \sqrt{sensor_x^2 + sensor_y^2 + sensor_z^2}$$

After calculating the magnitude value of accelerator, gyroscope and magnetometer, we normalise the input frequency by interpolating at a rate of 1 kHz. These are grouped in a time window of one second and associated with one position (deploying interpolation in between marked ground truth location labels to generate continuous values of latitude and longitude) to each time window. The sensor data is then overlapped with 90% and down-sampled from (1000 * 3) to (10 * 3) per sample with one second time interval.

2) WiFi Fingerprints Dataset: Inputs to neural networks are provided as vectors of the RSS values for each AP mounted inside the building. To construct this vector, we first scan the whole WiFi logfile to identify all unique APs observed throughout the data collection process (total of n APs observed inside the building), as well as the minimal and maximum signal strength encountered throughout, which are used to normalise the vector input to [0,1] interval by linear scaling. By observation, the min-max interval is [-100,-40] in dBm. Hence, for missing APs in WiFi scans, a value of -100 is associated with their representation in the n-dimensional vectors as input to the neural network. To keep the original features of the sampled data without unnecessary human intervention, we keep those occasionally seen personal hotspots, considered as noise, in the dataset to add complexity, which simulates the real environments of changeable WiFi signal distribution.

3) Cross-sensor Dataset: As two synchronously-logged datasets contain not only inertial sensor and WiFi RSS samples but also ground truth location information within the same time duration, the time record is utilised for matching multimodalities with geographical labels. Table II illustrates the components of the cross-sensor dataset after alignment.

TABLE II
CROSS-SENSOR DATASET

Time	Acc	Gyro	Mag	AP_0	AP_1	...	AP_n	X	Y
t_0	a_0	g_0	m_0	-100	-85	...	-100	X_0	Y_0
t_1	a_1	g_1	m_1	-100	-100	...	-100	X_1	Y_1
t_2	a_2	g_2	m_2	-70	-100	...	-65	X_2	Y_2
...	...	...	...	...	...	...	...	...	...
t_n	a_n	g_n	m_n	-100	-100	...	-100	X_n	Y_n

It is noticed that the WiFi sampling frequency is significantly slower than the inertial sensing rate due the hardware limitations. For some timesteps, if there only contains inertial sensor samplings, same as the strategy applied on the missing WiFi, we use -100 dBm to represent missing WiFi scans for all APs. They are added in parallel for a uniform size of multimodal dataset. Meanwhile, location labels are normalised to the extreme boundaries chosen for the building and scaled to [0,1] interval. Estimations of neural network models are converted back into latitude and longitude coordinates in meters to measure the estimation errors as the euclidean distance between predict and target locations.

V. EXPERIMENTS

In this section, we implement single-modality location estimators and our multimodal sensor fusion estimator (MM-Loc). We test these models on senor and WiFi datasets collected from two buildings (deployment scenarios). To explore the opportunity to reduce energy consumption further, we vary the WiFi scan frequency in our proposed MM-Loc system.

We perform the following split of the datasets: 65%, 25% and 10% for training, validation and testing respectively. Results are here presented using Cumulative Distribution Function (CDF) charts.

A. Single Modality Implementations

1) Sensor based Recurrent Neural Network: Similar to our previous work on the end-to-end independent LSTM model, we use a recurrent neural network to extract observations from inertial sensor data [7]. However, instead of estimating the location directly from the model, the RNN integrated into the multimodal architecture behaves as a robust feature extractor for sensor fusion. Here, to evaluate the independent sensor model, the head of the network accepts a fully-connected layer to estimate the locations. The parameter settings of the LSTM model are shown in table III.

TABLE III
SENSOR NEURAL NETWORK PARAMETER SETTINGS

Parameter	Settings
Epoch	100
Batch Size	100
LSTM Hidden Units	128
LSTM Layer	1 Layer
Learning Rate	0.005
Learning Rules	RMSprop

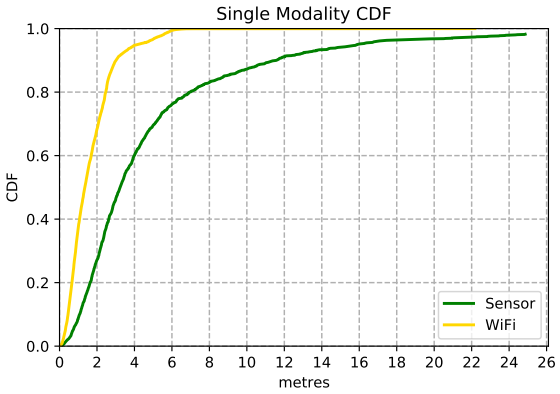
2) WiFi based Deep Neural Network: WiFi scans are received through the Android API at an irregular frequency, which average update rate of about one second. Here we evaluate the performance of the WiFi model based on the processed WiFi fingerprint dataset containing -100dBm representing missing values. For the multimodal architecture, we construct data tuples for every 100ms. Those inputs with missing WiFi samples are masked with an input vector of zero values after normalisation. The WiFi based estimator is modelled with a three-layer neural network regression model. The model details are shown in table IV. The only variation of the model structure is the input sizes caused by the number of APs seen in a building. For our two evaluation scenarios, there are 102 APs and 750 APs respectively.

TABLE IV
WIFI NEURAL NETWORK PARAMETER SETTINGS

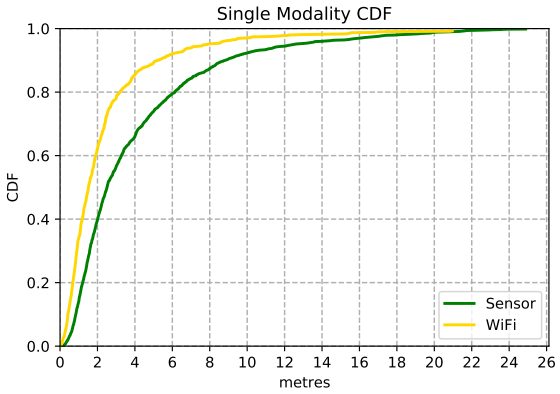
Parameter	Settings
Epoch	100
Batch Size	100
DNN Hidden Units	128
DNN Layer	3 Layers
Dropout Rate	0.5
Learning Rate	0.001
Learning Rules	RMSprop

B. Single Modality Model Evaluation

Figure 4(a) shows the performances of the sensor model and WiFi model on scenario A. It is noticed that the WiFi model perform significantly better than the sensor model with 2.6× better accuracy. Within the estimation accuracy of 80%, the WiFi model has a precision of 2.6 metres error while the sensor model holds 6.9 metres prediction error. Training the model on the dataset collected in scenario B, we observe similar performances as presented in figure 4(b). The WiFi model has an error of 3 metres and the sensor based model

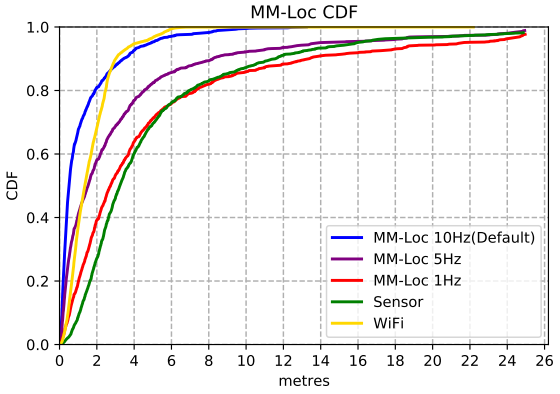


(a) Single Modality Model Performance CDF on Scenario A

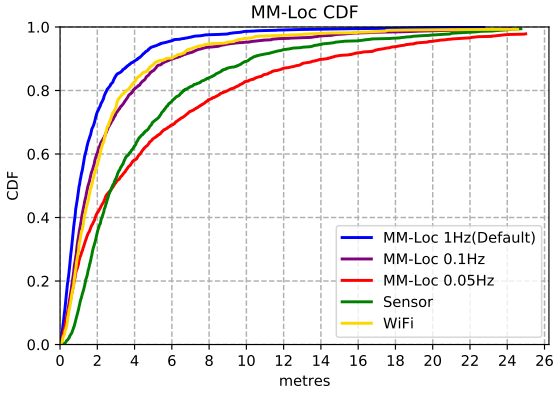


(b) Single Modality Model Performance CDF on Scenario B

Fig. 4. Single-modality model performances over two scenarios.



(a) MM-Loc Performance CDF on Scenario A



(b) MM-Loc Performance CDF on Scenario B

Fig. 5. The comparison between MM-Loc models running at different WiFi sampling rates and two independent single-modality models.

TABLE V
MULTIMODAL DEEP NEURAL NETWORK ARCHITECTURE

Layers			**Output Shape**
LSTM Layer (sensor)			(Batch Size,128)
FC Layer.1 (WiFi)			(Batch Size,128)
Dropout Layer.1 (WiFi)			(Batch Size,128)
FC Layer.2 (WiFi)			(Batch Size,128)
Dropout Layer.2 (WiFi)			(Batch Size,128)
FC Layer.3 (WiFi)			(Batch Size,128)
Fusion Layer (joint)			(Batch Size,256)
FC Layer.4 (joint)			(Batch Size,128)
FC Layer.5 (joint)			(Batch Size,64)
FC Layer.6 (joint)			(Batch Size,2)
Batch Size	**Learning Rate**	**Learning Rules**	**Dropout Rate**
100	0.001	RMSprop	0.5

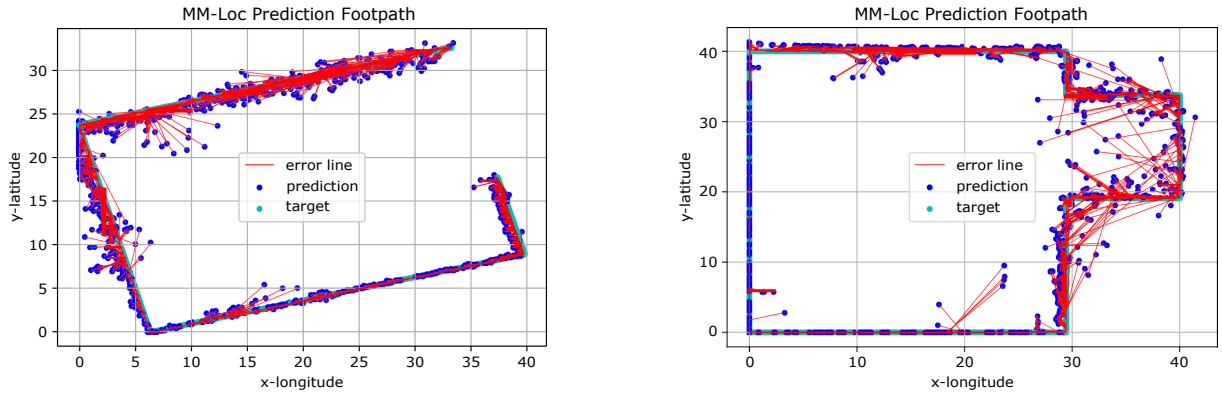
has an error of 6.2 metres. The similar performances in the two scenarios give us confidence in the generalisation power of the two models.

C. MM-Loc Implementation

Table V indicates the architecture and parameter settings of our proposed multimodal model (MM-Loc). Here, we evaluate the performances of all the models on data collected from the two scenarios, but using different WiFi sampling frequency. Specifically, for scenario A, as the default WiFi sampling rate is about 10Hz as sourced from the system, we reduce the scanning frequency of the dataset from 10Hz to 5Hz and 1Hz with a filter. The purpose of adjusting WiFi frequency is to assess the impact on location estimation accuracy of this energy-saving strategy of scanning at lower frequencies. It also shows how this would behave on systems where a high refresh rate is not available. For Scenario B, we decrease the WiFi sampling frequency from the original 1Hz to 0.1Hz and even 0.05Hz for the same reasons.

D. MM-Loc Evaluation

Figure 5 presents the comparison between the performances of MM-Loc running at different WiFi sampling frequencies, and the single-modality baseline models. Generally, MM-Loc with the highest sampling rate performs best. MM-Loc median accuracy is within 2 metres error for 80% of the prediction cases, which is 3.5× better than the sensor baseline model. By



(a) MM-Loc predicted footpath on Scenario A

(b) MM-Loc predicted footpath on Scenario B

Fig. 6. MM-Loc Footpath Visualisation

comparing figure 5(a) and 5(b), we observe that in scenario A, MM-Loc reaches a better accuracy of 2.6 meters median precision at the intersection point with WiFi model. However, the WiFi model has consistently good performance even for the extreme cases, with the maximum error being better than that of any other models. In scenario B, MM-Loc performs better than any other models. Another observation is that with decreasing sampling rates, the multimodal model prediction accuracy experiences the same trend. MM-Loc with intermediate refreshing rate data still predicts with approximately 4 meters precision. Hence, a proper sampling rate setting contributes to minimum on-device computing cost and power consumption with reliable positioning accuracy.

E. MM-Loc Visualisation

Figure 6 visualises the predicted footpath of MM-Loc in both scenarios. The red error lines indicate the distance between the coordinates of the ground truth and the predicted coordinates of MM-Loc. In scenario A, the accuracy at first quartile (Q_1), second quartile (Q_2) and third quartile (Q_3) is 0.332, 0.697 and 1.562 meters respectively; While in scenario B, Q_1, Q_2 and Q_3 are 0.521, 1.028 and 2.145 metres respectively. This is also visible from the CDFs.

We observe that MM-Loc predicts the footpath along the corridor with high quality, having clear estimation boundaries. However, some predictions are over 5 metres away from the ground truth, especially at the corners of corridors. We put this on the difficulty of observations in the WiFi component near corners. The other aspect introducing errors is the magnetic interference present in some places on the pathway (elevators and heavy iron materials in building materials).

VI. Discussion

We showed that traditional smartphone indoor localization methods could be modelled through deep neural network architectures, both as individual components with specific modality neural architectures (RNN and DNN) and also as sensor-fusion with multimodal neural networks. Through this, we are moving the effort from engineering each component, step counting, direction and creating integration methods (particle filters, Kalman filter and graph-based constraints) to a purely data-driven approach, relying entirely on an end-to-end neural network solution.

This approach is not tuned to the conditions of a single building, but being trained on data from the target building. This data-driven approach generalises better than algorithms designed for specific buildings and their local conditions. Although the effort is moved entirely on the quality of training data, we believe that in the age of big data, access to such datasets will become much easier, potentially through automated systems of labelling position from vision-based systems [19], and adopting transfer learning.

Our multimodal architecture shows its ability to capture fine-grained observations distilled in features on each sensing modality branch and brings complementary perspectives together through the cross-modality network structure in the joint architecture.

VII. Conclusions

In this work, we present how the task of performing indoor localization can be modelled with a multimodal deep neural network. First, we assess candidates for modality-specific neural network architectures to model two popular localization techniques, Dead Reckoning through recurrent neural networks and WiFi Fingerprinting using deep neural networks. We observe that these can perform location estimation independently with a median error rate of 2.6 and 6.9 metres respectively. Their combination with a multimodal architecture captures the perspective from cross-sensor modalities to reduce median error to under 2 metres. Our multimodal deep neural network can efficiently combine diverse sensing modalities with automatic training purely from data to achieve high accuracy instead of relying on heavily engineered localization components.

REFERENCES

[1] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *International Conference on Data Mining*. IEEE, 2012.

[2] He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury. No need to war-drive: Unsupervised indoor localization. In *Proceedings of MobiSys*. ACM, 2012.

[3] Valentin Radu and Mahesh K. Marina. Himloc: Indoor smartphone localization via activity aware pedestrian dead reckoning with selective crowdsourced wifi fingerprinting. In *Proc. IPIN 2013*. IEEE, 2013.

[4] Zheng Yang, Chenshu Wu, and Yunhao Liu. Locating in fingerprint space: wireless indoor localization with little human intervention. In *Proceedings of the 18th annual international conference on Mobile computing and networking*, pages 269–280. ACM, 2012.

[5] Anshul Ari, Krishna Kant Chintalapudi, Venkata N. Padmanabhan, and Rijurekha Sen. Zee: Zero-effort crowdsourcing for indoor localization. In *Proc. MobiCom*. ACM, 2012.

[6] Zhuoling Xiao, Hongkai Wen, Andrew Markham, and Niki Trigoni. Robust pedestrian dead reckoning (r-pdr) for arbitrary mobile device placement. In *Proc. IPIN*. IEEE, 2014.

[7] Xijia Wei and Valentin Radu. Calibrating recurrent neural networks on smartphone inertial sensors for location tracking. In *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8. IEEE, 2019.

[8] Yin Chen, Dimitrios Lymberopoulos, Jie Liu, , and Bodhi Priyantha. Fm-based indoor localization. In *Proc. MobiSys 2012*. ACM, 2012.

[9] ASanorita Dey, Nirupam Roy, Wenyuan Xu, Romit Roy Choudhury, and Srihari Nelakuditi. Accelprint: Imperfections of accelerometers make smartphones trackable. In *NDSS*, 2014.

[10] Robert Harle. A survey of indoor inertial positioning systems for pedestrians. *Communications Surveys and Tutorials*, 15(3), 2013.

[11] Moustafa Alzantot and Moustafa Youssef. Uptime: Ubiquitous pedestrian tracking using mobile phones. In *Proc. WCNC*. IEEE, 2012.

[12] Huan Dai, Wen hao Ying, and Jiang Xu. Multi-layer neural network for received signal strength-based indoor localisation. *Communications*, 10(6), 2016.

[13] Shehadi Dayekh, Sofiene Affes, Nahi Kandil, and Chahe Nerguizian. Cooperative localization in mines using fingerprinting and neural networks. In *Proc. WCNC*. IEEE, 2010.

[14] Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N Padmanabhan. Indoor localization without the pain. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, pages 173–184. ACM, 2010.

[15] Xuyu Wang, Lingjun Gao, Shiwen Mao, and Santosh Pandey. Csi-based fingerprinting for indoor localization: A deep learning approach. *Transactions on Vehicular Technology*, 66(1), 2017.

[16] Zhuoling Xiao, Hongkai Wen, Andrew Markham, and Niki Trigoni. Lightweight map matching for indoor localisation using conditional random fields. In *Proc. IPSN*. IEEE, 2014.

[17] Chenshu Wu, Zheng Yang, Yunhao Liu, and Wei Xi. Will: Wireless indoor localization without site survey. *IEEE Transactions on Parallel and Distributed Systems*, 24(4):839–848, 2013.

[18] V Radu, C Tong, S Bhattacharya, ND Lane, C Mascolo, MK Marina, and F Kawsar. Multimodal deep learning for activity and context recognition. *IMWUT*, 1(4), 2018.

[19] Adrian Cosma, Ion Emilian Radoi, and Valentin Radu. Camloc: Pedestrian location detection from pose estimation on resource-constrained smart cameras. In *arXiv:1812.11209*. preprint, 2018.