Mobile Device Usage Recommendation based on User Context Inference Using Embedded Sensors

Cong Shi*, Xiaonan Guo†, Ting Yu‡, Yingying Chen* Yucheng Xie† and Jian Liu§
*WINLAB, Rutgers University, North Brunswick, NJ 08902, USA
†Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA
‡Qatar Computing Research Institute, Ar-Rayyan, Qatar
§The University of Tennessee, Knoxville, TN 37996, USA

***Abstract*—The proliferation of mobile devices along with their rich functionalities/applications have made people form addictive and potentially harmful usage behaviors. Though this problem has drawn considerable attention, existing solutions (e.g., text notification or setting usage limits) are insufficient and cannot provide timely recommendations or control of inappropriate usage of mobile devices. This paper proposes a generalized context inference framework, which supports timely usage recommendations using low-power sensors in mobile devices Comparing to existing schemes that rely on detection of single type user contexts (e.g., merely on location or activity), our framework derives a much larger-scale of user contexts that characterize the phone usages, especially those causing distraction or leading to dangerous situations. We propose to uniformly describe the general user context with *context fundamentals*, i.e., physical environments, social situations, and human motions, which are the underlying constituent units of diverse general user contexts. To mitigate the profiling efforts across different environments, devices, and individuals, we develop a deep learning-based architecture to learn transferable representations derived from sensor readings associated with the context fundamentals. Based on the derived context fundamentals, our framework quantifies how likely an inferred user context would lead to distractions/dangerous situations, and provides timely recommendations for mobile device access/usage. Extensive experiments during a period of 7 months demonstrate that the system can achieve 95% accuracy on user context inference while offering the transferability among different environments, devices, and users.**

***Index Terms*—Mobile Device Usage, Deep Learning**

I. Introduction

Mobile devices forever change our lives, for better and for worse. On the one hand, almost any service could be accessed through a touch on the screen, which provides great convenience to our daily lives. On the other hand, we are increasingly addicted to mobile devices, forming annoying behaviors (e.g., gluing to a phone when having dinner with families and friends) or even dangerous habits (e.g., texting while driving). Such pathological behaviors could cause distraction from daily works and studies, leading to social isolation or even life-threatening injuries. To mitigate the negative impacts of such behaviors and ensure the wide adoption of mobile devices, it becomes increasingly important to detect inappropriate usage scenarios of mobile devices. Existing solutions [1], [2] rely on long-term analysis, such as using screen time of multiple days/weeks to analyze user behaviors. These approaches are not sufficient in handling many of today's usage scenarios as they fail to detect annoying behaviors in real time (that cause social uncomfortness) and dangerous usages that require immediate actions. Detection of inappropriate usage scenarios of mobile devices is not a single-dimensional decision that only based on a user's habits, but more importantly, it is tightly coupled with the current and immediate context/environment that the user is in. For instance, texting is fine while sitting at home, but it is inappropriate when attending a meeting and even could become dangerous when operating a vehicle. Therefore, understanding current and immediate user contexts is a critical key factor for determining the inappropriate usage scenarios of mobile devices.

Existing studies use different sensors to infer user contexts, allowing mobile devices to adapt their settings according to the user's immediate situation. For instance, existing studies could derive the user's location [3], physical activities [4], and surrounding environments [5]. These solutions rely on detecting single type of user contexts (e.g., coordinate of a user, type of an activity, or indoor/outdoor environment). However, such information is insufficient to determine inappropriate usage scenarios of mobile devices, which are usually determined by a much larger-scale of user contexts. For instance, attending a business meeting can be characterized as "indoor", "multiple people participation", and "static motion". While attending a party could be described with "multiple people participation" along with "outdoor" and "dynamic motion". We refer to these constituent units as context fundamentals, which are shared in a wide-spectrum of general user contexts.

To uniformly characterize general user contexts, as illustrated in Figure 1, we propose a context inference framework to derive three types of context fundamentals including physical environment (i.e., indoor or outdoor), social situation (i.e., single or multiple users), and human motion (i.e., static or dynamic motion) by using low-power sensors (i.e., WiFi module, microphone, and accelerometer) that are readily available on mobile devices. Existing context sensing approaches either require building profiles for each environment, device, and individual, or need dedicated equipment (e.g., shoe-mounted initial sensors [6]) or surrounding infrastructures (e.g., cell tower, wireless access point [7]). To reduce profiling efforts, we propose to derive transferable representations from sensor readings associated with context fundamentals across different environments, devices, and individuals. By using a small set

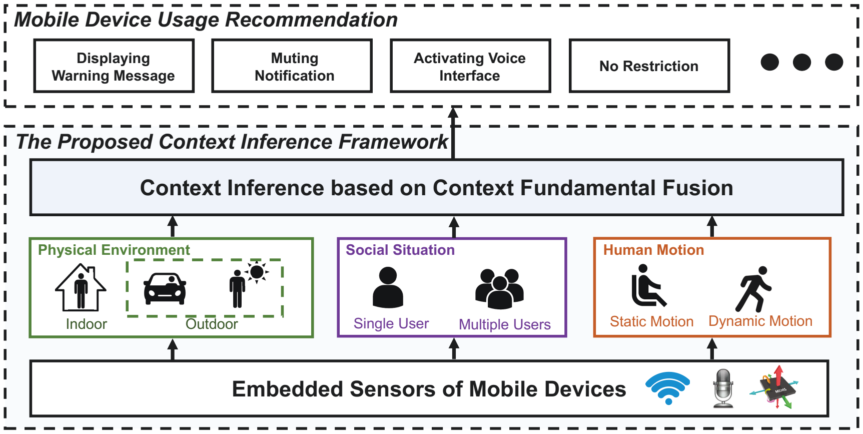

Mobile Device Usage Recommendation
Displaying Warning Message
Muting Notification
Activating Voice Interface
No Restriction
The Proposed Context Inference Framework
Context Inference based on Context Fundamental Fusion
Physical Environment
Indoor
Outdoor
Social Situation
Single User
Multiple Users
Human Motion
Static Motion
Dynamic Motion
Embedded Sensors of Mobile Devices

Fig. 1. An illustration of the proposed context inference framework to provide mobile device usage recommendations based on the derived context fundamentals.

of widely-available sensors in mobile devices, our context inference frameworks could be easily deployed to provide timely usage recommendations.

To design such a framework, several challenges need to be addressed: (i) It is challenging to extract effective features from sensor readings to consistently characterize the context fundamentals, which exist in large-scale general user contexts. (ii) Sensor data usually carry substantial information that is specific to environments, devices, and human subjects, the designed framework needs to be able to derive transferable representations that are resilient to the impacts of such variations; (iii) Constrained by the energy budget/limited computational capabilities of mobile devices, it is also challenging to provide timely context inference without affecting the functionalities of other mobile applications.

To effectively characterize the context fundamentals while restraining energy consumption on a mobile device, we propose to derive representative features from a small set of low-power sensors which are capable of measuring general contextual information such as the physical environment around the user, interactions between people, and human movements. Furthermore, to derive transferable representations without introducing much computation overhead, we design a lightweight deep neural network (DNN) architecture that could efficiently learn feature abstractions that are resilient to diverse variations. By deriving the transferable representations, the user can simply download models already built by mobile vendors (e.g., Apple, Samsung) to avoid the power-consuming training process on his own device. In particular, the DNN model exploits a bidirectional recurrent neural network (BRNN) to derive representations embedded with the temporal dynamics of the sensor features, which maximize the correlations between the derived representations of the same type of context fundamental. It then aggregates the representations of all three sensors to produce comprehensive and transferable representations. Additionally, as shown in Figure 2, our framework quantifies how likely a usage scenario would lead to distractions/dangerous situations based on the derived context fundamentals and then provides timely usage recommendations. For example, it restricts the usage of apps when the context fundamentals of inside vehicle, multi-user participation, and dynamic&active motion are detected. The main contributions of our work are summarized as follows:

- We propose context fundamentals to uniformly describe large-scale user contexts that characterize phone usage scenarios, especially those causing distractions or leading to dangerous situations.
- Our context inference framework leverages a small set of low-power embedded sensors to facilitate the context fundamental recognition. A lightweight DNN architecture is developed to learn reliable and transferable representations, which could maximize the transferability across diverse environments, devices, and users.
- Based on the derived context fundamentals, we quantify how likely the current usage scenario would lead to distractions or dangerous situations, enabling timely mobile device usage recommendations.
- We conduct extensive experiments during a period of 7 months. The results demonstrate that the proposed framework could recognize context fundamentals with high accuracy while offering transferability.

II. Related Work

Traditional approaches to restraining inappropriate uses of the mobile device mainly rely on setting usage limits [2] and installing smartphone apps to curb addictive behaviors [8]. For example, Screen Time [2] on iOS devices analyzes user behaviors based on weekly reports of app uses, suggesting usage limits for most-used apps. However, such long-term analyses fail to timely detect immediate usage scenarios that could cause distraction or lead to life-threatening situations.

To enable timely recommendations, it is necessary to promptly and accurately infer the user context of a mobile device usage scenario. Existing studies of context inference use various sensors to derive user contexts such as locations [3], physical activities [4], and the surrounding environment [5], [9]. For instance, CUPID [3] explores WiFi signals to estimate Angle-of-Arrival for calculating the user's coordination

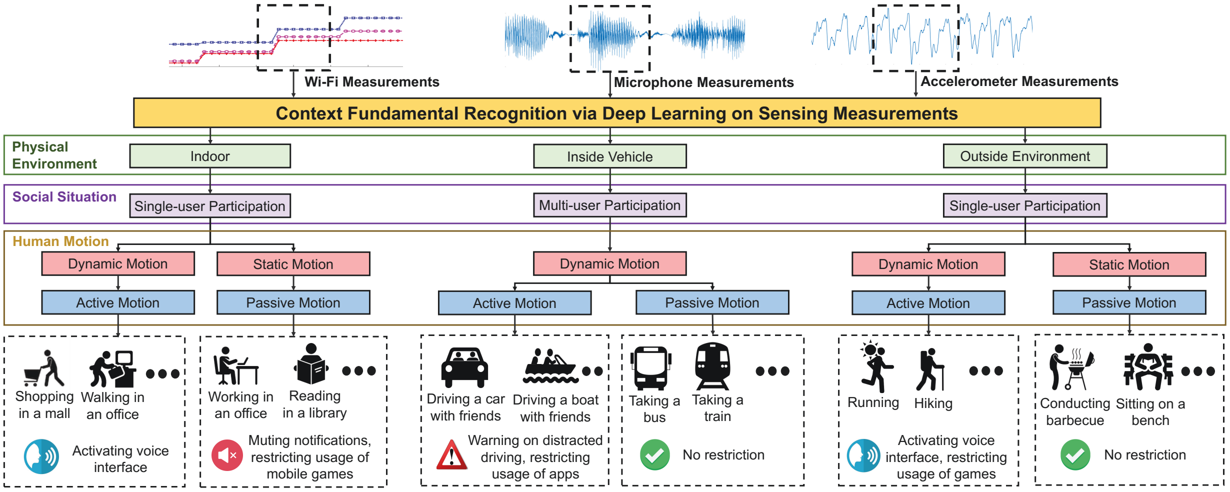


Fig. 2. Examples of providing device usage recommendations through recognizing context fundamentals.

in indoor environments. As another instance, FitCoach [4] exploits sensing measurements from the accelerometer of wearables/smartphones to recognize exercise types and further provide fitness assistance. Each of these studies can recognize a single type user context with great details. However, these approaches could not characterize the users' overall situations toward controlling their mobile device usage.

To perform context inference, existing techniques mainly rely on building profiles of built-in sensor readings (e.g., light sensor, accelerometer, gyroscope, and magnetometer) to train machine-learning-based models [10], [11]. However, since the sensor readings carry substantial information specific to the user/environment, a trained model usually will not work well if moving to a new environment, which involves extra profiling efforts in practical deployment. Another line of research works proposes to measure physical properties (e.g., Geolocation, delay of RF signals) embedded in user contexts. These approaches however involve energy-hungry sensors (e.g., GPS [12]), dedicated equipment (e.g., shoe-mounted initial sensors [6]), or infrastructures in the environment (e.g., wireless access points [7]).

Different from previous studies, we develop a generalized context inference framework that derives a much larger-scale of user contexts via identifying underlying context fundamentals, i.e., physical environment, social situation, and human motion. Our solution also derives transferable representations from sensor readings associated with various context fundamentals, which significantly reduces the profiling efforts across different environments, devices, and individuals. By using a set of low-power readily-available sensors, our framework could be easily integrated into any mobile devices to provide timely device usage recommendations.

III. System Design

A. Challenges

Deriving Context Fundamentals. User contexts of mobile device usage scenarios are rich, unpredictable, and hard to define. It is thus necessary to characterize these contexts with more general contextual information, i.e., context fundamentals, which uniformly characterize diverse usage behaviors. However, deriving reliable sensing features to characterize such general contextual information is challenging.

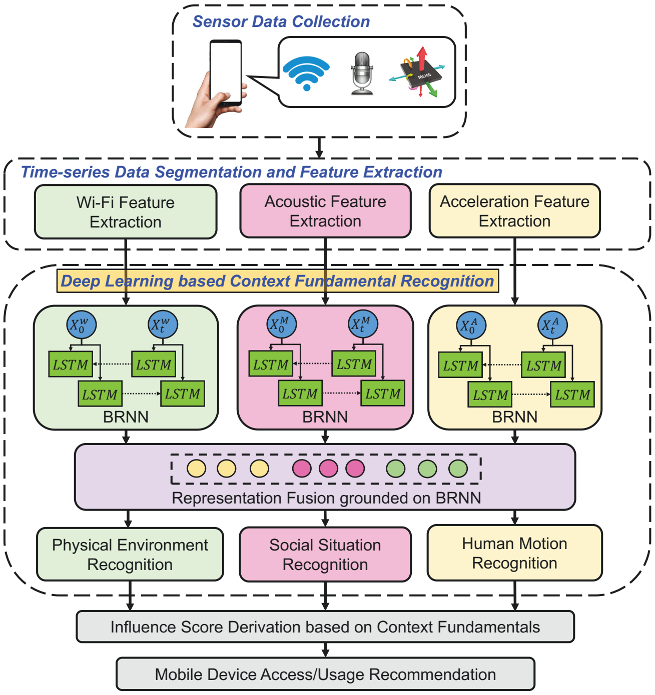

Sensor Data Collection
Time-series Data Segmentation and Feature Extraction
Wi-Fi Feature Extraction
Acoustic Feature Extraction
Acceleration Feature Extraction
Deep Learning based Context Fundamental Recognition
X_0^W
X_t^W
X_0^M
X_t^M
X_0^A
X_t^A
LSTM
BRNN
BRNN
BRNN
Representation Fusion grounded on BRNN
Physical Environment Recognition
Social Situation Recognition
Human Motion Recognition
Influence Score Derivation based on Context Fundamentals
Mobile Device Access/Usage Recommendation

Fig. 3. Overview of the proposed context inference framework.

Deriving Transferable Representations. The sensor data used for constructing the DNN model usually carries substantial information that is specific to environments, devices, and human subjects, making the model ineffective when applied to new environments/devices/users. To address this issue, the framework should be able to derive transferable representations for identifying context fundamentals.

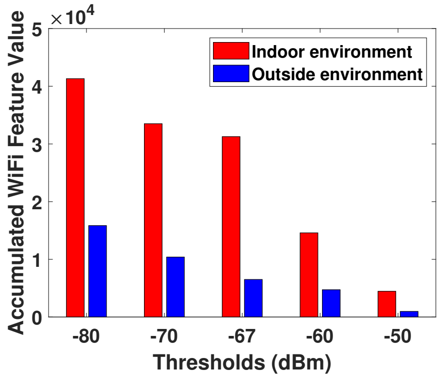


Fig. 4. Accumulated WiFi feature values under indoor and outside environments in 30 minutes.

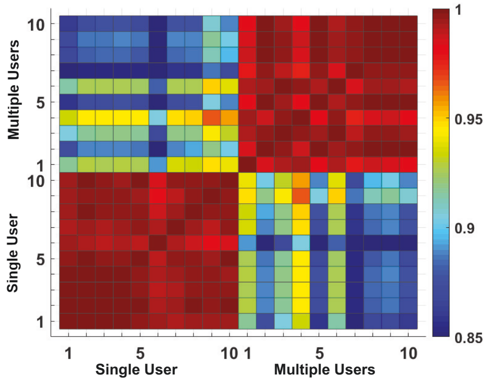


Fig. 5. Pairwise Pearson Correlation of MFCC features for single/multi-user.

Robustness to Input Variations. To enable timely context inference, it is essential to correctly identify context fundamentals from any segment of sensor data of a general user context. However, the temporal pattern in a data segment could be mismatched with that learned by the DNN model. To enable robust context inference, the proposed framework should have the capability to extract representations that are resilient to such pattern variations.

B. System Overview

We aim to design a generalized context inference framework that can uniformly interpret the immediate context of the usage scenario of a mobile device. This can enable timely device access recommendations (e.g., restricting access to games, muting notifications) when inappropriate usage scenarios are detected. As illustrated in Figure 3, the proposed framework monitors sensing measurements of WiFi module, microphone, and accelerometer, which possess extremely low-power consumption and are always enabled in the background processes [13]. To enable timely context inference, the framework exploits a sliding time window to segment and cache the latest observed sensor data. Then, it divides the sensor readings within the time window into a set of data clips and organizes these clips in a time sequence to preserve embedded temporal patterns. To derive context fundamentals, representative features regarding the three sensors are extracted to characterize physical environment around the user, interactions between people, and human movements.

The extracted features are then fed into the core component of our framework, *Deep Learning-based Context Fundamental Recognition*. Compared to existing context inference approaches, the proposed DNN architecture derives context fundamentals by learning reliable representations and enables the transferability over diverse environments, devices, and people. In particular, the DNN architecture consists: 1) bidirectional recurrent neural network (BRNN), 2) a fusion layer to concatenate the representations of different sensor modalities, 3) classifiers to recognize the context fundamentals (i.e., physical environment, social situation, and human motion). To ensure low power consumption on mobile devices, our DNN architecture exploits a lightweight structure without stacked BRNN or fully-connected layers. The lightweight structure could also help to preserve the transferability which tends to drop significantly in higher layers with increasing domain discrepancy [14].

Specifically, the BRNN connects a set of Long Short-Term Memory (LSTM) units to model and extract the temporal patterns embedded in the context fundamentals. Such a recursive structure of BRNN also enables temporal pattern recognition on arbitrary input feature sequences that might either be shifted or scaled. It thus enables context fundamental recognition with any segments of sensor data from the user context. To comprehensively interpret users' contexts, our DNN architecture further fuses the representations of the three sensing modalities by concatenating the outputs of the BRNNs. Based on the fused representations, three classifiers are employed to recognize the context fundamentals. Finally, using mobile device usage recommendation as an application example, our framework quantifies the degree of influence based on the derived context fundamentals, and provides suggestions on mobile device access through a recommendation policy.

IV. Context Fundamental for Defining Inappropriate Mobile Device Usage Scenarios

A. Inappropriate Mobile Device Usage Scenarios

The usage scenario of a mobile device is defined as inappropriate when it is involved at least one of the following situations: 1) physically inappropriate environment [15], 2) socially inappropriate situation [16], and 3) human motion required considerable attention [17] (e.g., driving a vehicle). All three situations are associated with the context that the user is in, and thus it is crucial to understand the context before judging whether a usage scenario is appropriate or not. Different from existing mobile device overuse study [2], which relies on long-term analysis, our work focuses on determining inappropriate usage scenarios via inferring immediate user contexts, which can further be used to provide usage recommendation. To systematically detect a broad range of inappropriate usage scenarios, our framework needs to derive large-scale user contexts.

B. Context Fundamentals

We propose to use *context fundamentals*, i.e., physical environment, social situation, and human motion, to jointly describe a user context. These context fundamentals are designed based on the inappropriate usage scenarios of mobile devices [16]. As illustrated in Figure 2, our framework first derives the context fundamentals from collected sensor data via deep learning and then jointly considers the derived context fundamentals for context inference. A variety of device access recommendations (e.g., restricting access to mobile games/apps) could then be provided.

Physical Environment. To reflect the general environments around users, we define the physical environment as the space/surrounding where the mobile user resides in. In particular, we categorize the user's physical environments as *Indoor* (e.g., library, shopping mall), *Inside Vehicle* (e.g., car, bus, train) and *Outside Environment* (e.g., street).

Social Situation. To differentiate the social settings that the mobile user is involving in, we categorize the associated situations as *Single-user participation* (e.g., working, studying alone), or *Multi-user Participation* (e.g., attending a party, meeting with colleagues [18]).

Human Motion. Human motion is used to characterize users' involvement and location displacement. In particular, we categorize human motion as *Dynamic/Static Motion* (e.g., sitting/sleeping or walking/running) and *Active/Passive Motion* (e.g., traveling as a passenger or operating a vehicle).

V. Time Series Data Segmentation and Feature Extraction

A. Data Segmentation

To enable timely context inference, our framework applies a sliding time window on the data streams of the three sensors (i.e., Wi-Fi sensor, accelerometer, and microphone) to cache the latest observed sensing measurements. We choose a window length of 5-second based on an empirical study on the computation overhead and the context inference accuracy. In order to preserve temporal patterns of the context fundamentals, our framework divides the sensing measurements of the time window into u data clips and organizes these clips in a time sequence. From each data clip, one set of features respecting to the three sensing modalities are extracted.

B. Feature Extraction

Characterizing context fundamentals within diverse general user contexts is challenging due to the unpredictable usage behaviors and noisy sensor readings. To consistently capture contextual fundamentals, we propose to derive three sets of representative features from WiFi module, microphone, and accelerometer. Exacting features could also compress the high-volume data clips, reducing computational complexity of data processing in DNN. We denote the extracted WiFi, acoustic, and acceleration features from the data clips as $\{W, M, A\}$, e.g., $W = \{w_1, w_2, ..., w_u\}$.

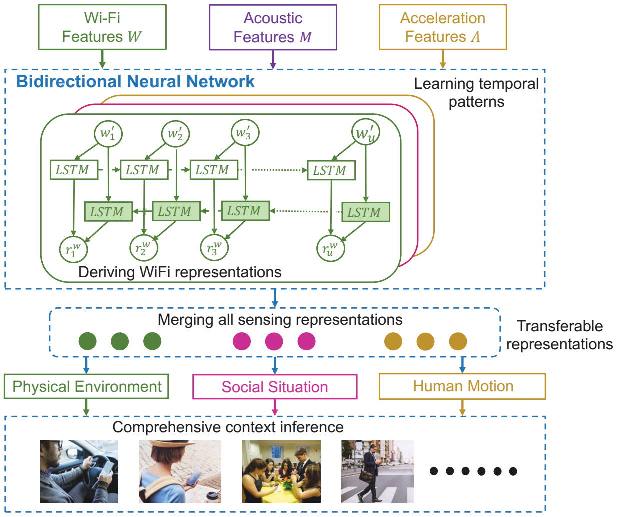

Wi-Fi Features W
Acoustic Features M
Acceleration Features A
Bidirectional Neural Network
Learning temporal patterns
w'_1
w'_2
w'_3
w'_u
LSTM
r^w_1
r^w_2
r^w_3
r^w_u
Deriving WiFi representations
Merging all sensing representations
Transferable representations
Physical Environment
Social Situation
Human Motion
Comprehensive context inference
• • • • • •

Fig. 6. The proposed deep learning architecture derives sensing representations with bidirectional long short-term memory (LSTM).

WiFi Feature Extraction. Almost all indoor environments are equipped with WiFi infrastructures (e.g., access point, IoT devices), which periodically generate beacon signals to announce their presence. The magnitude of the beacon signal could reveal the physical environments (e.g., indoor and outside environment), while its pattern in time domain could capture the movements (e.g., walking and remaining stationary) of a user. Unlike the previous approach [19] that relies on profiling MAC addresses of access points in specific environments, our framework extracts general intensity of the WiFi traffic via analyzing received signal strength (RSS) of the beacon signals. The WiFi feature can be derived as: $n(\delta) = \sum_{i=1}^{N} m(i)$, where i is the index of a WiFi infrastructure and $m(i) = 1$ if $RSS(i)$ is greater than a predefined threshold δ and $m(i) = 0$ otherwise. N denotes the total number of detected WiFi infrastructures. Note that we exclude public WiFi APs (e.g., Xfinity, Spectrum WiFi, optimum) to avoid the cases where outdoor environments have higher WiFi densities than indoor environments. Based on different intensity of WiFi traffic [20], we extract 5 WiFi features by using 5 thresholds, -50, -60, -67, -70, -80. Figure 4 shows the accumulated values of WiFi features in 30 minutes. We observe that the accumulated feature values of indoor environments (i.e., a university building) are much greater than that of the outdoor environments (i.e., a park).

Acoustic Feature Extraction. Our framework utilizes Mel-frequency cepstral coefficient (MFCC) as the acoustic feature to characterize social situations and physical environments. Existing work [21] shows that MFCC-based features can be used to estimate the number of speakers in a conversation, it thus has the capability to reveal social interactions. It could also capture acoustic characteristics of surrounding environments (e.g., engine noise of vehicles). The number of filterbank is set to 16, and the 8^{th} order cepstral coefficients are derived in each Hanning window with the same size as the audio clip. Figure 5 shows the Pearson correlation coefficient between any two MFCC features derived under single-user and multiple-user participation. We observe that the MFCC

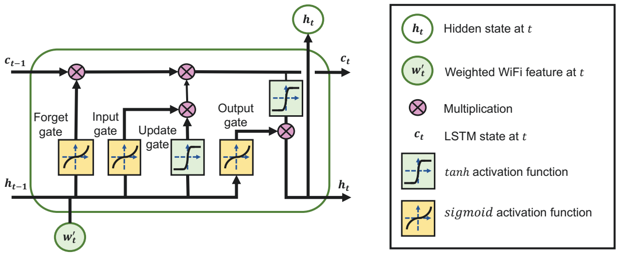


Fig. 7. An illustration of the LSTM structure to learn temporal patterns embedded in the weighted WiFi features.

features for the same group present a higher correlation than that between different groups.

Acceleration Feature Extraction. To reveal human motion and physical environments (e.g., inside a moving vehicle), three acceleration features, velocity, range, variance, are extracted from each accelerometer axis. Specifically, velocity represents the mobility and the displacement of human body. Range is the absolute difference between maximum and minimum values in the data clip. It reflects the intensity of the human body motion, which is distinctive between active (i.e., riding a bicycle) and passive motion (i.e., taking public transportation). Moreover, we use variance as a complement to characterize human motion.

VI. Deep Learning Based Context Fundamental Recognition

It is highly desired to have a context inference framework that could be directly applied to a user's mobile device without extra profiling efforts. To develop such a framework, we propose to derive transferable representations of context fundamentals that are resilient to the distortions caused by changes of environments, devices, and people. In addition, to avoid causing computational overhead to mobile devices (e.g., smartphone), the proposed DNN model needs to have a low run-time computational cost. As depicted in Figure 6, we present a lightweight DNN architecture to derive transferable representations by learning temporal patterns embedded in feature sequences. The architecture utilizes bidirectional a recurrent neural network (BRNN) with long short-term memory (LSTM) to learn temporal patterns of context fundamentals. Finally, the representations are merged with a fusion mechanism and fed to three classifiers for identifying the context fundamentals.

A. Bidirectional Recurrent Neural Network

The proposed BRNN module exploits a recurrent structure that connects consecutive LSTM units to learn temporal patterns which are transferable among different general user contexts. With this structure, the derived representations could maximize the correlations between the same type of context fundamentals even when the input feature sequences have significant misalignments with the profile. As shown in Figure 6, the BRNN module takes weighted WiFi features W' as input and derives WiFi representations with two hidden layers, i.e., a forward layer and a backward layer defined as:

$$\begin{aligned}\vec{h}_t &= H(\overset{\rightarrow}{\lambda}_1 w_t + \overset{\rightarrow}{\lambda}_2 \vec{h}_{t-1} + \vec{b}),\\ \overset{\leftarrow}{h}_t &= H(\overset{\leftarrow}{\lambda}_1 w_t + \overset{\leftarrow}{\lambda}_2 \overset{\leftarrow}{h}_{t-1} + \overset{\leftarrow}{b}).\end{aligned} \tag{1}$$

The derived WiFi representation for the clip at t could be represented as $r_t^w = \overline{\lambda}_1 \rightarrow h_t + \overline{\lambda}_2 \leftarrow h_t + \overline{b}$. The λ terms (e.g., $\{\overset{\rightarrow}{\lambda}_1\}$) denote weight matrices, and the b terms (e.g., $\{\vec{b}\}$) represent bias vectors. H is an activation function implemented with Long-Short Term Memory (LSTM) units, which is well-suited to capture inherited temporal patterns and addresses the vanishing gradient and error blowing up problems [22]. Figure 7 shows the structure of LSTM which takes the last LSTM state c_{t-1} and the hidden state h_{t-1} from the last clip, along with a weighted WiFi feature w'_t as input. By utilizing the input gate, forget gate, update gate and output gate, the LSTM unit could remember values over arbitrary time intervals, which could facilitate learning on temporal patterns. To avoid computational complexity while ensuring context inference performance, we use only 64 LSTM units for both forward and backward layers of the BRNN. We train the model by using ADAM optimizer with a learning rate of 10% and 300 epochs.

TABLE I
Examples of Influence scores based on the context fundamentals.

Context Fundamental	Score	Context Fundamental	Score
Indoor Environment	$S_1 = 0$	Inside Vehicle	$S_1 = 2$
Outside Environment	$S_1 = 1$	Single-user participation	$S_2 = 0$
Multiple-user participation	$S_2 = 2$	Static Motion	$S_3 = 0$
Dynamic Motion	$S_3 = 1$	Active Motion	$S_4 = 2$
Passive Motion	$S_4 = 0$		

B. Transferable Representation Derivation

Representation Fusion. To comprehensively interpret a user context, we exploit a fusion layer to combine the representations of the three sensing modalities for context inference. The fusion layer concatenates the output of the three BRNNs:

$$R = R^w \oplus R^a \oplus R^m, \tag{2}$$

where R^w, R^a, and R^m denote the representations of WiFi, acoustic, acceleration features, respectively. $\oplus$ represents the concatenation process and R is the fused representation.

Context Fundamental Recognition. The BRNN can only derive sensing representations, and thus classifiers are needed to recognize the context fundamentals. Specifically, we use softmax layer using the output of fusion layer for classification. The outputs of each softmax function characterize the probability distribution over context fundamentals illustrated in Section IV. Then a representation R will be classified as class k if it has the largest probability.

VII. Mobile Device Usage Recommendation

There are many mobile applications (e.g., health care, pervasive games, multimedia apps) could benefit from our generalized user context inference framework. In this section,

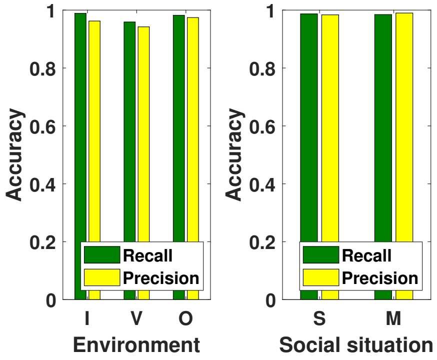


Fig. 8. Performance of our system on recognizing physical environments/social situations.

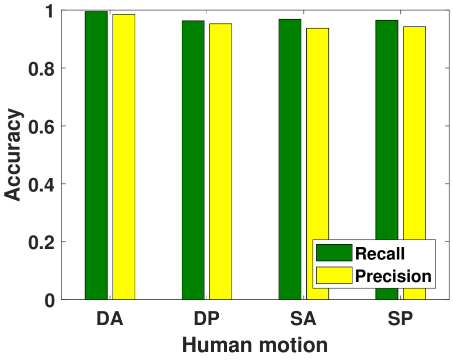


Fig. 9. Performance of our system on recognizing human motions.

we use mobile device usage recommendation as an application example. Our goal is to design a policy that provides recommendation based on the current usage scenario of a mobile device. An inference score is designed to quantify that how likely an inferred user context would lead to distractions or dangerous situations. For example, a high inference score is expected in the context of texting or answering phone calls while driving a vehicle, which could lead to life-threatening accidents. While a user should be allowed to access all mobile functionalities in his leisure time. We define the inference score as follow: $I = \sum_{k=1}^{K} S_k$, where $K = 4$ represents the four sets of context fundamentals. An example of inference score based on the context fundamentals is detailed in Table I. Different access recommendations can be provided based on the influence score with the threshold γ. For example, if the inference score is over the threshold $\gamma_2 = 5$, our framework would warn that the mobile device usage in the current user context (e.g., driving a car) is strictly prohibited.

VIII. Performance Evaluation

A. Experimental Methodology

Devices. We evaluate the proposed framework with two types of mobile devices, smartphones (i.e., Motorola Nexus 6) and tablets (i.e., Amazon FireHD), that run on the Android operating system. Both devices are equipped with a WiFi module, a microphone, a 3-axes accelerometer. The data collection processes are implemented based on the Android platform with different built-in libraries, i.e., SensorManager, MediaRecorder, and WifiManager, respecting to the three sensing modalities. Specifically, the sampling rates for the three sensors are set as $100Hz$ (i.e., accelerometer), $8000Hz$ (i.e., microphone), and $10Hz$ (i.e., WiFi module), respectively.

Data Collection. Our framework is evaluated with 10 volunteers (8 males and 2 females) over a period of 7 months. The volunteers are of ages from 25 to 35 years old and are mainly graduate students/university researchers. In particular, we collect data of 12 representative user contexts, where the context fundamentals are detailed in Table II. The volunteers are asked to record the ground truth of the user contexts/context fundamentals in the experiments. During the data collection, the volunteers can either hold the device or place it in the pocket. In total, we collect $84,386$ data segments, which are obtained via applying a sliding time window (i.e., duration 5s, step size 1s) on the time-series sensing measurements. All data segments are shuffled for evaluating the system's capability given arbitrary data segments of the user contexts.

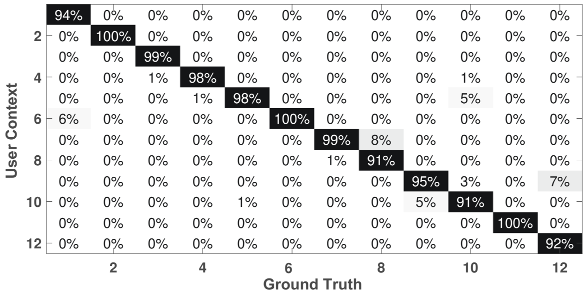
Metrics. The feature extraction mechanism and the DNN model are implemented by using Python programming language with Keras API which allows building customized neural networks. To evaluate system performance, we define three different metrics: *precision/accuracy*; *recall*; *confusion matrix*. Particularly, precision of a context fundamental c is defined as $Precision_c = N_c^T/(N_c^T + M_c^F)$, where N_c^T is the number of sensor segments correctly recognized as the context fundamental c and M_c^F represents the number of segments corresponding to other context fundamentals which are mistakenly recognized as c. Recall of a context fundamental c is defined as $Recall_c = N_c^T/N_c$, where N_c is the number of all segments from the context fundamental c. For the confusion matrix, each row indicates the derived user contexts and each column represents the ground truth. The entries located in the diagonal of the matrix is the percentage of correct predictions.

B. Performance of Context Fundamental/General user Context Inference

Context Fundamental Derivation. We first evaluate the performance of our system on deriving context fundamentals. Figure 8 depicts the precision/recall of recognizing physical environments and social situations. The physical environments include indoor (I), inside vehicle (V), and outside environment (O), while the social situations involve single-user participation (S) and multi-user participation (M). It is encouraging to find that the physical environments and social situations can be accurately recognized with the lowest precision and recall as 95.7% and 97.4%, respectively. In Figure 9, we show precision and recall of four motion contexts: dynamic and active context (DA), dynamic and passive context (DP), static and active context (SA), static and passive context (SP). We observe that our system achieves high average precision (i.e., 95.32%) and recall (i.e., 96.41%). Through the above examination, we find

TABLE II
CHARACTERIZING 12 REPRESENTATIVE USER CONTEXTS WITH CONTEXT FUNDAMENTALS.

ID	User Context	Physical Environment	Social Situation	Human Motion
1	Study/working alone in an office	Indoor	Single User	Static&Passive Motion
2	Running/conducting exercises in a Gym	Indoor	Single User	Static&Active Motion
3	Attending a meeting/talking to friends in an office	Indoor	Multi-user	Static&Passive Motion
4	Walking in a campus building with friends	Indoor	Multi-user	Dynamic&Active Motion
5	Taking a rest in a park	Outside Environment	Single User	Static&Passive Motion
6	Running in a park	Outside Environment	Single User	Dynamic&Active Motion
7	walking on the street with friends	Outside Environment	Multi-user	Dynamic&Active Motion
8	Attending an outdoor party	Outside Environment	Multi-user	Static&Passive Motion
9	Driving a car	Inside Vehicle	Single User	Dynamic&Active Motion
10	Taking a bus/train	Inside Vehicle	Single User	Dynamic&Passive Motion
11	Driving a car with friends	Inside Vehicle	Multi-user	Dynamic&Active Motion
12	Taking a bus/train with friends	Inside Vehicle	Multi-user	Dynamic&Passive Motion



		2		4		6		8		10		12
	94%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	0%	0%	99%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4	0%	0%	1%	98%	0%	0%	0%	0%	0%	1%	0%	0%
	0%	0%	0%	1%	98%	0%	0%	0%	0%	5%	0%	0%
6	6%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%
	0%	0%	0%	0%	0%	0%	99%	8%	0%	0%	0%	0%
8	0%	0%	0%	0%	0%	0%	1%	91%	0%	0%	0%	0%
	0%	0%	0%	0%	0%	0%	0%	0%	95%	3%	0%	7%
10	0%	0%	0%	0%	1%	0%	0%	0%	5%	91%	0%	0%
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%
12	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%

Fig. 10. Performance of our system on inferring general user contexts by fusing the derived context fundamentals.

that the proposed system is very effective in recognizing the context fundamentals.

Impacts of Window Length. Generally, a larger length sliding window would improve the system performance, while it could also increase the computational cost in context inference. To study the impacts of window length, we use lengths of 2-10 (seconds) and examine performance with each window size for 10 trials, where the data for training and testing are randomly selected in each trial. Given the window length of $5s$, our system achieves the best performance on deriving context fundamentals of physical environment, social situation, and human motion, i.e., 97.82%, 96.6%, 94.71%, respectively.

General user Context Inference. We present the performance of the proposed system on inferring user contexts, where each context is uniquely inferred by integrating the derived context fundamentals. For example, working along in an office, can be characterized as indoor, single user, and static&passive motion. As shown in Figure 10, we observe that our system achieves over 91% context inference accuracies for recognizing all 12 contexts. Especially, the accuracies of context 2, 6, and 11 reach 100%. The average context inference accuracy is 95.96% with a standard deviation of 3.84%. The above results confirm that our system is highly effective in inferring general user contexts.

C. Transferability of the Proposed Framework

We iteratively exclude samples of one context which is one type of user environment from the training dataset, and then use data of the other contexts to train the DNN model.

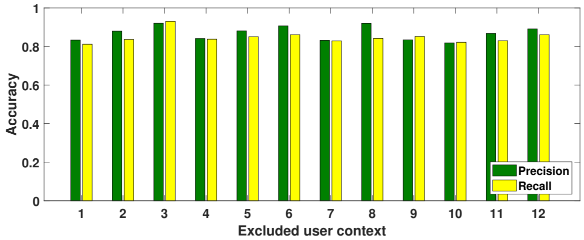


Fig. 11. Assessing the transferability over user contexts by iteratively excluding data of one user context from the training dataset and using the representations learned from the other user contexts for context inference.

A user context can only be correctly inferred if all the underlying context fundamentals are correctly identified. As shown in Figure 11, we find that most of the user contexts can be correctly inferred even corresponding data are excluded from the training dataset. The lowest recall and precision are 82.85% and 81.89%, and the average recall and precision are 86.71% and 84.54%. The results validate that even without profiling a user environment, our framework could still infer a user context with transferable representations learned from the other environments.

D. Effectiveness of Representation Fusion

Representation Fusion. Next, we study the effectiveness of the proposed representation fusion grounded on BRNN. Figure 12 depicts the accuracies of context inference with and without the fusion mechanism. When the fusion mechanism is removed, the system utilizes representations of the WiFi, acoustic, and acceleration features to infer physical environment, social situation, and human motion, respectively. We can observe that the fusion mechanism can help to achieve higher context inference accuracies. This is because the fused representations could capture more comprehensive characteristics of each type of context fundamental.

IX. DISCUSSION

A. Framework Scalability

Serving as an initial study on deriving context fundamentals, our framework has a high potential to be extended to various context-driven applications such as phone addition studies, augmented reality games, and multimedia services. By adding

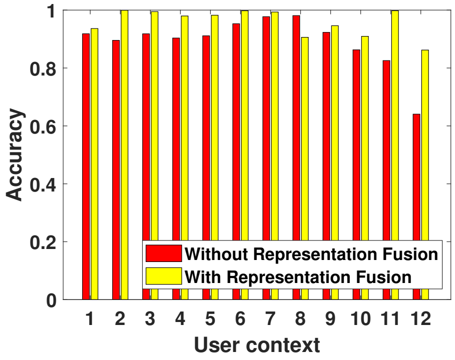


Fig. 12. Effectiveness of the proposed representation fusion mechanism.

a single entry of context fundamental based on the application requirement, the proposed framework could be easily applied to an additional set of general user contexts, which are the combinations of the new entry and all existing context fundamentals. This greatly reduces the profiling efforts compared to existing approaches.

B. Energy Consumption

Our system is a lightweight context inference framework with low energy consumption. The most power-consuming tasks of our system are sensor data collection and DNN-based context inference. Specifically, the power consumption of WiFi module, microphone, and accelerometer are $21mW$, $135mW$, $101mW$ [23], respectably. The overall power consumption is $257mW$, which is much lower than that of the GPS sensor, i.e., $400mW$. With the transferability, the DNN model avoids the power-consuming training process. Since our DNN model adopts a light structure without stacking BRNN/ReLU layers, it has low run-time energy consumption, especially when using the new generation of AI chips (e.g., A12 Bionic) on the latest smartphones.

X. Conclusion

In this paper, we propose the first generalized context inference framework, which supports timely usage recommendations using low-power sensors in mobile devices. By identifying context fundamentals, our framework can uniformly describe large-scale user contexts associated with inappropriate usage scenarios of a mobile device. Furthermore, we design a DNN architecture to learn transferable representations of the context fundamentals to mitigate training efforts across various contexts, devices, and individuals. Additionally, the framework quantifies the degree of influence for the derived context fundamentals and provides warnings/feedbacks for mobile device usages. Extensive experiments demonstrate that our framework could achieve remarkable performance on context inference while offering the transferability over different environments, devices, and users.

References

[1] S. Perez, “Siempo’s new app will break your smartphone addiction,” 2018, https://techcrunch.com/2018/05/19/siempos-new-app-will-break-your-smartphone-addiction/.

[2] iPhone User Guide, “Set screen time, allowances, and limits on iphone,” https://support.apple.com/guide/iphone/set-screen-time-allowances-and-limits-iph9b66575d5/ios.

[3] S. Sen, J. Lee, K.-H. Kim, and P. Congdon, “Avoiding multipath to revive inbuilding wifi localization,” in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 2013, pp. 249–262.

[4] X. Guo, J. Liu, and Y. Chen, “Fitcoach: Virtual fitness coach empowered by wearable mobile devices,” in *2017 IEEE International Conference on Computer Communications (IEEE INFOCOM)*, 2017, pp. 1–9.

[5] P. Zhou, Y. Zheng, Z. Li, M. Li, and G. Shen, “Iodetector: A generic service for indoor outdoor detection,” in *Proceedings of the 10th acm conference on embedded network sensor systems*. ACM, 2012, pp. 113–126.

[6] S. Jain, C. Borgiattino, Y. Ren, M. Gruteser, Y. Chen, and C. F. Chiasserini, “Lookup: Enabling pedestrian safety services via shoe sensing,” in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2015, pp. 257–271.

[7] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, “Widir: walking direction estimation using wireless signals,” in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2016, pp. 351–362.

[8] E. Livni, “Cut your phone dependence with an app that plants trees as a reward,” 2017, https://qz.com/1112713/urb-tech-dependence-with-an-app-that-plants-trees-as-a-reward/.

[9] D. H. Kim, J. Hightower, R. Govindan, and D. Estrin, “Discovering semantically meaningful places from pervasive rf-beacons,” in *Proceedings of the 11th international conference on Ubiquitous computing (ACM UbiComp)*, 2009, pp. 21–30.

[10] G. M. Weiss and J. Lockhart, “The impact of personalization on smartphone-based activity recognition,” in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[11] S. Hemminki, P. Nurmi, and S. Tarkoma, “Accelerometer-based transportation mode detection on smartphones,” in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (ACM SenSys)*, 2013, p. 13.

[12] G. Xiao, Z. Juan, and C. Zhang, “Travel mode detection based on gps track data and bayesian networks,” *Computers, Environment and Urban Systems*, vol. 54, pp. 14–22, 2015.

[13] G. Milette and A. Stroud, *Professional Android sensor programming*. John Wiley & Sons, 2012.

[14] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *International Conference on Machine Learning (ICML)*, 2015, pp. 97–105.

[15] M. Kwon, J.-Y. Lee, W.-Y. Won, J.-W. Park, J.-A. Min, C. Hahn, X. Gu, J.-H. Choi, and D.-J. Kim, “Development and validation of a smartphone addiction scale (sas),” *PloS one*, vol. 8, no. 2, p. e56936, 2013.

[16] M. Takao, S. Takahashi, and M. Kitamura, “Addictive personality and problematic mobile phone use,” *CyberPsychology & Behavior*, vol. 12, no. 5, pp. 501–507, 2009.

[17] D. C. Schwebel, D. Stavrinos, K. W. Byington, T. Davis, E. E. O’Neal, and D. De Jong, “Distraction and pedestrian safety: how talking on the phone, texting, and listening to music impact crossing the street,” *Accident Analysis & Prevention*, vol. 45, pp. 266–271, 2012.

[18] C. Moser, S. Y. Schoenebeck, and K. Reinecke, “Technology at the table: Attitudes about mobile phone use at mealtimes,” in *Proceedings of the Conference on Human Factors in Computing Systems (ACM CHI)*, 2016, pp. 1881–1892.

[19] M. Azizyan, I. Constandache, and R. R. Choudhury, “Surroundsense: Mobile phone localization via ambience fingerprinting,” in *Proceedings of the 15th Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*, 2009, pp. 261–272.

[20] eyeSaaS, “Wi-fi signal strength: What is a good signal and how do you measure it,” 2019, https://eyesaas.com/wi-fi-signal-strength/.

[21] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.-F. Chen, J. Li, and B. Firner, “Crowd++: unsupervised speaker count with smartphones,” in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 43–52.

[22] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.”

[23] S. Tarkoma, M. Siekkinen, E. Lagerspetz, and Y. Xiao, *Smartphone energy consumption: modeling and optimization*. Cambridge University Press, 2014.