mFingerprint: Privacy-Preserving User Modeling with Multimodal Mobile Device Footprints

Haipeng Zhang*[1], Zhixian Yan[2], Jun Yang[2], Emmanuel Munguia Tapia[2], and David J. Crandall[1]

[1] School of Informatics & Computing, Indiana University, Bloomington, IN, USA
{zhanhaip,djcran}@indiana.edu
[2] Samsung Research America, San Jose, CA, USA
{zhixian.yan,j3.yang,e.tapia}@samsung.com

Abstract. Mobile devices collect a variety of information about their environments, recording "digital footprints" about the locations and activities of their human owners. These footprints come from physical sensors such as GPS, WiFi, and Bluetooth, as well as social behavior logs like phone calls, application usage, etc. Existing studies analyze mobile device footprints to infer daily activities like driving/running/walking, etc. and social contexts such as personality traits and emotional states. In this paper, we propose a different approach that uses multimodal mobile sensor and log data to build a novel user modeling framework called mFingerprint that can effectively and uniquely depict users. mFingerprint does not expose raw sensitive information from the mobile device, e.g., the exact location, WiFi access points, or apps installed, but computes privacy-preserving statistical features to model the user. These descriptive features obscure sensitive information, and thus can be shared, transmitted, and reused with fewer privacy concerns. By testing on 22 users' mobile phone data collected over 2 months, we demonstrate the effectiveness of mFingerprint in user modeling and identification, with our proposed statistics achieving 81% accuracy across 22 users over 10-day intervals.

1 Introduction

Mobile devices such as smartphones and tablets have become powerful people-centric sensing devices thanks to embedded sensors such as GPS, Bluetooth, WiFi, accelerometer, touch, light, and many others. The devices have also advanced significantly in terms of computational capacity, memory and storage. These improvements have stimulated people-centric mobile applications, ranging from inferring and sharing real-time contexts such as location and activities [3,8] to identifying heterogeneous social behaviors of mobile users [6,9,11,15]. Most of these studies focus on using phone data to infer physical and social contexts for a particular user at a specific point in time. Additional studies have concentrated on analyzing long-term data from mobile devices to monitor trends and to establish predictive models of location [7] and app usage [12].

[*] This work was done while this author was a research intern at Samsung Research America - Silicon Valley.

In this paper, we take a significantly different view of mobile device data. Instead of focusing on real-time inference of contexts or social behaviors from phone sensors, we analyze multimodal mobile usage data to extract simple yet effective statistics that can uniquely represent mobile users. We construct a novel user modeling framework called 'mFingerprint' to define 'fingerprints' that try to uniquely identify users from mobile device data. Our experiments show the effectiveness of mFingerprint in modeling users and identifying them uniquely. In contrast to many user identification studies that try to identify users based on raw sensor data such as touch screen [2] and cell towers [5], mFingerprint computes high-level statistical features that do not disclose sensitive information from the phone, such as raw location, browser history, and application names. Therefore, applications can share, transmit, and use these descriptive privacy-preserving feature vectors to enable personalized services on mobile devices with fewer privacy concerns.

Research Challenges. It is non-trivial to build a system to identify users based on high level statistics of their mobile usage data, due to the following challenges: (1) Mobile devices collect a variety of multimodal data including logs from physical sensors such as GPS and accelerometer and application data like app usage and web browser history; how to choose sources to construct effective mobile fingerprints remains a question. (2) Mobile data is generated under complex real-life settings, introducing significant noise that demands robustness in processing. (3) In contrast to most existing offline data analysis approaches deployed on the server side, our mFingerprint framework focuses on designing lightweight energy-efficient algorithms that are able to run on mobile devices directly. (4) mFingerprint must avoid disclosing sensitive mobile data, such as geo locations from GPS, the applications installed, and URLs of website visited, to protect user's privacy.

Main Contributions. To address these challenges, this paper presents a novel approach for user modeling and identification based on digital footprints from mobile devices, with the following key contributions: (1) We build mFingerprint, a novel framework to analyze data from mobile devices and model users via digital footprints; (2) mFingerprint generates fingerprints from heterogeneous hardware sensors such as GPS, WiFi, and Bluetooth and soft sensors including app usage logs; (3) By designing a discriminative set of statistical features to capture mobile footprints, mFingerprint is able to identify users while preserving their privacy.

2 Related Work

Recently, user's digital footprints from mobile device have gained significant attention in various research areas such as mobile computing, data mining and social analysis.

In mobile computing and data mining, several studies build mobile systems to infer context offline or in real-time, such as detecting semantic locations (home and office) from GPS [14], identifying physical activities from accelerometer data [8], or estimating user's environmental properties such as crowdedness from Bluetooth [13] and noise-level [10]. These studies however focus on analyzing only single physical sensors to infer specific types of daily contexts and activities.

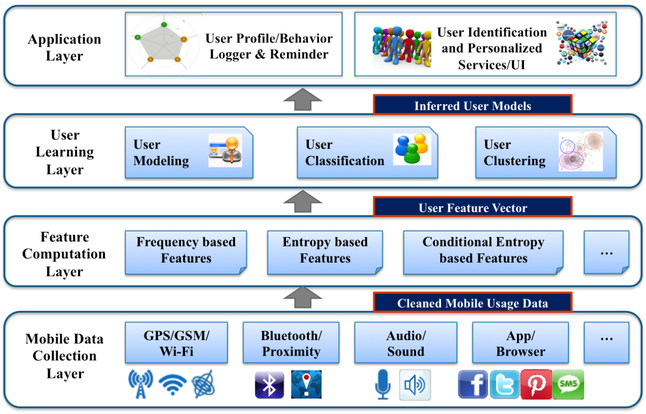


Fig. 1: The mFingerprint framework.

For social and behavioral analysis on mobile data, the recent focus is mainly on continuously monitoring user's daily social contexts, such as inferring emotions from audio [11], detecting mood from communication history and application usage patterns [9], predicting user's personality using various phone usage data [4,6], and estimating user's profile and sociability level using Bluetooth crowdedness [15]. The mFingerprint framework further extends these studies to analyze multimodal mobile footprints (including both soft usage data and physical sensor data) by extracting discriminative statistical features as a fingerprint to model the user and provide accurate user identification with privacy preservation.

3 mFingerprint System Overview

Fig. 1 shows the four layers of the mFingerprint framework. The bottom layer collects various sensor readings using hardware sensors that trace location (e.g., GPS, WiFi, and cell tower), proximity (Bluetooth and microphone). Furthermore, soft sensor data such as application usage and web browser history are also recorded. The second layer of mFingerprint computes a set of privacy-preserving statistical features from these digital footprints. In this paper, we particularly focus on designing frequency and entropy based statistical features to capture mobile device usage patterns. Such features flow up to the third layer for user learning, which includes building user models via the feature vectors, identifying users via classification methods, and grouping users into meaningful clusters via unsupervised learning. The forth layer is the application layer where various applications can be established, such as inferring user profile, logging mobile behaviors, and creating personalized services and user interfaces.

4 Footprint Feature Computation and User Identification

We focus on designing simple frequency and entropy-based statistical features to create users' "fingerprints" and evaluate the features' performance in user identification.

4.1 Frequency based Footprint Features

The number of devices and cell towers that are observed by a phone throughout the day provides information about the owner's environment. For example, a phone in a busy public place will likely observe many wireless devices while a phone in a moving car observes different cell towers over time. Meanwhile, a user's app usage patterns throughout the day tell us something about his or her daily routine. We thus propose simple frequency-based features that measure how much activity of four different types (Wifi, Cell towers, Bluetooth, and App usage) is observed at different time intervals throughout the day. More specifically, we divide time into T-minute time periods, and make observations about the phone's state every M minutes, with $M < T$ so that there are multiple observations per time period. In the i-th observation of time period t, we record: (1) the number of Wifi devices that are observed ($W^{t,i}$), (2) the number of cell phone towers that the phone is connected to ($C^{t,i}$), (3) the number of bluetooth devices that are seen ($B^{t,i}$), and the number of unique apps that have been used over the last m minutes ($A^{t,i}$). We then aggregate each of these observation types to produce four features in each time period:

$$F_W^t = \sum_i W^{t,i}, \qquad F_C^t = \sum_i C^{t,i}, \qquad F_B^t = \sum_i B^{t,i}, \qquad \text{and} \qquad F_A^t = \sum_i A^{t,i}.$$

A feature incorporating all of these features is simply the vector $F^t = [F_W^t \; F_C^t \; F_B^t \; F_A^t]$.

4.2 Entropy based Footprint Features

While the simple frequency features above give some insight into the environment of the phone, they ignore important evidence like the distribution of this activity. For example, in some environments a phone may see the same Wifi hotspot repeatedly through the day, while other environments may have an ever-changing set of nearby Wifi networks. To illustrate this, Fig. 2 compares observed frequency versus anonymized device IDs for two users across each of the four observation types, for a period of 10 days. We can observe that User 2 is less active in WiFi and cell mobility compared to User 1, but has more Bluetooth encounters and uses more diverse apps.

We thus propose using entropy of these distributions as an additional feature of our user fingerprints. The entropy feature summarizes the distribution over device IDs, but in a coarse way such that privacy concerns are minimized. For Wifi, let W_j^t denote the

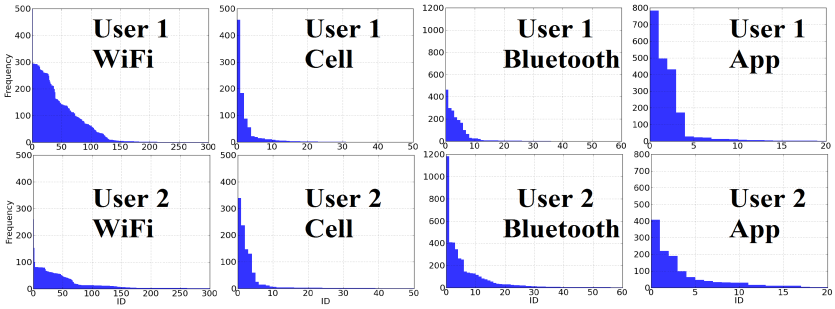


Fig. 2: Comparison of activity histograms for 2 users over 10 days. Y-axes are frequencies; X-axes are WiFi, Cell, Bluetooth, and App IDs.

number of times we observe wifi hotspot j during time period t. Then we define the Wifi entropy during time period t as,

$$E_W^t = -\sum_j \frac{W_j^t}{F_W^t} \log \frac{W_j^t}{F_W^t}.$$

Entropy features for Cell towers, Bluetooth, and Apps (E_C^t, E_B^t, and E_A^t, respectively) are computed in the same way, and we define a multimodal entropy feature vector $E^t = [E_W^t \ E_C^t \ E_B^t \ E_A^t]$, which incorporates all four perspectives.

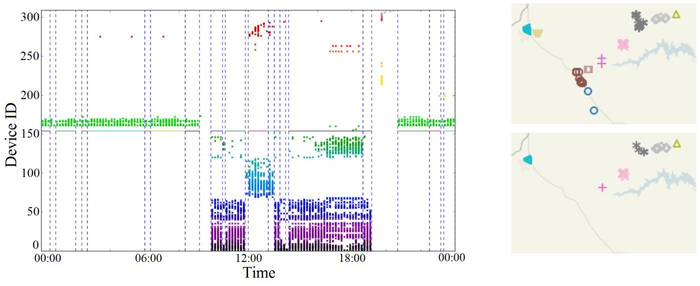
4.3 Conditional Entropy and Frequency based Footprint Features

In our system, we calculate the above entropy and frequency features conditioned on time and location. Intuitively at different times and at different locations, users have different patterns of application usage and surrounding devices (Bluetooth, WiFi, cell, etc.). For example, two users might have similar overall apps usage but one user always uses apps in the mornings, while the other uses them only in the afternoon. Two other users might have similar overall Bluetooth entropies but one might have more surrounding devices at work while the other observes the variety at a coffeeshop. Conditioning on time and space is thus useful to better differentiate users.

Conditional features on time. For the frequencies and entropies conditioned on time, we differentiate on time of a day and day of a week. Currently we distinguish between three fixed daily time intervals, mornings (0:00 - 8:59), working hours (9:00 - 17:59) and evenings (18:00 - 23:59), and two types of days, weekdays (Mon through Fri) and weekends (Sat and Sun). This gives five time periods over which we compute the conditional features. Future work might explore adaptive intervals instead.

Conditional features on location. We also compute frequency and entropy features conditioned on location. For each user, we filter and cluster their geo-locations in order to identify the top-k significant locations. From data collected at these k locations, we compute the conditional entropies and frequencies. There are two steps in finding significant locations: *Segmentation* and *Clustering*. In the segmentation step, we find periods of time when the phone appears to be stationary, by looking for time intervals when the IDs of surrounding devices are stable. In particular, we divide the data streams into 10-minute time frames and for each time frame, we record the IDs of the WiFi, Bluetooth and Cell towers. For adjacent time frames, we compute Jaccard similarity of the corresponding sets of IDs, $J(S_1, S_2) = |S_1 \cap S_2|/|S_1 \cup S_2|$, where S_1 and S_2 are sets of device IDs. If the similarity is larger than a threshold, we say that the device is stationary during the two time frames. Fig. 3(a) illustrates finding stationary and non-stationary periods according to WiFi readings. Similarly, we perform this on Bluetooth and Cell readings and take the union of all stationary time periods.

We then apply the DBSCAN clustering algorithm to the stationary segments in order to identify important locations. Fig. 3(b) shows the clustering results on the same data with and without non-stationary points. We see that noisy signals (e.g., location points moving along highways) have been removed by keeping only stationary data, which generates better and fewer clusters. Note that our mFingerprint system computes location based conditional features using anonymized cluster and device IDs, not the



(a) (b)

Fig. 3: Time segmentation and location clustering: (a) Finding stationary times using Wifi. (b) Location clustering using all (top) and only stationary (bottom) data.

original locations, to help preserve privacy. We choose $k = 2$ and statistics from the ith location will be compared with that from the location of the same rank across users.

5 Evaluation

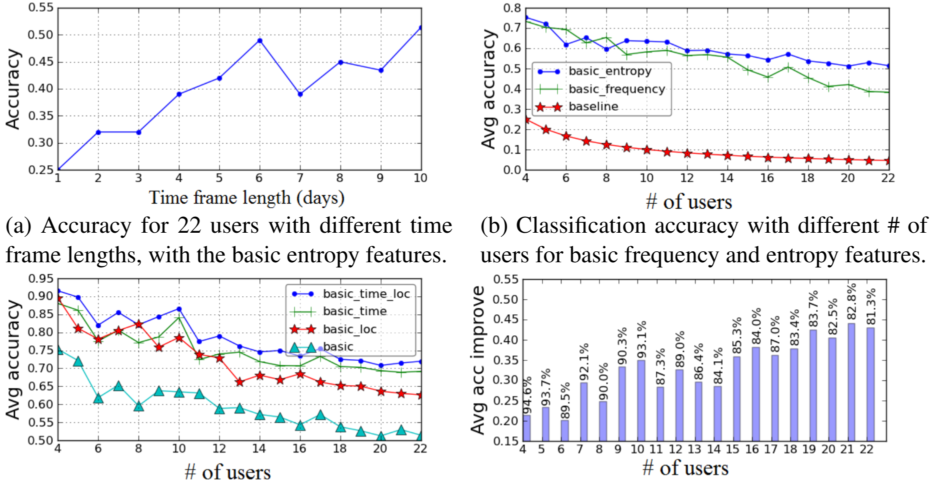
We define a user identification problem in order to test whether mFingerprint can uniquely characterize users. We pose this as a classification problem: *can we build a multi-class classifier trained on the entropy and frequency fingerprint features labeled with user IDs, such that it tells which user a certain fingerprint vector belongs to?*

5.1 Data collection and experimental settings

To collect data for our evaluation, we deployed an Android app called EasyTrack based on the Funf Open Sensing Framework [1]. This app has a customizable configuration with 17 data types, including WiFi, Bluetooth, cell tower, GPS, call log, app usage etc. We successfully recruited 22 users to install EasyTrack and collected their mobile footprints for about 2 months with some variation across users.

We first test the initial user identification performance using different time frame lengths. We uniformly sample the instances to make sure that the same number of instances is used to build the classifier for each time frame length. As shown in Fig. 4a, when the length of time frame increases, the classification accuracy generally improves, despite possible variations caused by weekday/weekend patterns. This suggests that in this range, longer time frames better capture the uniqueness. Since the data collection time span is about two months, longer time frames decrease the total number of time periods and thus there are fewer features for training the classifiers. In the following experiments, we fix the time period at 10 days.

With a 10-day time frame, we reach 107 time frames in total from 22 users. On average, each user has 4.86 time frames. The range is [2, 8] and the standard deviation is 2.2. In total, we have 64 features in mFingerprint. We test combinations of features (multimodal entropies/frequencies, conditional entropies/frequencies) on multiple classifiers including Naive Bayes, decision tree, SVM and Multilayer Perceptron. We report the results from the Multilayer Perceptron, which were best. The learning rate is 0.3, momentum is 0.2, the number of hidden layers is set to $\frac{\#features+\#classes}{2}$ where $\#features$ is 64 in mFingerpint and $\#classes$ is 22, which is the number of users. The number of



(a) Accuracy for 22 users with different time frame lengths, with the basic entropy features.

(b) Classification accuracy with different # of users for basic frequency and entropy features.

(c) Average classification accuracy with different number of users for entropy feature combinations.

(d) Avg accuracy improvement for all features, compared to baseline frequency. Absolute avg accuracy is marked on the top of each bar.

Fig. 4: User identification classification results.

epochs is set to 500. We use 10-fold cross validation for training and testing. We use accuracy as a performance measurement which is defined as $\frac{\#of\ correct\ predictions}{\#of\ instances}$.

5.2 User Identification Performance

We tested on varying numbers of users from 4 to 22, and observe that as the number of users increases, the average accuracy drops but is still significantly better than random guessing (see Fig. 4b and Fig. 4c). In these experiments, we randomly sample 10 from the $\binom{22}{n}$ possible combinations (where n is the number of users being tested), apply 10 fold cross validation on each of the samples, calculate the accuracy and then get the average over the 10 samples. On average, each sample group has 63 instances.

Performance of standalone frequency and entropy features. For basic multimodal frequency and entropy features, each vector has 4 dimensions, corresponding to WiFi, cell tower, Bluetooth, and apps, respectively. We compare their classification results as shown in Fig. 4b. Both frequency and entropy features outperform the random baseline significantly. Entropies have better performance for large numbers of users compared to frequencies: mean accuracy with entropies drops 22 percentage points from 4 users to 22 users, versus a 35 point drop using basic frequencies.

Performance of conditional features. We also compared various types of conditional features (Fig. 4c). When features including basic multi-modal entropies, location entropies and time entropies are all combined, the performance is the best. Time features perform slightly better than location features, perhaps because there are only 2 location-conditioned features but 5 time-conditioned features. With all three kinds of features combined, the accuracy is 71.96% for 22 users versus 91.54% for 4 users and 86.59% for 10 users. Though more features improve accuracy, more computation is required as well, especially for the clustering required by location conditioning.

Performance of all features. Finally, we compute the performance with all mFingerprint features including basic frequencies, entropies, and conditioned features. Fig.

4d shows the performance improvement against the basic frequency feature results shown in Fig. 4b, with the absolute accuracy marked on the top of each bar: 94.68% for 4 users, 93.14% for 10 users and remains 81.30% for 22 users.

6 Conclusion

We presented mFingerprint and showed that its statistics (frequencies and entropies) computed from the device usage data and sensor data can be used as a fingerprint for user identification while preserving privacy. This serves as the key idea of the proposed mFingerprint framework to collect multimodal mobile data, compute footprint features, build unique user models, and serve personalized applications. We tested on a user-identification task and achieved over 81% accuracy even when the number of users reaches 22. The feature computation is designed considering both simplicity and energy-efficiency, and thus can naturally be fit into the on-device framework.

References

1. The Funf Open Sensing Framework. http://www.funf.org/.
2. J. Angulo and E. Wästlund. Exploring touch-screen biometrics for user identification on smart phones. In *Privacy and Identity Management for Life*, pages 130–143. Springer, 2012.
3. A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn. The rise of people-centric sensing. *IEEE Internet Computing*, 12(4):12–21, 2008.
4. G. Chittaranjan, J. Blom, and D. Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *ISWC*, pages 29–36, 2011.
5. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
6. Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 48–55. 2013.
7. T. M. T. Do and D. Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Ubicomp*, pages 163–172, 2012.
8. J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011.
9. R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Moodscope: building a mood sensor from smartphone usage pattern. In *Mobisys*, 2013.
10. H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *MobiSys*. ACM, 2009.
11. K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Ubicomp*, pages 281–290, 2010.
12. C. Shin, J.-H. Hong, and A. K. Dey. Understanding and prediction of mobile application usage for smart phones. In *Ubicomp*, pages 173–182, 2012.
13. J. Weppner and P. Lukowicz. Collaborative crowd density estimation with mobile phones. In *SenSys*, 2011.
14. Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semitri: a framework for semantic annotation of heterogeneous trajectories. In *EDBT*, pages 259–270, 2011.
15. Z. Yan, J. Yang, and E. M. Tapia. Smartphone bluetooth based social sensing. In *Ubicomp*, pages 95–98, 2013.