ShopProfiler: Profiling Shops with Crowdsourcing Data

Xiaonan Guo[1], Eddie C. L. Chan[2], Ce Liu[3], Kaishun Wu[2,5], Siyuan Liu[4] and Lionel M. Ni[2]
[1]*Singapore Management University*
[2]*The Hong Kong University of Science and Technology*
[3]*University of Pittsburgh*
[4]*Carnegie Mellon University*
[5]*Corresponding author*
Email: guoxn@ust.hk, csclchan@cse.ust.hk, cel38@pitt.edu, kwinson@ust.hk, siyuan@cmu.edu, ni@cse.ust.hk

***Abstract*—Sensing data from mobile phones provide us exciting and profitable applications. Recent research focuses on sensing indoor environment, but suffers from inaccuracy because of the limited reachability of human traces or requires human intervention to perform sophisticated tasks. In this paper, we present ShopProfiler, a shop profiling system on crowdsourcing data. First, we extract customer movement patterns from traces. Second, we improve accuracy of building floor plan by adopting a gradient-based approach and then localize shops through WiFi heat map. Third, we categorize shops by designing an SVM classifier in shop space to support multi-label classification. Finally, we infer brand name from SSID by applying string similarity measurement. Based on over five thousand traces in three big malls in two different countries, we conclude that ShopProfiler achieves better accuracy in building refined floor plan, and characterizes shops in terms of location, category and name with little human intervention.**

I. Introduction

Thanks to the widespread proliferation of mobile phones, and their capability of recording, processing and transmitting image, audio and location data, we are offered new opportunity to sense urban environment and human daily behaviors [11][20][21]. Data from these mobile phones impact the foundation of exciting applications and profitable commercial activities such as shopping, advertising and dining. Examples of applications include CSP [3], SurroundSense [2], CrowdInside [1], and Walkie-Markie [19].

Urban sensing is usually targeting sensors at profiling places. In the previous work, CrowdInside [1] system was proposed to automatically construct an indoor map based on smartphone. Walkie-Markie [19] constructed pathway by leveraging WiFi-landmark. CSP [3] categorized places through opportunistically collecting images and audio clips crowdsourced from smartphone users. However, all the previous work have several limitations when applied to real practice. First, traditional approaches of constructing floor plan employ sensor readings to track people and then detect room boundary by applying density-based algorithm on traces. Such methods may suffer from the divergence of real layout (counters, shelves or other obstacles). Second, if utilizing audio or image assistance in categorizing places, such approaches lack of scalability in terms of data sources, computational complexity and privacy constraints (e.g. taking photo is strictly prohibited in many commercial places).

978-1-4799-3360-0/14/$31.00 © 2014 IEEE

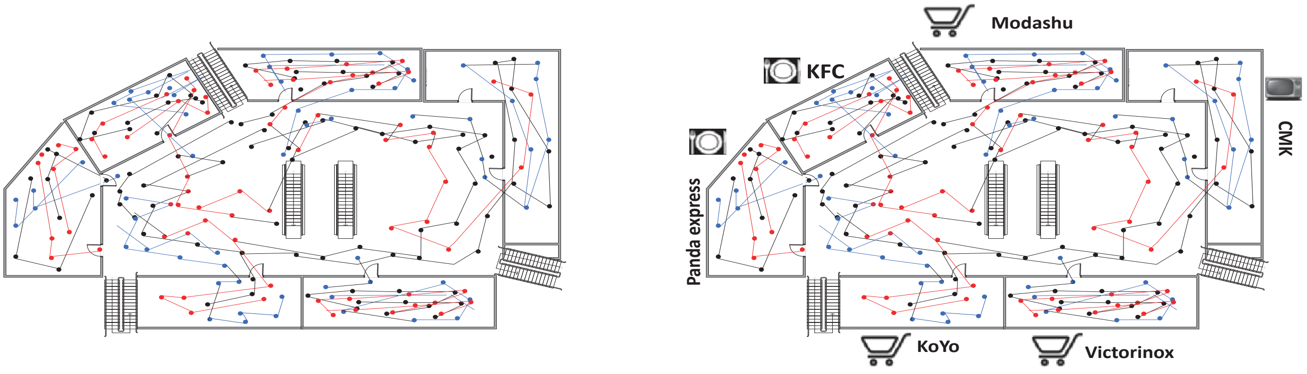
The goal of our work is to profile shops in mall by leveraging sensor readings from mobile devices as illustrated in Figure 1. Towards this goal, we empirically answer four questions.

First, how to characterize movement patterns from crowdsourcing data? In daily life, people follow certain movement patterns. For example, people may go to work in the morning, take lunch at noon and go home at night. Similarly, we explore customer movement patterns from sensor readings in mall. We collect sensing data from a variety of sources, such as accelerometer, compass, microphone and WiFi. Several interesting observations from empirical sensor readings can be used to characterize movement patterns: We are able to track people's moving speed from sensor readings. We notice that customers move at higher speed in corridor than in shop. Moreover, shops such as restaurant and cafe have higher population density at lunch and dinner time.

Second, how to differentiate basic units from movement patterns? Basic units in shopping malls are shops and corridors. In order to profile shops, we first differentiate basic units and then differentiate shop from each other by detecting shop boundary. For differentiating basic units, we extract useful features from customer movement pattern with location information and observe that movement pattern exhibits difference in different places. To differentiate shops, we leverage RSS (Received Signal Strength) information to design a boundary detection algorithm.

Third, how to categorize shops from multiple features? In a typical shopping mall, shops usually fall into some basic categories such as restaurant, book store, electronic store, supermarket, fashion related, etc. In shop type classification, we investigate the sensing data and extract several useful features for building classifier. Such as walking speed, population density at different times and the background sound intensity. Based on these features, we utilize Support Vector Machine (SVM) to categorize shops.

Finally, how to profile shops from SSID information? From our field study, we have an interesting observation that SSID (Service Set Identifier) can be considered as an indicator to infer shop names. We have investigated several typical shopping malls and noticed: First, APs are pervasive and nearly all shops in malls provide WiFi service either for customers or for their own employees. Second, SSID of these APs are more



(1) Floor plan

(2) Floor plan with more details

Fig. 1: Floor plan with real customers' traces and shops' details. (1) Traditional methods only build floor plan via customers' traces. (2) ShopProfiler system is able to profile shops in details (location, category, brand name, etc.)

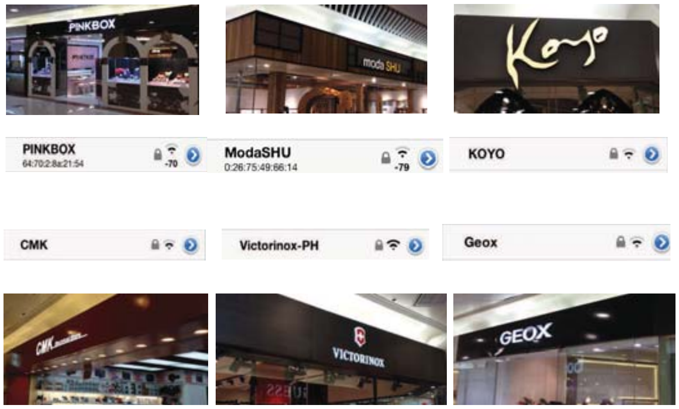


Fig. 2: Relationship between brand name and SSID. Even though the characters are not identical, we are able to infer shop names through SSID mining.

likely to be the characters of shop brand name. Third, APs belonging to a particular shop are more likely to reside in it. Our experimental results reveal that more than 80% shops set their SSID in this way. Examples are illustrated in Figure 2.

In this paper, we propose ShopProfiler as an automatically profiling system. ShopProfiler only relies on sensor readings from mobile devices. The process of data collection is automatic and running in the background. Inertial sensor readings reflect human movement information such as acceleration, heading and speed. Microphone and WiFi modules provide additional information of surrounding conditions. From the customer's point of view, mobile sensing data in a mall contain information of what shops that customers visit, how long they stay, and how fast they are walking. From shop's point of view, mobile sensing data reveal information about how many people visit the shops in a particular time period and what is the layout inside shops. Through mining and learning mobile sensing data, we capture unique features of different shops and categorize them into types (e.g., restaurant, book store, supermarket, electronic store, fashion and etc.) Furthermore, a complete profile of shops should contain specific location and precise brand name as well. Previous approaches usually leverage density-based algorithm in shop boundary detection to distinct location, and utilize image or audio to infer shop brand names. Our approach improves the accuracy of boundary detection by adopting gradient-based method and explore the relationship between SSID and brand names to characterize shops.

In summary, our contributions are as follows:

- We design and implement ShopProfiler as a system to automatically profile shops in mall with shop type and brand name. Our approach is able to achieve over 80% accuracy with little human intervention.
- We propose a gradient-based approach for shop boundary detection. ShopProfiler detects the shop boundary where the concrete wall is in-between and causes significant attenuation of signal.
- ShopProfiler system is able to categorize shops. By investigating human movement patterns in three big shopping malls from two different countries, we classify shops into different types with SVM.
- We are able to pinpoint shop location by constructing WiFi heat map via crowdsourcing data and discover shop name by extracting features from SSID through string similarity measurement.

The rest of this paper is organized as follows. In Section II, we present an overview of our ShopProfiler system. Section III discovers people movement patterns from mobile phone sensor readings. In Section IV, we will propose several novel approaches to distinguish different building units and classify shops into different types. In Section VI, we will present how to pinpoint shop location and discover brand name. We evaluate our system in Section VII. In Section VIII, we review the state of the art technologies related to this topic. Finally, we conclude our work and provide directions for future work.

II. ShopProfiler Overview and Design

ShopProfiler system is based on information from mobile phone sensors and automatically profile shop via crowdsouring. One hand, mobile phone sensors provide various information that depict human movement patterns in a typical shopping mall while sensing the environment around people. On the other hand, a large number of people's traces contain adequate

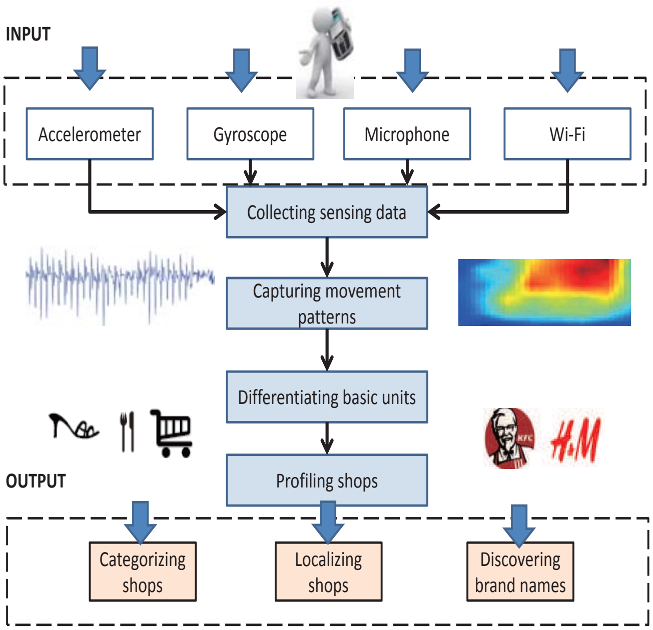


Fig. 3: ShopProfiler system.

descriptions of a building layout and shop information. As shown in Figure 3, ShopProfiler system has four modules.

1) **Collecting sensor data**: Sensor data are collected from inertial sensors, microphones and WiFi module while people walking around in shopping malls. Inertial sensors including accelerometer and gyroscope provide sufficient information for tracking customers. In addition, acoustic information reflects environmental features. Moreover, WiFi module captures data from all available APs including RSS, SSID and BSSID.
2) **Capturing movement pattern**: Customer traces in mall reveal features of what kind of shops the customers visit, how long they stay, and how fast they are walking. On the other hand, from shop's point of views, traces show that how many people visit the shops in a particular time period and what is the layout inside shops.
3) **Differentiating basic units**: We first differentiate shop and corridor, which are the two element units in typical malls. ShopProfiler extracts unique features from movement patterns to differentiate these basic units. Then, we propose a novel gradient-based room boundary detection algorithm that reduces false positive rate by traditionally applying density-based algorithm.
4) **Categorizing shops and discovering brand names**: ShopProfiler adopts SVM to categorize shops (i.e., restaurant, electronic store, book store, etc). This module is based on the data collected in module 1 and analysis of movement patterns in module 2. Furthermore, ShopProfiler performs AP localization algorithm via crowdsourcing and matches AP to floor outline. Finally, we label shop brand name using SSID mining.

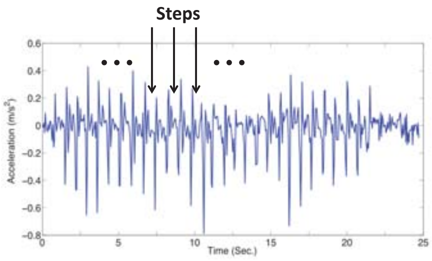


Fig. 4: We can count steps from acceleration data. One step is defined by two wave peaks.

In this paper we do not include details of how to track people because mobile based indoor localization has already been extensively studied [6][5][10].

III. Movement Pattern in Shopping Mall

To collect mobile sensing data, we develop an Android App to record inertial sensor readings, and background sound intensity. Moreover, mobile phone continuously scans all available nearby WiFi and records RSS, SSID and BSSID. To investigate movement pattern, we conducted three experiments in three big shopping malls in two different countries. In one experiment, we collected over 5,000 human traces in a shopping mall for two days. In the other two experiments, we had 20 volunteers in each shopping mall and collected data for one whole week. Several interesting observations from the field study and real life data are reported as follows.

First, walking speed is able to indicate location features. Refer to our system design, the data from inertial sensor describe people movement pattern from several aspects. From the experimental results, we notice that when huge amount of data traces are available, walking speed can indicate a shop or a corridor. Theoretically, the walking speed can be calculated by integrating acceleration with respect to time. However due to noise in poorly calibrated sensing data, calculating speed from inertial sensor readings leads to significant accumulation of errors [23]. As a result, we measure walking speed as the number of steps per minute. We count steps from acceleration patterns as shown in Figure 4.

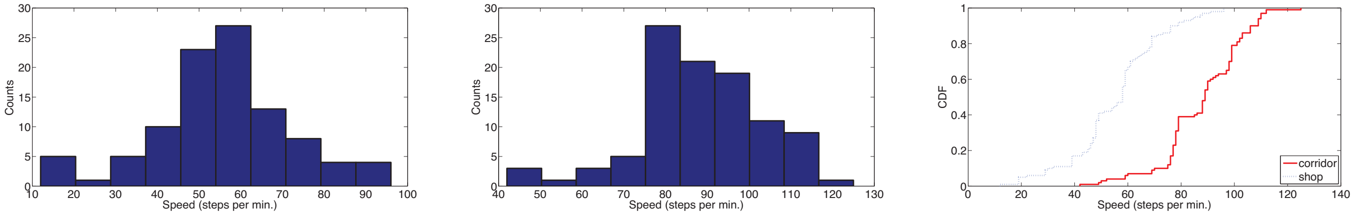
Second, customers frequently change walking direction and the time spent in different shops. For example, in a typical restaurant customers are likely to spend more time sitting at table for lunch/dinner and the heading change frequency is low. On the other hand, people in other shops have high frequency.

Third, population distribution reflects patterns at different time from a shop's point of view. Our observation meets the intuition that population distribution reveals a pattern type in a restaurant since the majority of people have lunch at noon and dinner at night. In Figure 5, we show the table occupancy of a restaurant in a shopping mall on Monday, Wednesday and Saturday respectively. Population density is much higher at noon/night than other times. Moreover, supermarkets appear another movement pattern where most of customers leave shopping malls or go home after shopping.

We summarize preliminary observations from our empirical studies in Table I and list important notions in Table II.

TABLE I: Observations (Customers exhibit different movement patterns in different shops.)

Features	Restaurant	Electronic store	Book store	Supermarket
Walking speed	stable	medium	low	fast
Spend time	stable	medium	low	fast
Sound loudness	medium	high	low	medium
Frequency of heading change	low	fast	low	fast
population density distribution	regular pattern	unknown	unknown	regular pattern



(1) Walking speed measurement in shop (2) Walking speed measurement in corridor (3) CDF of walking speed in two areas

Fig. 6: Difference of moving speed between shop and corridor. Moving speed in corridor is faster than in shop. The average moving speed in corridor is around 90 steps per minute and 60 steps per minute in shop.

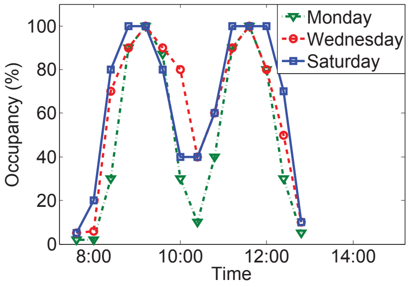


Fig. 5: Occupancy change (from 9:00 to 22:00) of a restaurant on Monday, Wednesday and Saturday.

TABLE II: Notions used in this paper

Notion	Description
(x_i, y_j)	Location
g_i	RSS gradient in adjacent area of location i
T	Gradient threshold
$\rho_{(x_i,y_j)}$	RSS measurement at location (x_i, y_j)
ϱ_t	RSS measurement at time t
ω	walking speed (steps per minute)
ψ	population density distribution indicator
ε	shop instance
C_j	shop types of instance ε
I	attribute
a_i	ith attribute
θ	moving direction
ζ	environment sound intensity
Ω	plane of a floor plan
N	number of People

IV. Categorizing Shops

In this section, we present our approach to categorize shops into different types. Before introducing the technical details, we first describe how to differentiate shop and corridor by leveraging speed variance. Then we design a novel gradient-based approach to detect room boundaries. Finally, we introduce classifier design for categorizing shops.

A. Differentiate Shop and Corridor

Corridors and shops are the key elements of shopping malls. As discussed in previous sections, we leverage walk speed to differentiate basic units. Note that some customers may walk at a slow speed in corridors while other may walk faster in shops. However, these instances do not affect our classification results as the adequate traces via crowdsoucring provide accurate prediction.

Assume we divide traces into segments by turn. Without loss of granularity, suppose there are n segments denoted by $s_1, s_2, s_3, \cdots, s_n$ and for each segment we have a score Γ. Every customer trace containing a particular segment contributes a score γ_{nm} which denotes that customer m contributes γ to segment n. Therefore, the score of each segment is

$$\Gamma_n = \sum_m \gamma_{nm} \tag{1}$$

Generally, when a customer's moving speed exceeds a threshold, then $\gamma_{nm} = 1$, while $\gamma_{nm} = -1$ when speed is under the threshold. Consequently, we label the segment s_n belongs to corridor as long as $\Gamma_n > 0$ and segments that belong to shop will have $\Gamma_n < 0$. The threshold is obtained by our empirical experiments, that is, 90 steps per minute. In Figure 6, we report the difference in moving speed between shop and corridor. When people walk in a corridor, the moving speed is around 90 steps per minute. On the contrary, in shops the moving speed is down to 60 steps per minute as customer usually stop to browse products.

B. Shop Boundary Detection

After basic unit retrieval, we separate shops for further classification. Previously in [1], CrowdInside adopts density-based clustering algorithm for room boundary detection. Nevertheless, this approach may result in serious false positive (FP) in shopping mall scenario. In a typical shop, there are counters for customers to pick up goods, and furniture or shelves for display products as well. These entities are

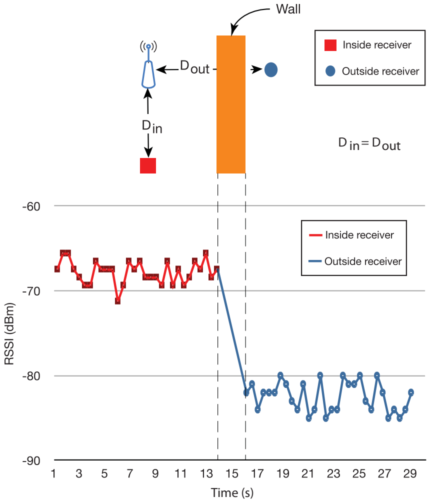


Fig. 7: Wall effect. Wall between the transmitter and the outsider receiver cause the signal attenuation. The plot shows RSS difference.

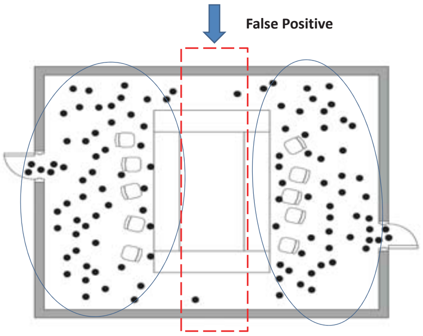


Fig. 8: Black dots represent samples of traces and the red rectangle shows the boundary returned by density-based algorithm. In this example, the customer reachability is limited by the counter. Using density-based algorithm in boundary detection may lead to serious false positive.

pervasive in shops and people's moving paths are restricted by these obstacles. Figure 8 illustrates the result of applying density-based algorithm to shopping mall scenario. Suppose a shop contains several counters in the center and customers cannot walk through or enter the counter. The dots represent customers' positions. We find that the algorithm classifies the traces into two categories and wrongly regard the counter as the room boundary.

To reduce FP rate, ShopProfiler adopts a novel approach in detecting room boundary. The method is motivated by the observation that the WiFi signal suffers from significant attenuation when passing through walls. Our empirical results support this observation as shown in Figure 7. In this experiment, we have two receivers r_{in} and r_{out}. The distances between receivers and AP are D_{in} and D_{out} respectively. There exists a Line-of-sight (LOS) path between r_{in} and the AP. While the LOS path between r_{out} and the AP is blocked by a concrete wall. We collect RSS from two receivers and find that without a concrete wall, the RSS at r_{in} is around $-66dBm$ and down to about $-80dBm$ at r_{out}. Generally, the gradient of a scalar field is a vector field that points in the direction of the greatest rate of increase of the scalar field. In ShopProfiler, the area with the fastest decrease indicates the boundary that separates two different shops.

Algorithm 1: Detecting room boundary.

```
input  : adjacent cluster pair $(i_l, i_r)$
output: room boundary indicator vector $\Lambda$
for all $(i_l, i_r)$ do
    for all (m, n) do
        for all $\theta$ do
            calculate $g_\theta$;
        end
        $g_n \leftarrow max(g_\theta)$;
    end
    if $g_n > threshold$ then
        $\Lambda \leftarrow 1$
    end
end
```

Generally, the gradient of a scalar function $f(x_1, x_2, x_3, ..., x_n)$ is denoted by ∇f, where ∇ denotes the vector differential operator. In a three-dimensional space, ∇f is given by

$$\nabla f = \frac{\partial f}{\partial x} i + \frac{\partial f}{\partial y} j + \frac{\partial f}{\partial z} k \tag{2}$$

where i, j, k are standard unit vectors.

The algorithm input is the result returned by density-based algorithm to obtain a candidate region for shop boundaries because density-based algorithm classifies trace data into different clusters. The candidate room boundaries are regions between two adjacent clusters with distance smaller than the general obstacle width. We assume that for each candidate region there is an indicator $\Lambda \in \{1, 0\}$ telling whether the region is a boundary or not and the width of obstacle is smaller than 5m. The distance is defined as $min(dis(m, n))$ where m, n are points from two clusters respectively. Suppose there are i clusters returned from density-based algorithm. First, we pick two points m and n subject to the distance constraint and generate a radial line from m and n then we generate 11 other radial lines every 30 degrees. Second, we calculate the gradient along each direction. We assume n is located at $(x_n, y_n)i \in \Re$ and record an RSS value ϱ_t at time t and the other point m record RSS value ϱ_i at time i. Then RSS gradient g_n considering RSS in adjacent area within distance σ_2 is

$$g_n = \sum_{\|(x_n,y_n)-(x_m,y_m)\|<\sigma_2} (\varrho_t - \rho_i) \frac{(x_n, y_n) - (x_m, y_m)}{\|(x_n, y_n) - (x_m, y_m)\|^2} \tag{3}$$

where ρ_i is RSS measurement of visited location. Finally, if the gradient of a direction surpasses a threshold and then we notice there exists a wall orthogonal to the radial line. According to Eq. (3), threshold T is set to 6.3 because of the fact that for concrete wall, signal at 2.4GHz represents about 12dB attenuation[16]. The whole process is described in Algorithm 1.

Algorithm 2: Categorizing shops.

```
input  : (training set, type, testing set)
output: types
        ▷ Suppose there are $j$ instances in training set
for $k = 1, 2, 3, 4, 5$ do
    for all $j$ do
        if $(C_j = k)$ then
            $C_j \leftarrow 1$;
            else $C_j \leftarrow 0$;
        end
    end
    Building $model_k$ based on 2-class SVM ;
end
        ▷ Suppose there are $i$ instances in testing set
for all $i$ do
    for $k = 1, 2, 3, 4, 5$ do
        if (applying $model_k$ on i)==TURE then
            break
        end
    end
    return $C_j$;
end
```

C. Classifier Design

After separating shops from each other, in this part, we introduce the approach to categorize shops. Our empirical experiments have shown that a restaurant is more likely have a dense population at lunch time and dinner time. Moreover, people are likely to spend more time in a restaurant than in other shops, and the moving speed is relatively stable. An electricity store has some unique acoustic fingerprints because of its high sound intensity due to display appliances, such as TVs, radios, fans and etc. A book store is much quieter than other places and most of the people are likely to walk at slow speed or be stable for a while. Based on the above observations from our dataset, we are trying to extract features to separate shops from each other. In ShopProfiler, we define six different classes ({Restaurants}, {electronics stores}, {book stores}, {supermarkets}, {fashion related}, {others}). Therefore, for each instance with observed attributes, we classify an instance into one of them.

We utilize support vector machine (SVM) to categorize shops. SVM is widely used for classification and regressions. The input of SVM model is obtained directly or extracted from mobile phone sensor readings, including velocity (steps per minute), change of direction ($\frac{d\theta}{dt}$), sound intensity (ζ),WiFi RSSI (ρ), timestamp (t), and people count (N).

The output of the SVM model are the types of shops. Since we have six different classes, we adopt a multi-class SVM technique. In our case, multi-class SVM aims to assign shop types to instances by using SVM, where the number of types is six. In ShopProfiler, we reduce the multi-class problem to multiple binary classification problem. We describe the procedure in Algorithm 2: First, we build k binary classifiers which distinguish between one certain type and the rest. Second, for each instance from testing set, we use 2-class SVM to perform the classification.

Algorithm 3: Retrieving brand name from SSID.

```
input  : SSID $w_i$, Brand database $D$
output: Brand name $B_i$
for all $w_i$ do
    $K \leftarrow \phi$
    for all $w_j$ in $D$ do
        calculate $lev(w_i, w_j)$;
        if $lev(w_i, w_j) < threshold$ then
            $K_i \leftarrow lev(w_i, w_j)$
        end
    end
    if $K \neq \phi$ then
        $B_i \leftarrow Min(K_i)$
    end
    else
        $B_i \leftarrow w_i$
    end
end
Output Brand name $B_i$
```

V. Retrieving Brand Name from SSID

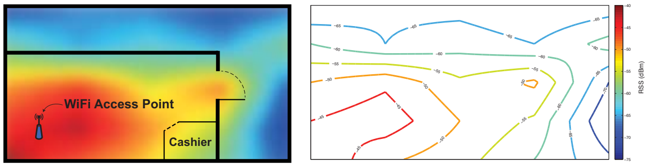
Our empirical studies show that SSID is a good indicator to retrieve brand name. First, in a shopping mall, most of shops have wireless connection either for customers or for their own employees. Second, shops are more likely to use characters that contain brand name as SSID because it helps customers to identify WiFi and the shop. Third, the AP of a particular shop is more likely located in that shop. Based on these observations, we extract band name features from SSID by comparing the string in SSID with brand names in database. ShopProfiler system adopts a string similarity search algorithm with edit distance (a.k.a, Levenshtein distance) threshold. Generally, the edit distance between two strings w_1 and w_2 is the minimum number of edit operations of single character needed to transform w_1 to w_2. Edit operations include insertion, deletion, and substitution. We denote the edit distance between the two strings as lev(w_1, w_2).

Algorithm 3 elaborates our threshold-based approach. For each SSID w_i, we scan the brand name in database D and calculate the edit distance accordingly. If the edit distance is smaller than the threshold, we put w_j in candidate array K. After traversing all the brand names in database, we select w_j with the smallest $lev(w_i, w_j)$ value and set it as the band name. If all the $lev(w_i, w_j)$ are larger than the threshold, we output w_i as the brand name. The computational complexity is $O(n)$, where n is the number of the entries in brand name database.

VI. Pinpointing Shop Location

RSS can be used for AP location inference, since signal will attenuate when the distance between AP and sampling position increase and suffer from significant attenuation when there is a wall in between. Moreover, through crowdsourcing technique, we can build WiFi heat map accordingly and pinpoint shop location even in a rich multi-path environment.

We conducted an experiment in a shop where we know the AP location by asking the shop owner. We divided the test



(a) WiFi heat map (b) Isothermal curve

Fig. 9: WiFi heat map with isothermal curve. We locate AP in room level by matching the hot spots with shops.

area into cells and for each cell we measure RSS from the AP. Based on site survey, we built a WiFi heat map as shown in Figure 9 (a). Such a WiFi heat map is constructed by site survey. In practice, with the help of crowdsourcing, we are able to obtain a huge amount of traces. Therefore, the site survey procedure is conducted by implicit crowdsourcing. Customers walk around with recording RSS and then we build a WiFi heat map accordingly.

The experimental result shows that wall affects WiFi signal as well. We survey the site by measuring RSS and build heat map according to SSID, and then consider the *hot spot* as the location of the AP. Hot spot is defined as a region within a certain *isothermal curve* as shown in Figure 9 (b). Assume the ith customer measures RSS ρ_i at point (x_i, y_i) in a floor plan. For each shop ε we have a vector that contains RSS measurement $\rho_{(}x_i, y_j)$ in the customer traces. When a new customer enters the shop, the corresponding RSS is added to the vector. Therefore, we build a heat map at time t_c with all available RSS measurements $\{\rho_t | t \leq t_c\}$. ShopProfiler then draws an isothermal curve and locates the hot spot in a particular shop.

VII. Empirical Experiment Result and Discussion

A. Experiment Setup

We conduct extensive experiments on three real life datasets from three different experiment scenarios. **Dataset A**: Customer shopping behavior data [12]. We collected 5,000 customers' traces for two days in one level of a big shopping mall A in Singapore. **Dataset B**: Volunteer shopping behavior data as one reference experiment. We recruited 20 volunteers and record 250 traces in a shopping mall B in Hong Kong. **Dataset C**: Volunteer shopping behavior data as the other reference experiment. We recruited 20 volunteers and record 250 traces in another shopping mall C in Hong Kong. For shopping mall B, most of the shops are large. The other, shopping mall C, is smaller and each shop is smaller. The features of the three shopping malls are reported in Table III.

Dataset B and **Dataset C** have close-loop control setup while there is no such setup in **Dataset A**. For **Dataset B** and **Dataset C**, first, each of the volunteer held a mobile devices installed with our App for sensor reading collection and a power bank to provide extra energy in case of running down of battery. The volunteers were required to stay in malls for shopping, eating and recreation. In this set of experiments, three mobile device models (Nexus S, HTC G10 and Galaxy Tab) were used. This configuration for **Dataset B** and **Dataset C** help us analyze the impact of different parameters.

TABLE III: Statistics of three shopping malls (A & B & C).

Categories	A	B	C
# shops	15	13	10
# restaurants	5	3	3
# electronic stores	1	1	2
# book stores	1	Null	1
# supermarkets	1	Null	1
# other stores	7	9	3
# different BSSID	20	17	14

TABLE IV: Prediction accuracy on different kernels.

Kernel function	Prediction accuracy (SVM)
Radial Basis Function	46.15%
Polynomial	76.92%
Multilayer perceptrons	70.23%
Quadratic	81.58%

B. Impact of Different Kernel Functions

In this part, we show the impact of different kernel functions on **Dataset A**. We show the comparison results among Radial Basis Function kernel, Polynomial kernel, Multilayer perceptron kernel and Quadratic kernel. The prediction accuracy of different kernels are shown in Table IV. Among four kernel functions, Quadratic kernel presented the highest prediction accuracy based on the training data. Since the issue of how to choose the kernel function is out of the scope of this paper, ShopProfiler simply chooses Quadratic kernel function in categorizing shops.

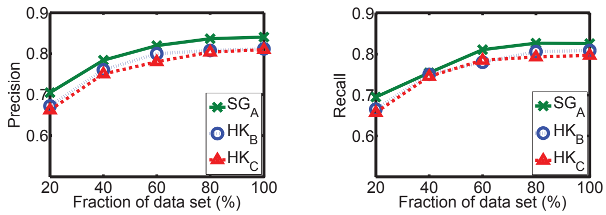
C. Prediction in Different Scenarios

We compare three prediction results on traces from three different experiment scenarios. As shown in Figure 10, ShopProfiler is tolerant to different trace sizes. The results show that 1) the more traces, the better accuracy (A is better than B and C) and 2) the larger room unit size, the better accuracy (B is better than C). Note that overall the prediction accuracy begins to converge with more traces.

D. System Performance

We evaluate ShopProfiler performance in two malls (B & C). First, we compare gradient-based algorithm with density-based algorithm on shop boundaries detection. Figure 11 shows the experimental results. The y-axis represents the number of shops given by the output of the two algorithms. The ground truth of the number of shops for each shop equals one. Suppose within a shop the output of density-based algorithm is two clusters, which means density-based algorithm regards the two clusters as belonging to two different shops. In shopping mall B, the proposed gradient-based algorithm outperforms density-based algorithm. There are only one FP occur. On the contrary, FP of density-based algorithm is three. In shopping mall C, the FP of the gradient-based algorithm and density-based algorithm are the same.

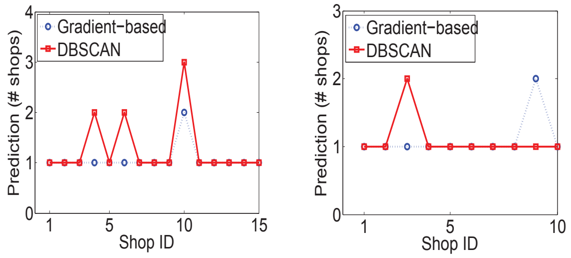
We report prediction result on shop brand name and category in Table V. There are 15 different shops and 13 of the shops contain characters of brand name in SSID. The prediction accuracy of category is over 80%.



(a) Precision in mall test

(b) Recall in mall test

Fig. 10: Effectiveness in mall test. The result from the big trace set (A) outperforms the small trace set (B and C).



(a) Mall B

(b) Mall C

Fig. 11: Prediction on boundary. (1) Result of a large shop in mall B. The gradient-based algorithm outperforms the density-based algorithm, because a large size shop usually has counter or shelf between customer traces. (2) Result of boundary detection in mall C. The two algorithms have the same FP.

E. Findings and Discussion

We evaluate ShopProfiler on multiple datasets from two different countries to investigate the impact of 1) different sizes of traces, 2) different scales of room units and 3) different behavior patterns of different persons. The characteristics of the first dataset can be described as follows. 1) **Scale.** 5,000 real customer traces in a mall from Singapore. 2) **Diversity.** Customers involve young and old, men and women. No restrict in using of mobile devices. But for the second and the third trace data collected by volunteers in two malls from Hong Kong, there are only 250 traces each. The volunteers are men and women (age from 24 to 29). We only provide three different types of mobile devices. The localization accuracy is high because we record the true positions.

TABLE V: Shop brand name, SSID and category.

Shop brand name	SSID	category (prediction)	category (ground truth)
modashu	ModaSHU	fashion related	cloth and bags
PINKBOX	PINKBOX	fashion related	cloth and bags
VICTORRINOX	Victorinox-PH	book store	Knife and bags
GEOX	Geox	others	cloth and bags
KOYO	KOYO	restaurant	restaurant
Bauhaus	BAUHAUSSHOP	electronic store	cloth and bags
Coxell	Coxell02	electronic store	electronic store
CMK	CMK	restaurant	restaurant
SUZURANBED	SUZURAN BED	restaurant	bedclothes
Catalo	Cisco03870	restaurant	restaurant
MacDonald	MacDonald	restaurant	restaurant
Marks & Spencer	MS	supermarket	supermarket
Mimosa	mimosaAP	book store	book store
Charles & Keith	CharlesKeith	fashion related	cloth & bags
STARBUCKS	Starbucks-free	restaurant	restaurant

Based on experimental evaluation, ShopProfiler achieved high prediction accuracy despite of different experiment scenarios. The large trace set (A) outperforms the ones (B and C) from volunteers. Consequently, we illustrate crowdsourcing technique is robust and localization accuracy has no significant impact on system. Furthermore, the accuracy of shop type classification and shop brand discovery could be improved by human-aid. We can return the classification result to customers and they can investigate the accuracy during site visiting. Customers can help correct the possible wrong types by voting a shop into a different classification depending on site survey.

VIII. Related Work

Constructing floor plan The floor plan construction falls into the simultaneous localization and mapping (SLAM) system, which has been extensively studied in robotic fields. SLAM relies on accurate control of a robot [13], [22], which is equipped with sensors, ultrasonic, or cameras to build maps. In [18] FootSLAM used shoe-mounted inertial sensors to construct the map. In [17] PlaceSLAM derived a Bayesian formulation and particle filtering implementation for manually annotated. Recently, researchers begin to use commodity mobile phones to construct a floor plan or path way. CrowdInside [1] used sensor readings from mobile phones to track people by leveraging common environment anchor points to reset the errors obtained by dead-reckoning. And then CrowdInside applied density-based algorithm to users' traces to perform boundary detection. In [19], instead of looking at the face RSS values, the authors leveraged the trend of RSS changes by defining WiFi-marks and used it on building path way. Our approach is different from the previous methods in room boundary detection. In stead of using density-based algorithm, ShopProfiler leveraging RSS information to detect boundary with high accuracy, since previous density-based algorithm regards the area that traces cannot be reached as wall, such as shelves, counters and other obstacles.

Characterizing places Techniques for characterizing places have been proposed from different perspectives. [8] proposed opportunistic feature vector merging, and the social network-driven sharing of training data and models between users. [14] aimed to use information on the sequence of visiting each product zone to find how they affected purchasing. But our approach is from another perspective that we use mobile sensing data to profile shops. In [7][25], data from personally carried devices were augmented with users either providing or confirming location semantics. Techniques were proposed in [24] to leverage FourSquare check-in data to determine place categories. But the previous approaches more or less need human to participate in. Our approach is different from the previous techniques by automatically characterizing shops with little human intervention. Furthermore, on learning the name of location, previous work mainly infer from traces data and GPS information. [9] proposed the relational Markov network to label locations with the activities that have occurred in these places. [15] proposed a Hierarchical Hidden Markov Model on raw GPS data to extract significant locations and mode transfer locations and then to learn users' transportation routines. [3] proposed an approach that combines location and traces with sensor hints through opportunistically captured images and audio clips from smartphones. In indoor scenarios, however, concrete wall prevents GPS signal being received and taking photo is prohibited in some place especially in malls. Our approach characterizes shops from mobile sensing data and infers shop brand names from SSID information.

In summary, ShopProfiler is unique in leveraging customer traces and mobile sensing data to categorize shops and infer brand names from SSID. Moreover, it allows for automatically detecting boundary in malls.

IX. Conclusion and Future Work

In this paper, we presented ShopProfiler as a automatic profiling system. ShopProfiler adopts novel approaches to depict details inside shopping malls, including shop type and brand name. In addition, our proposed methods are able to improve the accuracy of building floor plan by significantly reducing FP rate compared with previous methods. The whole process is autonomous without human intervention and only relies on sensor readings from mobile phones via crowdsourcing. First, we extract customer movement patterns from real traces. Second, we adopt a gradient-based approach to detect shop boundary and pinpoint shop location through WiFi heat map construction. Third, we categorize shops by design a SVM classifier. Finally, we find that SSID is a good indicator to infer shop names. Therefore, we applying string similarity measurement to label shops. Through real traces from three malls, we confirmed the merits of our system.

Our system can be improved in the following aspects. First, currently there are six classes in our ShopProfiler. More classes can be defined by exploring people's behavior in depth. Second, power consumption is another importance issue in mobile phone based applications[4]. Battery drains very fast when continuously using build-in sensors and WiFi modules.

Acknowledgement

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA); Hong Kong RGC Grant HKUST617212, HKUST617710 and HKUST617811; Major State Basic Research Development Program (973 Program) under Grant 2013CB329006; New Century Excellent Talents in University (NCET-13-0908); Guangdong Natural Science Funds for Distinguished Young Scholar (No.S20120011468); New Star of Pearl River on Science and Technology of Guangzhou (No.2012J2200081); Guangdong NSF Grant (No.S2012010010427), China NSFC Grant 61202454, 60933011, 61170077, 61170076, 61033009, 61003272, 61027009 and 61103001; Science and Technology Planning Project of Guangdong Province, China under Grant No. 2011A011302001.

References

[1] M. Alzantot and M. Youssef. Crowdinside: automatic construction of indoor floorplans. In *Proc. of ACM SIGSPATIAL*, 2012.

[2] M. Azizyan, I. Constandache, and R. Roy Choudhury. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proc. of ACM Mobicom*, 2009.

[3] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proc. of ACM UbiComp*, 2012.

[4] I. Constandache, S. Gaonkar, M. Sayler, R. Choudhury, and L. Cox. Enloc: Energy-efficient localization for mobile phones. In *Proc. of INFOCOM*, 2009.

[5] X. Guo, D. Zhang, and L. Ni. Localizing multiple objects in an rf-based dynamic environment. In *Proc. of IEEE ICDCS*, 2012.

[6] T. W. Hnat, E. Griffiths, R. Dawson, and K. Whitehouse. Doorjamb: unobtrusive room-level tracking of people in homes using doorway sensors. In *Proc. of Sensys*, 2012.

[7] D. H. Kim, K. Han, and D. Estrin. Employing user feedback for semantic location services. In *Proc. of ACM UbiComp*, 2011.

[8] N. D. Lane, H. Lu, S. B. Eisenman, and A. T. Campbell. Cooperative techniques supporting sensor-based people-centric inferencing. In *Proc. of IEEE Percom*, 2008.

[9] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational markov networks. In *Proc. of ACM IJCAI*, 2005.

[10] H. Lim, L.-C. Kung, J. C. Hou, and H. Luo. Zero-configuration indoor localization over ieee 802.11 wireless infrastructure. *Wireless Network*, 2010.

[11] S. Liu, L. Kang, L. Chen, and L. Ni. Distributed incomplete pattern matching via a novel weighted bloom filter. In *Proc. of IEEE ICDCS*, 2012.

[12] S. Liu, S. Wang, K. Jayarajah, A. Misra, and R. Krishnan. Todmis: Mining communities from trajectories. In *Proc. of ACM CIKM*, 2013.

[13] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. Fastslam: a factored solution to the simultaneous localization and mapping problem. In *Proc. of ACM AAAI*, 2002.

[14] T. Nakahara and K. Yada. Analyzing consumers' shopping behavior using rfid data and pattern mining. *Adv. Data Analysis and Classification*, pages 355–365, 2012.

[15] D. Patterson, L. Liao, D. Fox, and H. Kautz. Inferring high-level behavior from low-level sensors. In *Proc. of ACM UbiComp*, 2003.

[16] D. M. Pozar. *Microwave Engineering*. Addison-Wesley, 2004.

[17] P. Robertson, M. Angermann, and M. Khider. Improving simultaneous localization and mapping for pedestrian navigation and automatic mapping of buildings by using online human-based feature labeling. In *Proc. of IEEE PLANS 2010*, 2010.

[18] P. Robertson, M. Angermann, and B. Krach. Simultaneous localization and mapping for pedestrians using only foot-mounted inertial sensors. In *Proc. of ACM UbiComp*, 2009.

[19] G. Shen, C. Zhuo, Z. Peichao, M. Thomas, and Z. Yongguang. Walkie-markie: Indoor pathway mapping made easy. In *Proc. of NSDI*, 2013.

[20] Y. Tong, L. Chen, Y. Cheng, and P. S. Yu. Mining frequent itemsets over uncertain databases. *PVLDB*, 5(11):1650–1661, 2012.

[21] Y. Tong, L. Chen, and B. Ding. Discovering threshold-based frequent closed itemsets over probabilistic data. In *Proc. of IEEE ICDE*, 2012.

[22] R. W. Wenzel. Giving a compass to a robot - probabilistic techniques for simultaneous localisation and map building (slam) in mobile robotics. Technical report, 2002.

[23] O. Woodman and R. Harle. Pedestrian localisation for indoor environments. In *Proc. of ACM Ubicom*, 2008.

[24] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *Proc. of ACM SIGKDD*, 2011.

[25] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst.*, 25(3), July 2007.