Automatically Characterizing Places with Opportunistic CrowdSensing using Smartphones

Yohan Chon†, Nicholas D. Lane‡, Fan Li‡, Hojung Cha†, Feng Zhao‡

†Yonsei University ‡Microsoft Research Asia
Seoul, Korea Beijing, China

ABSTRACT

Automated and scalable approaches for understanding the semantics of places are critical to improving both existing and emerging mobile services. In this paper, we present *CrowdSense@Place* (CSP), a framework that exploits a previously untapped resource – opportunistically captured images and audio clips from smartphones – to link place visits with place categories (e.g., store, restaurant). CSP combines signals based on location and user trajectories (using WiFi/GPS) along with various visual and audio place "hints" mined from opportunistic sensor data. Place hints include words spoken by people, text written on signs or objects recognized in the environment. We evaluate CSP with a seven-week, 36-user experiment involving 1,241 places in five locations around the world. Our results show that CSP can classify places into a variety of categories with an overall accuracy of 69%, outperforming currently available alternative solutions.

Author Keywords

Semantic Location, Crowdsourcing, Smartphone Sensing, Location-Based Services

ACM Classification Keywords

I.2.6 Artificial Intelligence: Learning; J.4 Computer Applications: Social and Behavior Sciences.

General Terms

Algorithms, Design, Experimentation, Human Factors

INTRODUCTION

Smartphones embedded with a growing diversity of new sensors continue to capture media headlines and the attention of both consumers and researchers alike. However, location remains the most successful and widely used contextual signal in everyday usage. Awareness of user location underpins many popular and emerging mobile applications, including local search, point-of-interest recommendation services, navigation, and geo-tagging for photographs and tweets. Still, just like most forms of low-level sensor data, many of the potential uses of location require that we extract high-level pieces of information. A key abstraction when interpreting location sensor data is *place* – that is, logical locations meaningful to users, such as where they work, live, exercise, or shop. Prior work has shown how places can be discovered from temporal streams of user location coordinates [5, 19, 16, 13]. However, if we can automatically *characterize* places by linking them with attributes, such as place categories (e.g., "clothing store," "gym") or likely associated user activities (e.g., "eating," "work"), we can realize powerful location- and context-based scenarios. For example, mobile applications such as location-based reminders [29] or content delivery [22] can become aware of place semantics. Beyond potential mobile applications, scalable techniques for characterizing the places people visit can act as a valuable signal for activity recognition, allowing greater understanding of large-scale human behavioral patterns.

In this paper, we propose *CrowdSense@Place* (CSP) a framework for categorizing places that relies on a previously unused source of sensor data – opportunistically collected images and audio clips crowdsourced from smartphone users. CSP users install a smartphone application that exploits intermittent opportunities throughout the day to sample the microphone and camera whenever the device is exposed to the environment, as when users receive calls, check email, or browse the Web. These sampled images and audio clips contain a rich collection of hints about each place the user visits, including written text (e.g., menus, store signage, posters), spoken words (e.g., when a customer purchases a cup of coffee) or physical objects (e.g., cups, cars). To extract these hints from the environment, CSP incorporates a variety of image- and audio-based classifiers (e.g., scene classification, optical-character-recognition, object and speech recognition, sound classification). The output of these classifiers is merged with conventional location-based signals from WiFi and GPS sensors to segment user trajectories into separate place visits as well as provide additional features that discriminate places. In CSP, place characteristics are learned using topic models [8] typically applied to text collections. With this approach, image and audio classifier outputs and location-based signals are represented as discrete tokens (words) grouped by place visit (documents). Topics learned from the model correspond approximately to a place category, with individual places represented as a weighted combination of place categories. CSP can automatically categorize previously unseen places by inferring the topic distribution for the new place and assigning a category based on the dominant topic.

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

Our paper makes the following contributions:

- CrowdSense@Place is, to the best of our knowledge, the first framework for characterizing places that exploits opportunistically crowdsourced images and audio from smartphones – in addition to using more conventional sensors (e.g., GPS/WiFi). By intelligently leveraging this new source of sensor data, we can differentiate a greater number of place categories than currently possible using existing techniques.
- We propose a topic-model-based approach to modeling places that can effectively combine a variety of image- and audio-based classifiers (e.g., scene recognition, OCR, speech recognition, etc.) along with mobility-based signals from GPS/WiFi sensors. Our design and evaluation indicates which classifiers are effective for place categorization, with some classifiers being tuned for this particular application.
- We have evaluated CSP with a seven-week, 36-person deployment using commodity smartphones. Our primary finding demonstrates that CSP can classify the 1,241 places study participants encountered into a total of seven place categories, while still maintaining high levels of accuracy – with 69% achieved across all place categories.

TOWARDS UNDERSTANDING PLACES

In this section we overview existing approaches to recognizing and categorizing places before discussing how opportunistically sampled images and audio can be used to characterize the everyday places that users encounter.

Existing Approaches. The study of places – locations that are semantically meaningful to everyday people – has primarily focused on two main aspects: (1) the discovery of place visits mined from user trajectories (e.g., a time-series of GPS coordinates); and (2) the allocation of descriptors to places that are discovered, such as place categories (e.g., "theatre," "drug store"), informal labels (e.g., "parents' house"), or activities associated with the location (e.g., "eating," "exercise"). Place-discovery techniques [5, 19, 16, 13] commonly rely on location information based on GPS or WiFi sensors to determine features, such as the duration a user remained in the same logical location.

Techniques for allocating descriptors to places have employed a relatively more diverse range of data sources, either relying on data collected in situ while users are visiting places (e.g., [19, 31]) or exploiting existing large-scale data collections, such as point-of-interest databases (e.g., Bing, Yelp) or location-based community-generated content (e.g., Twitter, FourSquare). In [14, 36], data from personally carried devices is augmented by incorporating the user into the loop, with users either providing or confirming location semantics. Techniques proposed in [20, 34, 24] leverage FourSquare check-in activity to determine place categories.

Are Location-based Lookups the Answer? An intuitive approach to accumulating information about many places is to rely on the increasingly rich place information available on the Internet. One example would be performing a search of a location-based service (e.g., local search, recommendation services, Web search) based on the user's location coordinates. However, in practice, it is not always possible to accurately know which place a user is located based purely on their location estimate. The error in GPS-, GSM-, or WiFi-based location estimates often ranges between 10 and 400 meters. Within this margin of error, the user may be present in one of several different places. [20] studies precisely this issue in the Beijing area and reports, for example, that the average 50-square-meter region has more than four distinct places. Similarly, we find that during CSP deployment, 426 of the 1,241 total place visits cannot be correctly associated with a place based solely on the location estimate of the user's smartphone. We observe that this occurs, for example, when users visit multiple places within a single large building (e.g., shopping mall). Because they are indoors, their location estimate cannot update, making it difficult to determine which place they are visiting. In the Evaluation section of this paper, we compare CSP's performance to a baseline approach that leverages solely location estimates and a large-scale location database; our results show that CSP outperforms this technique by 40% when performing place categorization.

Leveraging Rich Visual and Acoustic Place Hints. Different places, such as restaurants, stores, homes, and workplaces often contain a variety of visual and acoustic clues that allow people to intuitively understand a surprising amount about a location, even if they have never been there before. To better illustrate the types of hints that are available with this approach, we manually examine the images and sounds sampled from different types of places in a large dataset collected during the evaluation of CSP (see the Evaluation section for additional details). Figure 1 shows a set of captured images from diverse places located in Los Angeles, Beijing, Seoul, and San Francisco. In Figure 1, we see a coffee cup, a distinctive coffee store brand logo, and words associated with coffee (e.g., "blend," "roast") that appear to have been taken near the cash register during payment. Figure 1 also shows shoes mounted on the wall and an assortment of signs describing the store (e.g., "city chain," "converse"). Our experiment logs which smartphone applications are being used when images are captured. We find that these particular images are taken as users place calls, send text messages, and interact with their music applications. What Figure 1 cannot illustrate are the additional acoustic hints present in these locations, which capture not only a place's characteristics but also how the people in the place behave. Audio clips from the coffee shop capture the exchanges between the customers and employees as coffee is ordered and payed for, or words spoken by baristas when orders are ready for pickup (e.g., "coffee," "macchiato," "non-fat"). Similarly, within the clothing stores, audio clips capture employees answering customers questions as to clothing sizes or colors, and welcoming them to the store (while often stating the store name). Finally, as can be seen in the images and overheard from the audio clips, much of the data collected is unusable.

Due to the uncontrolled nature of collection, which is trans-

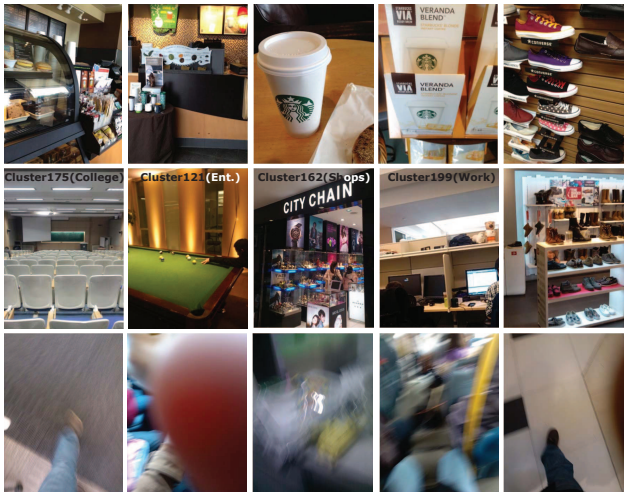

Figure 1. Example of opportunistically captured images. Images on the top two rows show hints for inferring the type of place, such as objects (coffee cup or shoes) or text (signs or brand names). In the bottom row, we see noisy images caused by blurring or camera direction.

parent to the user, images are frequently blurry or capture unhelpful scenes (e.g., the floor, roof, or sky). Unsurprisingly, this pattern is repeated in the audio clips, which frequently contain too much background noise to be intelligible or simply capture silence. CSP currently overcomes this problem with simple brute force: users collectively capture large volumes of both image and audio data daily and repeatedly visit places that are important to them. Crowdsourcing allows CSP to circumvent the limitations of data quality even if only a fraction of collected data is ultimately usable.

Our exploitation of opportunistically captured sensor data is related to the more general concept of *opportunistic sensing* [10], which proposes to collectively leverage sensors in consumer devices to form large-scale sensor networks. The smartphone application CenceMe [1] adopts this opportunistic approach and collects images during phone calls or sensor-based triggers towards a larger goal of using phone sensors to automate user participation within social networks. CenceMe and CSP differ because they have completely different objectives (place understanding compared to social networking); additionally, with CenceMe, no inference is performed on the collected images (although inference is applied to other sensors such as the accelerometer). Attempting to understand the user environment from body-worn sensors including cameras and microphones is also similar in spirit to projects, such as SenseCam [23] and various wearable sensor systems [30] used to build "life-logging" applications. Unlike these projects, which capture sensor data relatively continuously using purpose-built devices deliberately deployed by the user, CSP only has sporadic opportunities to capture data and must rely on crowdsourcing to accumulate enough "clean" data to achieve its application objectives.

The use of a wider range of sensing modalities to improve location services has been previously considered, as in [7], which improved localization accuracy by exploiting smartphone sensors, including the camera. However, the objective in [7] was to determine the physical boundaries of a logical location (e.g., a McDonalds outlet). CSP is only concerned with place classification and relies on existing methods (e.g., WiFi) to perform place segmentation; as such, both projects are complementary to each other. VibN [35] is a smartphone application that improves point-of-interest search and recommendation using both manually and opportunistically collected phone sensor data. Through the collection of microphone data, along with user surveys and mobility patterns, VibN identifies popular points of interest in the city. In contrast to CSP, VibN performs no analysis over collected audio and requires the user to manually listen to audio clips to decide if the place is of interest. Potentially, the techniques developed in CSP could be applied within VibN to automate some of these manual stages. CSP has a closer relationship with sensor fusion frameworks that attempt to understand the physical environments developed by the robotics community. For example, [33] attempts to utilize cameras along with other sensors (e.g., laser based range-finding) to categorize physical environments (e.g., kitchen, living room). However, these techniques assume carefully positioned and calibrated sensors, and are concerned with different types of classification that can assist with the navigation and interaction of the robot within these locations.

CROWDSENSE@PLACE

In the following section, we describe the overall architecture of CrowdSense@Place and detail the key processing stages performed when categorizing places using crowdsourced smartphone sensor data.

Overview

CSP is split between two software components – namely, a smartphone application and offline server-side processing of the collected data. The smartphone application operates as a background service that recognizes places using radio fingerprinting of nearby WiFi access points. CSP opportunistically captures images and audio clips at this location – unless the user has previously prevented data collection at this particular place or for a period of time (e.g., disabling sensor collection for six hours). Based on hints about place category mined from this collected data, and combined with data collected by other users, CSP can automatically determine the type of place (e.g., restaurant) without user intervention. By using CSP, a smartphone can be aware of the place category of the user's location, sharing this information with any installed location-based/context-sensitive applications.

To bootstrap the place category, recognition models employed by CSP users can annotate the category of place they are in, which allows CSP to learn over time which collection of place hints (e.g., spoken words, keywords seen on signs) most often correspond to a particular place category. Not all users need to provide place annotations because the training examples from all users are shared to build a single place category model. Similarly, not all places need to be annotated – place category models are designed to generalize to never-before-seen places. Finally, even if users disable the collection of images and audio data, they can still benefit because places they visit might have already been categorized

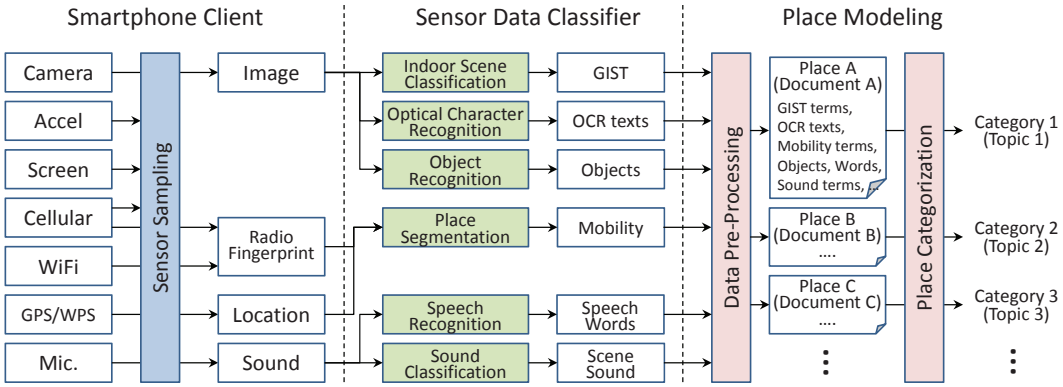

Figure 2. CrowdSense@Place processing stages.

by CSP (using the data contributed by other users).

Figure 2 shows the overall architecture and dataflow that occurs within CSP. It shows mobility information, collected by WiFi and GPS sensors, along with image and audio clips being uploaded from user smartphones to server-side infrastructure for further processing and, ultimately, place category modeling. Data from the smartphone is not immediately uploaded, but rather waits for a period of 24 hours, letting the user decide whether to delete collected data. Furthermore, by waiting, the smartphone client can exploit opportunities to upload the data at a potentially lower energy cost by transmitting when the phone is line-powered and/or WiFi connectivity is available – which commonly occurs while the phone is recharging. During server-side processing of the collected data, CSP applies a variety of classifiers to mine hints as to the place category. CSP employs object recognition, indoor scene classification, and optical character recognition to process collected images. Similarly, a speech recognizer and sound event classifier are applied to collected audio clips. To model place categories, CSP adopts a topic-modeling approach that incorporates the output of these classifiers, along with user trajectory information. Places encountered by CSP users are modeled as documents, and the output of classifiers along with user mobility patterns are discretized into terms that populate each document. A subset of all documents (places) are labeled by users with a single overall document topic (place category). Through topic modeling, each topic is related to a distribution of terms and each place is related to a distribution of topics. We find that, for most places, a dominant topic emerges that represents the category for this place. As new places are presented to CSP, the learned topic model is applied, enabling the place category to be inferred.

Smartphone Client

The CSP smartphone client performs the following primary functions: i) place segmentation, which uses WiFi fingerprints and GPS to discover places and later recognize them again upon subsequent visits; ii) opportunistic crowdsensing, which gathers image and audio sensor data about the places the user visits; and iii) privacy configuration, offering users complete control over all data collected and the ability to stop any collected sensor data from leaving the phone. Secondary application functions include logic for i) uploading collected sensor data and ii) interacting with the CSP servers to receive the predicted place category. We developed our prototype application for Android smartphones and implemented it as two software components: the first is a background service responsible for sensor sampling and place segmentation; the second is a simple user interface and is largely responsible for offering privacy controls and allowing users to manually label places.

Place Segmentation. As users move from location to location, each distinct place is recognized based on its unique WiFi fingerprint. This is a standard approach for place discovery, commonly used in the literature [9, 18]. During standard operation, our smartphone client regularly performs WiFi scans to identify nearby WiFi access points. Whenever a WiFi fingerprint is encountered that is unlike those previously seen, a new place is assumed to have been discovered. Similarly, previously visited places are recognized based on their WiFi fingerprint being sufficiently similar to fingerprints that have been observed earlier. More formally, our WiFi fingerprint similarity function $\mathcal{S}$ is defined using the Tanimoto Coefficient:

$$\mathcal{S} = \begin{cases} \text{different (move)}, & \text{if } \frac{\vec{f}_{t_1} \cdot \vec{f}_{t_2}}{\|\vec{f}_{t_1}\|^2 + \|\vec{f}_{t_2}\|^2 - \vec{f}_{t_1} \cdot \vec{f}_{t_2}} \leq \varphi \\ \text{same (stationary)}, & \text{else} \end{cases}$$

where $\vec{f}_{t_i}$ is a vector of WiFi SSIDs (i.e., WiFi access point names) scanned at t_i for a certain duration, and φ is the similarity threshold. The output of $\mathcal{S}$ is a similarity metric ranging between 0.0 and 1.0. Place changes are detected by evaluating $\mathcal{S}(\vec{f}_{t-1}, \vec{f}_t)$. If $\mathcal{S}$ exceeds $place_{thres}$ the two places are determined to be the same; otherwise they are assumed to be different. Discovered places are associated with the most recent GPS estimate, allowing the WiFi-fingerprint-defined place to be tied to a physical location.

Sensor Sampling. Our smartphone client adopts a simple heuristic to improve the quality of image and audio data collected; sampling occurs after a small random delay once the user starts an application or uses an important phone func-

tion (e.g., receiving a phone call). By adopting this practice, the phone samples when it is exposed to the environment. However, data quality is still highly variable and often poor (e.g, images captured of the floor or audio clips overwhelmed by background noise). To provide some limited awareness of phone resources, our client maintains a coarse sampling budget of a fixed number of images and audio clips that is reset when prolonged periods of recharging occur (monitored by system events that indicate the phone is line-powered). Moreover, available storage is also monitored, and the application never samples when the phone is below a minimum amount of available storage space.

Privacy. Given the sensitivity of the sensor data CSP collects, providing the users with control over their own data is paramount. All data is forced to reside on the smartphone for at least 24 hours, during which time, users can delete any data they are uncomfortable with CSP using. For this purpose, our client incorporates a simple interface that allows users to view all images and play all audio clips, which they can then manually choose to delete. To further simplify this process, with the press of a single button, users can decide to purge all collected sensor data for the previous 1, 6, or 24 hours. Finally, as a preventative measure, users can also pause data collection for an upcoming time interval (again 1, 6, or 24 hours) if they anticipate sensitive events occurring. Alternatively, users can inform the client to never collect data at a certain place (e.g., home, office).

Sensor Data Classifiers

All image and audio data collected by the CSP smartphone client is processed through a series of classifiers chosen to extract various place category hints about each place users visited. CSP currently utilizes five classifiers: three that operate on image data, and two that focus on audio. In the following subsection, we describe each of these in turn.

Optical Character Recognition. To mine written text found in posters or signs within places, CSP incorporates a commercial-grade OCR engine developed by Microsoft and in use in a number of consumer mobile applications (see [15] for more information). The engine provides well-defined APIs that allow us to determine both recognized words and the engine's confidence in each recognition result.

Indoor Scene Classification. We leverage the techniques developed in [26] to perform indoor scene classification. This approach attempts to recognize categories of indoor environments based on both global and local characteristics of indoor scenes (e.g., recognizing the strong horizontal visual patterns present in supermarket shelves). Experimentally, we discover that this classification technique works best in the CSP framework if we diverge from the original classifier design. With CSP, we first extract GIST[1] features [25] from each training image. GIST features are often used in the literature to capture scene characteristics. The images are clustered within a GIST-based feature space using standard k-means clustering. Then, when CSP receives a new image, we do not produce a classification result but instead produce a vector in which each element is determined by how close the image is to each cluster center after we have extracted the GIST features.

[1]GIST is not an acronym but was named because these features capture the *"gist"* of the scene

Objects Recognition. To recognize a variety of everyday objects observed within places, CSP adopts the *exemplar-svm* approach proposed in [21]. This hybrid technique offers state-of-the-performance by combining the benefits of an example-based nearest-neighbor approach with those of discriminative classifiers. We port a reference implementation made available by the authors as a processing stage within CSP. Classifier training is performed using a subset of the objects found in the PASCAL VOC 2007 dataset [11]. Objects are selected based on how likely they are to be found in everyday places. This processing stage can recognize the following 13 objects: {bus, bike, bottle, car, cat, chair, dinning table, dog, motorbike, person, potted plant, sofa, tv}.

Speech Recognition. CSP performs speech recognition using the open source CMU Sphinx recognizer [2]. We use speech recognition primarily to capture place hints found in the conversations of people as they interact (e.g., when a user purchases an item in a store). This recognition system is based on fully continuous Hidden Markov Models [6] and uses Mel-frequency Cepstral Coefficients [12] (MFCCs) as features. We use pre-trained acoustic and language models also provided as part of the Sphinx project.

Sound Classification. Our final classifier attempts to recognize simple acoustic events that occur in the background of audio clips – for example, music playing in the background in a home or store. We use a classifier developed in-house that models sounds using a Gaussian Mixture Model [6], and extracts MFCC features from the audio – just as was done in the speech recognizer. We collected training data for this classifier using a variety of smartphones over an extend time period under everyday settings. Our sound classifier is trained to recognize the following acoustic events: {music, voicing, car, large-crowd noise, alarm}.

Place Modeling

We conclude this section by describing how CSP applies the principles of topic modeling to leverage the output of all classifiers along with user mobility data to ultimately infer place categories (e.g., office, store, gym) for the locations users visit.

Data Pre-processing. CSP begins by building documents, one for each distinct place a user visits. All data collected at a particular place is mined to extract a series of terms, which can then be assigned to a document associated with that place. Terms come in two varieties, depending on whether they are sourced from either classifier or user mobility data.

Classifier Terms. The majority of CSP classifiers produce a sequence of class inferences (e.g., recognized words or objects), each with an accompanying classifier confidence measure. Each class inference corresponds to a different classifier term. All inferences below a certain level of con-

fidence are immediately filtered using an experimentally determined confidence threshold. Filtering is necessary because a lot of collected data is noisy; we must filter uncertain inferences, otherwise discriminative terms can be overwhelmed by noise. The exception to this process is our indoor scene classification stage, which produces a vector for each image. This vector is discretized into a series of terms, each of which correspond to a cluster set of vectors. Before terms are finally added to documents, we apply conventional term frequency analysis [28] to remove any non-discriminative terms (i.e., terms that are common across all places/documents).

Mobility Terms. The underlying assumption in our use of user mobility is that the visit duration and the time of day when people visit certain place categories has a consistent pattern. Intuitive examples of this in practice include a person spending mealtimes at food-related places or being found on weekdays at their workplace from 9 to 5. Encoding user trajectories into terms begins in CSP by extracting the stay-duration and arrival time for each place for each user. Using this data, a residence-time distribution is created for each place in the form of a discrete histogram. Each histogram bin represents a 10-minute period during a single day (i.e., 144 bins). CSP builds two sets of residence-time distributions, one for the weekend and one for weekdays, as suggested in [32]. Consequently, the vocabulary of trajectory terms is 288 (e.g., weekday001, $\cdots$, weekday144, weekend001, $\cdots$, weekend144). A subset of terms are only used if they rarely appear across all places visited by a user, which is determined this time by term frequency-inverse document frequency [28].

Place Categorization. CSP employs the Labeled Latent Dirichlet Allocation (L-LDA) model [27] to categorize places using the documents and terms generated from the crowd-sourced data. L-LDA is an extension of traditional LDA [8]; it allows topic models to be trained with labeled documents and even supports documents with more than one label. Topics are learned from the co-occurring terms in places from the same category, with topics approximately capturing different place categories. A separate L-LDA model is trained for each place category, and can be used to infer the category of new, previously unseen places.

We now briefly overview the training process of the L-LDA model, which CSP uses to extract topics (place categories) from our collection of documents (places). Let each document d be represented by a tuple consisting of a list of word indices $w^{(d)} = (w_1, \cdots, w_{N_d})$ and a list of binary topic presence/absence indicators $\Lambda^{(d)} = (l_1, \cdots, l_K)$ where each $w_i \in \{1, \cdots, V\}$, and each $l_k \in \{0, 1\}$. Here, N_d is the document length, V is the size of the vocabulary, which includes all classifier terms and user trajectory terms, and K is the total number of unique labels in the corpus. The model generates multinomial topic distributions over vocabulary $\beta_k = (\beta_{k,1}, \cdots, \beta_{k,V})^T \sim \text{Dir}(\cdot \mid \eta)$ for each topic k, from a Dirichlet prior η. The L-LDA model then draws a multinomial mixture distribution $\theta^{(d)}$ over the topics that correspond to their labels $\Lambda^{(d)}$. For any document, the final topic distribution $\theta^{(d)}$ will correspond to the relevance of the topic within the document. In other words, $\theta^{(d)}$ indicates the strength of the place categories that are present in any place.

Category	# of place	# of visit	Stay duration (hour)	# of image	# of audio
College & Education	120	1,570	2,222	60	-
Arts & Entertainment	89	218	361	81	37
Food & Restaurant	578	1,426	926	534	236
Home	64	3,899	29,632	72	2208
Shops	112	255	175	1026	254
Workplace	116	4,882	12,306	386	1307
Others	162	656	491	156	121

Table 1. Description of collected data.

As new data accumulates, CSP can repeat the training process, which revises the relationship between topics and the occurrence of classifier terms and mobility terms in documents. Whenever a new place – previously unseen to CSP – enters the system, a new document is created, d_i and populated with terms based on the available data thus far. The current version of the L-LDA that CSP maintains will be applied to generate $\theta^{(d_i)}$, and CSP will assign a place category based on the topic with the highest relevance.

EVALUATION

In this section, we evaluate CSP's effectiveness in categorizing semantically meaningful places. Our primary result shows that CSP can link places to a wider range of categories than previously possible using existing techniques, while still maintaining high levels of accuracy.

Experimental Methodology

We evaluate CSP with a multi-country deployment using Android smartphones that includes 1,241 distinct places. We compare CSP with two benchmark techniques assuming place categories as defined by FourSquare.

Data Set. We recruit 36 users living in five locations around the world (Seoul, Seattle, Los Angeles, San Francisco, and Beijing). Table 1 describes the data collection, including statistics related to places and place visits. Users tend to gather most images while at stores and food-related places, and they often disable the camera while at home. We find that 22% of images are either blurred or completely black.

Metrics. To evaluate the place categorization performance, we adopt two metrics: (1) accuracy and (2) the distribution of place category topics. Our topic-model approach to modeling places generates a probability distribution of topics (i.e., place categories) at each place. Consequently, a single place can be associated strongly with multiple categories at the same time – which does reflect reality (e.g., a coffee shop can often have a dual secondary purpose as a restaurant). However, to simplify the understanding of our result, we largely rely on the accuracy metric. In this case, we assume the topic with the highest probability is the fi-

Category	Sub categories
College & Education	classroom, library, high school, educational institute
Arts & Entertainment	cinema, theater, museum, exhibit hall, gym, karaoke, gaming room, pool hall, stadium
Food & Restaurant	restaurant, fast food restaurant, cafe, dessert shops, ice cream shops, bakery
Home	home, friend/families' home, dormitory
Shops	bank, bookstore, clothing store, accessories store, shoe store, cosmetics shop, department store, convenience store, supermarket, salons, grocery store, jewelry store, high tech outlet
Workplace	workplace, office, meeting room, laboratory, conference room, seminar room, focus room
Others	transportation, church, temple, hospital, hotel, bars, pubs, clubs, street, unknown

Table 2. Definition of place categories

nal category for the place. Accuracy is then defined to be: $\frac{\sharp \text{ of correctly recognized places}}{\sharp \text{ of places}}$. Occasionally in our evaluation we use the topic probability distribution to more clearly illustrate an aspect that accuracy alone does not capture.

Baselines. Two baselines are used to benchmark the performance of CSP: (1) *GPS* and (2) *Mobility*. To compute GPS we simply give the FourSquare search API [4] the most recent location estimate of the user at the time the user visits a place. Multiple places are typically returned to the request, in which case, we select the closest place to the user's location estimate. Our second baseline, *Mobility*, is identical to CSP and classifies places using the same topic modeling approach; however, topics are built using only user trajectory information (i.e., histograms of residence-time distributions at a place). Existing approaches for determining place category rely on information of this nature.

Place Categories. To evaluate CSP, we use place categories defined by FourSquare [3] and adopt its top-level place category hierarchy. Our study ignores two of the original nine categories – Nightlife and Travel Spots. We find users made very few place visits to Nightlife locations, and we had insufficient data to train our model. The Travel Spots category is excluded because the focus of our work is in place classification, not recognizing mobility type. Table 2 lists all seven categories we use in this study.

The ground truth FourSquare category of each place visited during our study is based, when possible, on the category assigned by FourSquare itself. In some cases, FourSquare doesn't have a record for a place a user visited. For these locations, we rely on manual coding performed by five people, based on the standard FourSquare definitions. The people performing the coding used collected images, audio, and location (by consulting online mapping services to further verify the category). Coders' responses are merged to determine final categories based on majority decision.

Experiment Parameters and Implementation. We implement CSP's crowdsensing client using Android SDK 1.5. The WiFi scanning intervals and window size are 10 seconds and 30 seconds, respectively, and the similarity threshold of the WiFi vector is set to 0.7, as suggested in [9]. We implement the CSP backend on Microsoft Azure.

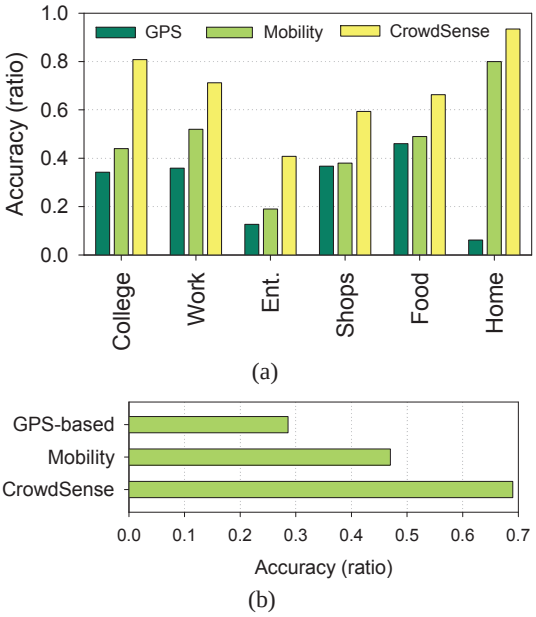

Figure 3. Accuracy of place categorization in (a) each category and (b) overall places

Place Categorization

We begin by investigating the accuracy of CSP when classifying places into the top-level category hierarchy of FourSquare. We used five-fold cross-validation to evaluate the performance of place categorization. Our results show that CSP is able to recognize place categories with 69% overall accuracy across these seven category types, outperforming both baseline comparison schemes. Comparable prior work only employed three or four categories [14, 36]; our use of an extended number of categories is both more challenging and practical for applications to use.

Figure 3 shows the overall accuracy for classifying all place visits in our dataset into the different FourSquare categories. This figure illustrates that CSP outperforms *GPS* and *Mobility* by around 22% to 40%. *GPS* has the lowest accuracy, 29%±16%; we suspect that this is due to poor indoor localization. In addition, *GPS* struggles to differentiate categories of places located near each other (e.g., stores at the same position but different floors). *Mobility* achieves 47%±20% accuracy. We find that mobility patterns have meaningful features that can differentiate some place categories. This is shown in Figure 4. For example, participants tend to spend their nights at home and most of their weekdays at the workplace. Strong peaks in the distribution of food places occur at lunch and dinner time. Across all categories, the home category is the easiest to recognize (and has the highest category average); it is recognized accurately 80% of the time.

To more closely examine the comparison between CSP and the best performing benchmark, *Mobility*, we consider not only whether the categorization is correct, but also which

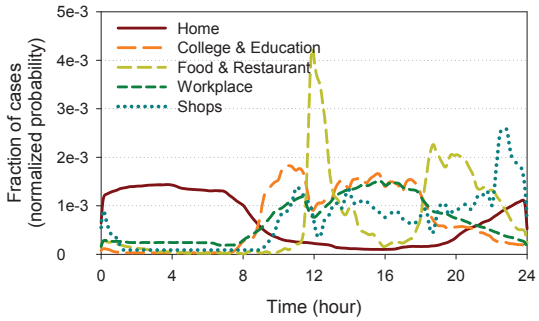

Figure 4. Mobility pattern of several categories

Mobility-based Method

Result / Label	Col.	Work	Ent.	Shops	Food	Home	Oth.
College	0.44	0.30	0.01	0.04	0.04	0.04	0.12
Work	0.33	0.52	0.01	0.03	0.07	0.01	0.03
Ent.	0.07	0.07	0.19	0.15	0.11	0.19	0.22
Shops	0.00	0.06	0.13	0.38	0.06	0.06	0.31
Food	0.10	0.04	0.02	0.08	0.49	0.05	0.20
Home	0.00	0.00	0.00	0.09	0.00	0.80	0.11
Others	0.06	0.14	0.17	0.14	0.04	0.16	0.30

CrowdSense@Place

Result / Label	Col.	Work	Ent.	Shops	Food	Home	Oth.
College	0.80	0.10	0.01	0.01	0.03	0.00	0.04
Work	0.05	0.71	0.03	0.01	0.02	0.01	0.03
Ent.	0.04	0.04	0.41	0.04	0.33	0.00	0.15
Shops	0.00	0.03	0.00	0.59	0.28	0.00	0.09
Food	0.02	0.11	0.05	0.09	0.66	0.00	0.06
Home	0.00	0.00	0.04	0.02	0.00	0.93	0.00
Others	0.05	0.09	0.09	0.20	0.12	0.10	0.36

Table 3. Confusion matrices of place categories for *Mobility* and CrowdSense@Place.

categories are confused with each other. Table 3 shows confusion matrices for CSP and *Mobility*. From this table we can see *Mobility* has trouble recognizing the workplace (44%) and college (52%) categories; this is due to the similarity of mobility patterns for students and office workers relative to colleges and workplaces. In contrast, CSP has high accuracy for these two categories: 80% and 71%, respectively. This is due to the assistance of distinctive place hints from image data even when the mobility patterns for two place categories share common traits. Similarly, we can see that the categories of entertainment and shops are confused under *Mobility*, whereas CSP does not suffer this same problem. In Table 3 we see the comparison between *Mobility* and CPS across all categories.

CSP's approach to place modeling captures the fact that some places can be related to more than one place category. Each place is modeled as a mixture of topics (i.e., place categories). In fact, we believe some of the "errors" in classification reported in the previously discussed results are due to some places being naturally associated with multiple place categories rather than just one. Figure 5 shows the average topic probability of places belonging to all of our supported place categories. For easy visualization, the figure shows just the top three highest-probability topics. We can see from the figure that CSP allocated the highest topic probability to the ground-truth place category. Furthermore, additional

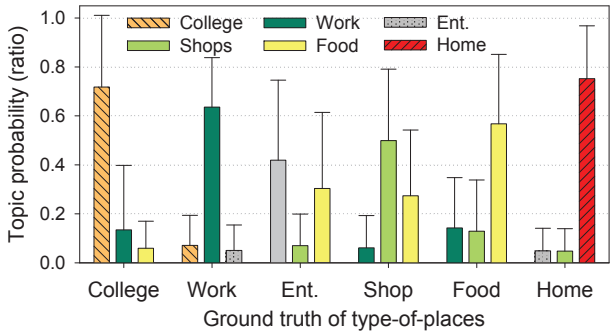

Figure 5. Top-three highest-probability topics for each category.

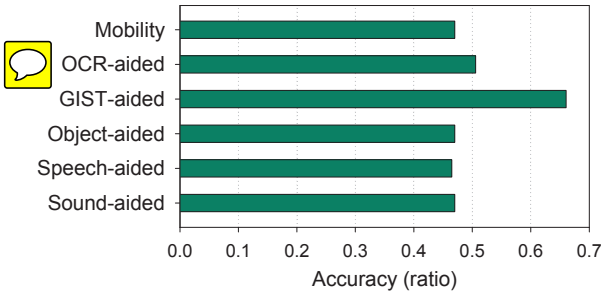

Figure 6. Accuracy of different classifiers used by isolation.

systems (e.g., location-based services, recommendation services) would likely benefit from using a place's topic mixture directly, rather than using a single place category.

Understanding the Benefits of Place Hints

We conclude our evaluation by studying the impact different varieties of place hints have on the performance of CSP place categorization. We find that certain classifiers (OCR and indoor scene classification – that is, GIST) are far more effective than others (e.g., speech recognition). The following set of results can guide future systems that adopt an opportunistic crowdsensing approach.

Figure 6 highlights the performance of CSP when using different classifiers and sources of data in isolation. This figure reports average classification accuracy across the entire dataset. All variations of CSP shown exploit user trajectory data (mobility data), just as the *Mobility* benchmark does. The use of indoor scene classification (i.e., GIST features) has the largest individual impact. OCR does not have a strong overall effect because written words are primarily observed in shopping and food-related places. The performance gains from using object detection, speech recognition, and sound classification are marginal. We find that while object detection is effective in outdoor environments (e.g., cars, buses) it operates poorly on our indoor focused dataset, so the output does not assist strongly with classification. Similarly, the results from speech recognition and sound classification do not have strong discriminative power between tested place types.

Because GIST- and OCR-based information offered the strongest discriminative value we, further investigated their usage in

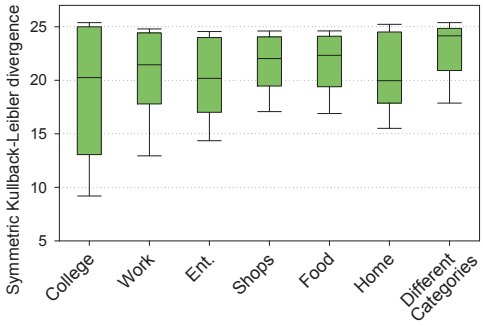

Figure 7. KL divergence between distribution of GIST features corresponding to categories.

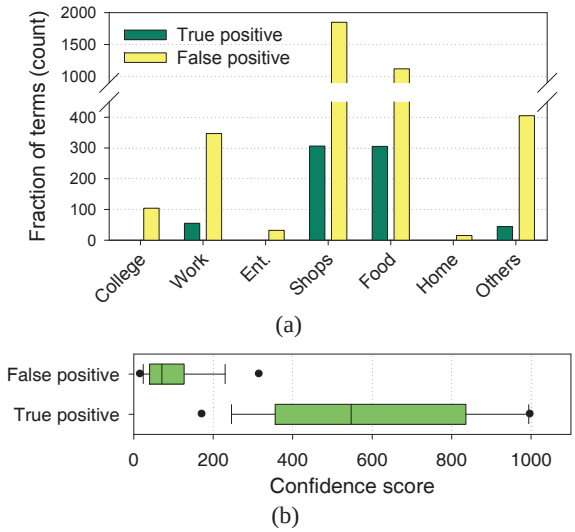

Figure 8. (a) Frequency and (b) confidence score of OCR terms in places.

CSP. Figure 7 presents the Kullback-Leibler divergence between distributions of GIST features each place category. KL divergence measures the distance between two distributions: the low value indicates the high similarity. This figure illustrates that places with same categories have higher similarities compare to those of places with different categories.

We only observe a high-frequency of OCR-recognized words in shops and food-related places. The result matches intuition, given that these environments are often filled with a variety of different signs and posters. Figure 8(a) shows that among the 4,158 words recognized by the OCR classifier, the number of correct words is 451. 86% of true-positive terms are observed in shopping and food places. Figure 8(b) illustrates the confidence score of OCR terms. The distribution of confidence scores is skewed low, in line with our manually checked accuracy rates. Thus, this result verifies the confidence scores of the OCR engine, which we use to filter words likely to have been incorrectly recognized.

Finally, we explored the relationship between the volume of data collected and place categorization accuracy. Intuitively, the more data collected should lead to a more accurate result. Figure 9 supports this finding by showing the increase in

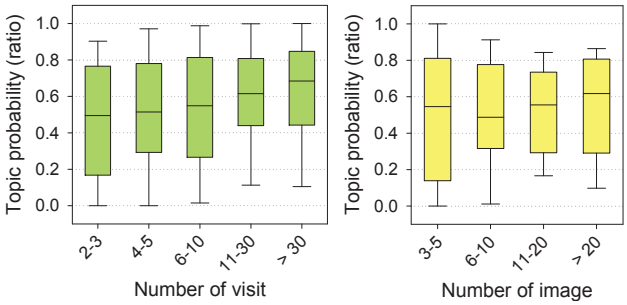

Figure 9. Relation between correctly allocated topic probability and number of visits and images.

topic probability as a function of the number of place visits, or the number of images collected.

DISCUSSION

In what follows, we describe CSP's limitations, along with future research directions, before concluding with potential applications of the CSP framework.

Limitations and Future Work. Our evaluation demonstrates that CSP is a promising, novel approach to performing place characterization. However, our findings also highlight a number of areas that require further investigation.

Finer Place Categorization. We were unable to accurately categorize places as precisely as we initially expected. A number of our classifiers (e.g., object and speech recognition) contributed little to our ability to classify places. However, after manually inspecting our deployment data, we notice that by recognizing a relatively small number of specific place hints, finer-grain place categorization may be possible. For example, we will test speech recognizers trained on a constrained vocabulary of discriminative words. By limiting the vocabulary, we expect higher recognition rates.

Privacy. Although we empowered users to delete (or never collect) data they felt was too sensitive to share, this clearly is insufficient for use by the general public. We plan to pursue a strategy of performing increased local processing of sensor data on the smartphone itself. For example, features will be extracted on the smartphone, with only features (and not raw data) being uploaded to the CSP server. While this does not offer watertight privacy protection, it significantly advances the existing design and is practical; existing privacy-preserving features can be tested, and prior smartphone sensing projects have shown that local processing of this complexity is possible [17].

Activity vs. Place Category. Our deployment study showed us that, in practice, high-quality place hints accumulate slowly. Often, people would not collect any data for hours, and high-quality hints are only collected when many factors coincide, such as a keyword overheard in a conversation or non-blurry image captured that includes a piece of signage. This makes our approach ill-suited to reliably make inferences from collected data on a visit-by-visit basis – for example, to perform some form of activity recognition. Opportunistic Crowd-

Sensing, due to its unpredictable nature, is better suited to incrementally learning static information over long time scales.

Application Scenarios. In the remainder of this section we briefly outline some of CSP's potential uses.

Enhanced Local Search & Recommendations. CSP can provide richer awareness of the types of places a user frequently visits. This information can act as an additional user profile attribute when providing mobile local search services. Similarly, CSP can improve how places are compared and recommended (e.g., searching for similar places). Instead of comparing two places solely based on discrete place categories (e.g., both places are coffee shops), places could be compared using place hints or topic distributions, allowing places that share common fine-grain traits (e.g., lighting conditions or frequent music) to be identified.

Rich Crowdsourced Point-of-Interest Category Maps. CSP can build "maps" that relate places (identified by WiFi fingerprints) to place categories. Such information is a general building block for many mobile and context-aware applications. For example, a targeted advertising application can determine the user's current place category based on a WiFi scan performed by his or her smartphone.

Understanding City-scale Behavior Patterns. Due to the popularity of mobile phones, we can collect large user trajectory datasets relatively easily. Powerful insights about ourselves and our cities have already been extracted from such datasets. By merging maps from CSP that link places to place categories, with user trajectory datasets, we can potentially increase the scope of analysis to include a greater awareness of user activities.

CONCLUSION

In this paper, we presented CrowdSense@Place, a framework for classifying places into place categories. This framework leverages place category hints mined from opportunistically sampled images and audio clips using smartphones. CSP models places using topic models, which allow visual and acoustic place hints to be combined with more conventional signals based on user trajectories. By merging these two sources of data, and exploiting crowdsourcing to gather large volumes of data, CSP is able to categorize places into a broader set of place categories than previously possible. To validate our framework, we tested CSP during a seven-week, 36 person study which that collected data at 1,241 places from five locations around the world. Our results showed that CSP can automatically classify these places into seven different categories, with an average accuracy of 69%.

REFERENCES

1. E. Miluzzo et al. Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application. In *Sensys'08*, pages 337–350. 2008. ACM.
2. CMU Sphinx Speech Recognition Engine. `http://cmusphinx.sourceforge.net/`.
3. FourSquare. `http://foursquare.com`.
4. FourSquare Search API. `https://developer.foursquare.com/docs/venues/search`.
5. D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement Across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.
6. C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.
7. M. Azizyan et al. Surroundsense: Mobile Phone Localization via Ambience Fingerprinting. In *Mobicom'09*, pages 261–272. 2009. ACM.
8. D. M. Blei et al. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
9. Y. Chon et al. Mobility Prediction-based Smartphone Energy Optimization for Everyday location monitoring. In *SenSys'11*, pages 82–95. 2011. ACM.
10. S. B. Eisenman et al. Techniques for Improving Opportunistic Sensor Networking Performance. In *DCOSS'08*, pages 157–175. 2008.
11. M. Everingham et al. The PASCAL VOC2007 Results. `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`.
12. Z. Fang et al. Comparison of different implementations of mfcc. *J. Comput. Sci. Technol.*, 16(6):582–589, 2001.
13. J. Hightower et al. Learning and recognizing the places we go. In *UbiComp'05*, pages 159–176. 2005.
14. D. H. Kim et al. Employing user feedback for semantic location services. In *Ubicomp'11*, pages 217–226. 2011. ACM.
15. Jun Du et al. Snap and Translate UsingWindows Phone. In *ICDAR'11*, pages 809–813. 2011.
16. D. H. Kim et al. Discovering semantically meaningful places from pervasive rf-beacons. In *UbiComp'09*, pages 21–30. 2009. ACM.
17. E. Miluzzo et al. Evaluating the iPhone as a mobile platform for people-centric sensing applications In *UrbanSense'08*, pages 41–45, 2008. ACM.
18. D. H. Kim et al. Sensloc: sensing everyday places and paths using less energy. In *SenSys'10*, pages 43–56, 2010. ACM.
19. N. D. Lane et al. Cooperative techniques supporting sensor-based people-centric inferencing. In *Pervasive'08*, pages 75–92. 2008.
20. D. Lian and X. Xie. Learning location naming from user check-in histories. In *GIS'11*, pages 112–121. 2011. ACM.
21. T. Malisiewicz et al. Ensemble of exemplar-svms for object detection and beyond. In *ICCV'11*, pages 89–96. 2011.
22. N. Marmasse and C. Schmandt. Location-aware information delivery with *commotion*. In *HUC'00*, pages 157–171. 2000.
23. D. H. Nguyen et al. Encountering sensecam: personal recording technologies in everyday life. In *UbiComp'09*, pages 165–174. 2009. ACM.
24. A. Noulas et al. An empirical study of geographic user activity patterns in foursquare. In *ICWSM'11*. 2011.
25. A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
26. A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR'09*, pages 413–420, 2009.
27. D. Ramage et al. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP'09*, pages 248–256, 2009.
28. B.-C. Salton, G. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
29. T. Sohn et al. Place-its: A study of location-based reminders on mobile phones. In *EMNLP'05*, pages 232–250. 2005.
30. T. Starner. *Wearable Computing and Conetextual Awareness*. PhD thesis, MIT Media Laboratory, April 30 1999.
31. D. Peebles et al. *Community-Guided Learning: Exploiting Mobile Sensor Users to Model Human Behavior*. In *AAAI'10*, 2010.
32. L. Vu, Q. Do, and K. Nahrstedt. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. In *PerCom'11*, pages 54–62. 2011. IEEE.
33. J. Wu et al. Visual place categorization: Problem, dataset, and algorithm. In *IROS'09*, pages 4763–4770. 2009.
34. M. Ye et al. On the semantic annotation of places in location-based social networks. In *KDD'11*, pages 520–528. 2011. ACM.
35. E. Miluzzo et al. Tapping into the vibe of the city using vibn, a continuous sensing application for smartphones. In *SCI'11*, pages 13-18. 2011.
36. C. Zhou et al. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst.*, 25(3), 2007.