Towards Scalable Activity Recognition: Adapting Zero-Effort Crowdsourced Acoustic Models

Long-Van Nguyen-Dinh, Ulf Blanke, and Gerhard Tröster
Wearable Computing Lab
ETH Zurich
Switzerland
{longvan,ulf.blanke,troester}@ife.ee.ethz.ch

ABSTRACT

Human activity recognition systems traditionally require a manual annotation of massive training data, which is laborious and non-scalable. An alternative approach is mining existing online crowd-sourced repositories for open-ended, free annotated training data. However, differences across data sources or in observed contexts prevent a crowd-sourced based model reaching user-dependent recognition rates.

To enhance the use of crowd-sourced data in activity recognition, we take an essential step forward by adapting a generic model based on crowd-sourced data to a personalized model. In this work, we investigate two adapting approaches: 1) a semi-supervised learning to combine crowd-sourced data and unlabeled user data, and 2) an active-learning to query the user for labeling samples where the crowd-sourced based model fails to recognize. We test our proposed approaches on 7 users using auditory modality on mobile phones with a total data of 14 days and up to 9 daily context classes. Experimental results indicate that the semi-supervised model can indeed improve the recognition accuracy up to 21% but is still significantly outperformed by a supervised model on user data. In the active learning scheme, the crowd-sourced model can reach the performance of the supervised model by requesting labels of 0.7% of user data only. Our work illustrates a promising first step towards an unobtrusive, efficient and open-ended context recognition system by adapting free online crowd-sourced data into a personalized model.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

Keywords

Activity recognition; ambient sound, mobile phone; crowdsourcing; personalization; adaptation; semi-supervised learning; active learning

MUM '13, December 02 - 05 2013, Luleå, Sweden

http://dx.doi.org/10.1145/2541831.2541832.

1. INTRODUCTION

Human activity recognition is important in developing context-aware systems. Being able to sense a user's routines (e.g. working in the office, staying at home) [4], his physical activities (e.g. standing, walking, cycling) [3], or his social context (e.g. having a conversation) [19], relevant information about the user can be captured. Based on the acquired information, actions can be executed. With the ubiquity of mobile phones, and their growing computational power and sensing capability, they enable new opportunities for developing personal context-aware systems in a large scale to perceive and act on what users are doing or experiencing. One example is the commercial application Google Now [1], which provides location-based reminders or arrival time estimates to reach a destination according to users' daily routines. Many more applications have been investigated in activity recognition research that span across a wide range of domains such as healthcare, sports, or entertainment that can profit from understanding the users' context [10]. Prominent examples include capturing daily life activities for health monitoring motivated by activities of daily living index (ADL) by Katz [9] and Bucks [5] or for the social rhythm measurement (SRM) [12]. Using context recognition systems activities can be detected automatically without burdening the user of keeping track of his activities.

Traditionally, most context recognition systems require a time-consuming preparation in which training data (e.g., sensor readings) is collected and manually annotated to build recognition models. Moreover, the daily life of a user often contains highly individual situations, activities, or environments. Also daily life situations can be expressed in a highly diverse way. For example, the – seemingly simple – activity of *working at the office* can consist of typing at the computer, reading, giving a talk, or attending a meeting. As a result, the recognition system has to be trained individually for a specific user and a large amount of training data has to be collected to cover the variability of the user-specific daily life. Clearly, this laborious and non-scalable requirement impede a real-world deployment in which a user can directly use the system, and prohibit the use in many applications such as in the healthcare-related scenarios mentioned above.

In this work we aim at facilitating the deployment of a context recognition system based on audio on mobile phones. To reduce the effort to collect training data, an alternative approach is to use crowd-sourced sound repositories (e.g., Freesound [2]) from the web [19]. The advantages of us-

[1] http://www.google.com/landing/now/
[2] www.freesound.org

ing web collected data are their free availability and a rich representation of possibly open-ended categories of sounds. However, crowd-sourced audio data can differ from data obtained on personal mobile phones. Differing characteristics of the user's surroundings or microphone responses can prevent the recognition when using a crowd-sourced model on user-specific data. In contrast, collecting labels of user data is a burden to every user. Table 1 shows the tradeoff between crowd-sourced audio data and user-centric data recorded from user's mobile phone.

	Crowd-sourced Audio data	**User-Centric data (Mobile phone)**
Annotation Cost	Free	Huge effort (by users or experts)
Length	Short-clips (seconds/minutes)	Long continuous recording (days-months-years)
Location	Unknown, heterogeneous	User's environment surroundings/activities
Device	Unknown, heterogeneous	User's device

Table 1: Comparison between crowd-sourced data and user-centric data

In this work, we take an essential step forward to combine the best properties from crowd-sourced data and user-centric data to obtain a high performing and yet scalable recognition system in terms of user labeling effort. We achieve our goal by adapting a generic model based on crowd-sourced data to a personalized model. To this end, we first collect crowd-sourced training data to bootstrap a context recognition system (as in [19]). Then, we refine model parameters with no to little interaction of the user to improve the recognition performance. We contribute an analysis of different methods for the adaptation. In the first approach, a semi-supervised learning scheme is used to combine labeled crowd-sourced audio data with unlabeled user-centric data. In the second approach, we use an active-learning scheme to detect the most informative user-specific data samples that the crowd-sourced model can not represent well and queries a user to label them. We analyze the tradeoff between labeling effort and accuracy of the recognition system. We provide a thorough evaluation on 7 users with a total data amount of 14 days. The results show that combining crowd-sourced data with user-specific data can achieve accuracies similar to a supervised approach built on user data, but lowering the labeling effort to a minimum. Thus, leveraging both crowd-sourced data and user-centric data can open a chance to build a scalable and efficient context recognition system.

The rest of the paper is organized as follows. Section 2 offers the literature review on auditory context recognition. Section 3 proposes our recognition system to combine the two sources of audio data. In Section 4, we discuss the probabilistic learning framework for context recognition used in our paper. The collected datasets are presented in Section 5. The proposed research is examined by extensive evaluations in Sections 6 and Section 7. Section 8 concludes our work and gives some potential research directions.

2. RELATED WORK

Environmental sound has been used as a rich source of information to infer person's activities and locations [8, 22, 2, 11, 6, 19]. For example, Stäger et al. [22] proposed a dedicated hardware to recognize a set of daily activities based on sound. Lu et al. [11] modeled and recognized sound events on mobile phones. While training data is essential for all these recognition systems, it is time-consuming and non-scalable to obtain sufficient amounts of data with annotations that represent daily life situations. Consequently, most of the previous work are limited to small datasets of sound daily life contexts that are manually collected and labeled under controlled conditions [8, 22, 6, 15].

Although a relatively new idea, mining online multimedia repositories for relevant training data for activity recognition has been proposed by researchers to reduce the effort to collect and label training data as well as increase the number of available context classes. Perkowitz et al. [16] presented the web-based activity discovery using text. Rossi et al. [19] proposed to use the online crowd-sourced Freesound database to obtain a heterogeneous and diverse training data to train sound models to recognize activities of daily living.

Semi-supervised learning and active learning are two different types of techniques in machine learning that minimize the need of labeled training data. Those techniques are highly-motivated where unlabeled data can be easily obtained but annotation is costly or time-consuming to obtain, thus, labels sparse. Semi-supervised learning make use of both labeled and unlabeled data to train a recognition system. Meanwhile, active learning selectively asks labels of the most informative training instances that can generalize the classifier maximally, and thus reduces user's burden of labeling, but still gets good performance. There are many variations of semi-supervised learning and active learning algorithms. A comprehensive survey can be found in [26] for semi-supervised learning and in [20] for active learning, respectively.

According to the best of our knowledge, there is no previous work that investigated adaptation techniques to optimally leverage labeled crowd-sourced audio data and user-centric data recorded from mobile phones to improve the recognition performance but reduce the effort to label user's data. In the work by Rossi et al. [19], Freesound has been used for context recognition with supervised learning. However, they do not consider user adaptation to improve the performance. Zhang et al. [25] used semi-supervised learning to improve sound event classification. In their work, however, labeled and unlabeled data are of the same data source and they did not work with personalized user context. Stikic et al [23] explored both semi-supervised learning and active learning in physical activity recognition, with focus only on user-centric data record in a highly instrumented home environment. In contrast to these works, we aim to improve the recognition of a classifier learned from one free data source – crowd-sourced repository – on another, the user personalized data on mobile phone.

3. CONTEXT RECOGNITION SYSTEM

While the web offers an abundance of labeled data, obtaining labels from a single user is often a tedious and time consuming task. However, with the option of obtaining an abundance of unlabeled data from the user, we employ techniques to optimally use available data. Figure 1 shows an overview of our sound-based context recognition system. In data preprocessing phase, we collect auditory training data from Freesound and user's mobile phone. We then extract acoustic features from the collected audio clips. In the learning phase, we apply machine learning techniques to learn and

adapt a context recognition model based on the two sources of data. In the recognition phase, the context recognition model will be used to infer user context from data recorded on user's mobile phone. We describe each component in our proposed system in the following.

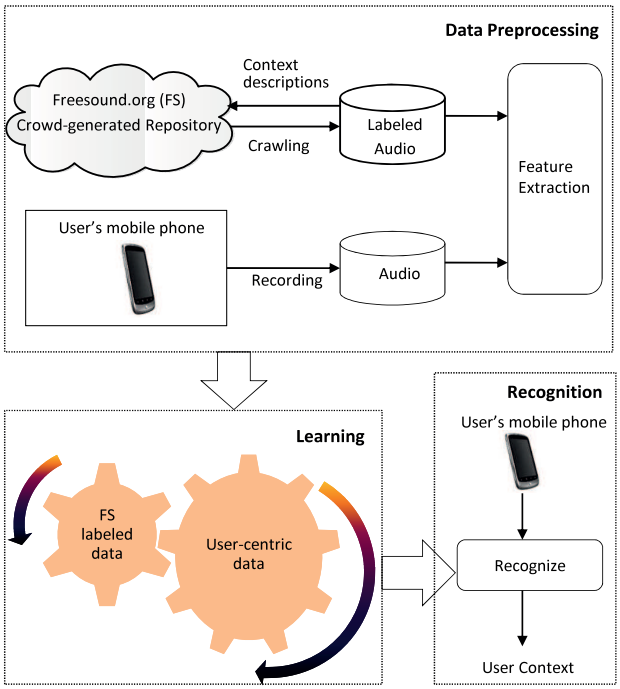

Figure 1: The data processing flow of our sound-based context recognition that combines Freesound data and user-centric audio data

3.1 Data Preprocessing

Freesound Repository. Freesound [1] is an online sharing repository of crowd contributed sound data. Sounds in Freesound are contributed by a very active online community and thus, the number of available sounds has increased rapidly. Currently, the database stores about 170000 samples uploaded by 6000 contributors. Sounds are often annotated in free-form styles and the tags come from very diverse vocabularies. Moreover, crowd-contributed sounds are recorded in a wide variety of situations, conditions, motivations, and skills.

Crawling labeled audio from Freesound. In our system, we focus on everyday situations such as *dining in a restaurant* or *transporting* For each context class, we use its name as a keyword (e.g., "restaurant") to search for the sound clips in Freesound that are tagged with the keyword. The list of context classes can be provided by a user who uses the context recognition system. In health monitoring systems, the list of context classes is usually defined by a specialist beforehand [12]. However, with the diversity of context classes in the crowd-sourced repository, it is easy to extend the recognition system by specifying new context classes and then extracting training data samples for those classes from the crowd-sourced repository. We retrieve only sound clips with the highest average rating (i.e., high quality) given to the sounds provided by the Freesound community. Multiple sound samples are then labeled with the corresponding context class. All the retrieved audio samples were converted to WAV format with a sampling frequency of 16 kHz and bit depth of 16 bits/sample. We manually filter the downloaded audio clips that are irrelevant to the assigned context class. For automatic filtering techniques see [19]. We do not apply an automatic filtering to remove irrelevant clips because the manual filtering showed the best results [19] and we assume that the small set of short audio clips that we retrieved from Freesound (30 sound clips per context) can be quickly and cheaply filtered by listening.

User Recordings. We record continuously audio data from users' smartphones with a sampling frequency of 16 kHz and bit depth of 16 bits/sample. As with the data from freesound, we store in WAV format.

Extracting audio features. We extract 12 coefficients mel-frequency cepstral coefficient (MFCC) and log-energy on a sliding window of 32 ms length of audio data. The same method are used to extract acoustic features for both audio data from Freesound and the smart phones.

3.2 Learning Phase

We propose to use semi-supervised learning and active learning schemes based on Gaussian Mixture Model (GMM) to combine two data sources: Freesound and user-centric data on mobile phones. We name two approaches as *Semi-supervised Adaptation* and *Active Learning Adaptation* respectively.

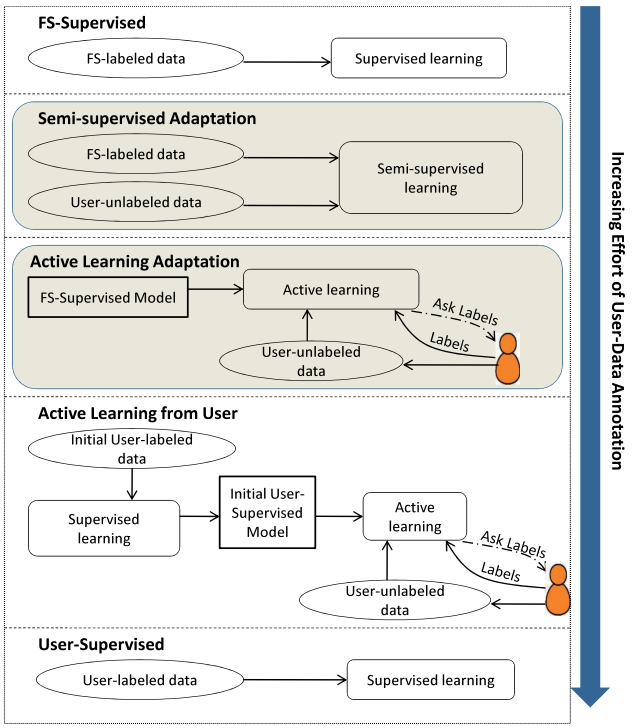

Figure 2: Learning approaches using either Freesound (FS) or user-centric data, or both of them to train the context recognition

Semi-supervised Adaptation. Semi-supervised learning is used to combine Freesound labeled data and user unlabeled data.

Active Learning (AL) Adaptation. We train a bootstrapped context classifier using Freesound labeled data. From that initial classifier, active learning proceeds and iteratively selects the most informative user-centric samples to query for labels. The classifier is then retrained and adapted with the new user-centric labeled data.

Semi-supervised Adaptation and *AL Adaptation* are illustrated in Figure 2. Details of the semi-supervised learning and active learning algorithms are discussed in Section 4.

3.3 Recognition Phase

The GMM models which are obtained from the learning phase are used to recognize user daily contexts based on audio data recorded from the smartphone. We construct a two-level classification. At the low level, audio instances extracted from windows of 32 ms are classified by the GMM models. The context class with the highest probability to generate an instance is assigned to that instance. At the high level, a decision is made on the longer segment (2 seconds) by taking a class with the highest frequency in the segment as a label.

4. PROBABILISTIC FRAMEWORK FOR CONTEXT RECOGNITION

As we mentioned before, immediate usage of crowd-sourced data to build a context recognition model is suboptimal due to lack of user-specific training data, therefore we propose two adaptation techniques based on GMM to tailor a context model build from crowd-sourced data to a personalized context model. In this section, we first briefly present the GMM probabilistic framework for context recognition used in this paper. Then the semi-supervised and active learning algorithms built on this framework are presented.

GMM is an effective generative classifier that has been used extensively in acoustic domains (e.g., speaker recognition [18, 17], environmental sound [6, 15, 2]). Let $\mathcal{D}$ be a set of N observed instances $\mathbf{x}_i \in \mathbb{R}^d$ and Ω be a Gaussian mixture model with K components, $c_1, ..., c_K$. Each component c_k (k = 1,...,K) is a Gaussian density conditional model, i.e., $p(\mathbf{x_i}|c_k) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)$, where $\boldsymbol{\mu}_k$ and Σ_k are the mean vector and covariance matrix of the component, respectively. Let us also denote Θ be the set of parameters of the model Ω, $\Theta = \{\boldsymbol{\mu}_k, \Sigma_k, \pi_k\}_{k=1}^K$, where π_k is the prior probability of the component c_k.

Given the data $\mathcal{D}$, the maximum log likelihood estimation (MLE) is used as a criteria to define the best model $\widehat{\Theta}$ to fit $\mathcal{D}$, i.e., $\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log p(\mathcal{D}|\Theta)$, with

$$\mathcal{L} = \log \mathrm{p}(\mathcal{D}|\Theta) = \log \prod_{i=1}^{N} p(\mathbf{x}_i|\Theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)$$

In our work, each context class can contain multiple Gaussian components. The probability that an instance x_i belonging to a context class y_i is computed as the sum of the probabilities of the mixture components belonging to the context class to generate the instance.

$$P(y_i|x_i;\Theta) = \frac{\sum_{j=1}^{K} \mathbb{1}\{c_j, y_i\} P(c_j|\mathbf{x}_i;\Theta)}{\sum_{k=1}^{K} P(c_k|\mathbf{x}_i;\Theta)},$$

with $\mathbb{1}\{c_j, y_i\} = 1$ if class y_i contains component c_j.

4.1 Semi-supervised Learning using EM

The goal is to find semi-supervised parameters Θ that maximize $\mathcal{L}$ to fit the labeled (i.e., Freesound data) and unlabeled (i.e., user-centric data) observations. The Expectation-Maximization (EM) approach [7] is a standard procedure to find the locally optimal parameter set $\widehat{\Theta}$. In our work, we consider that each context class can have multiple Gaussian components.

Inputs: Collections X^l of labeled data of l instances and X^u of unlabeled data of u instances. The training set $X = X^l \cup X^u$.

1. Initialization. For each class i, build a GMM model $\widehat{\Theta_i}$ from the labeled data X_i^l of that class. Merge all components of classes to have initial Θ.

2. Loop until converge. (i.e., the change in log likelihood of the training data X is less than 10^{-4}):

E-step: Use the current model to estimate the probability that each mixture component generated each instance (i.e., component membership).

$$\gamma_{ij} = P(c_j|\mathbf{x}_i;\widehat{\Theta}) = \frac{\pi_j \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)}$$

Restrict the membership probability estimates of labeled instances to be zero for components associated with other classes, and renormalize.

M-step: Re-estimate the GMM model, $\widehat{\Theta}$, given the estimated component membership of all labeled and unlabeled instances.

$$l_j = \sum_{i=1}^{l+u} \gamma_{ij}, \quad \pi_j = \frac{l_j}{l+u}, \quad \boldsymbol{\mu}_j = \frac{1}{l_j} \sum_{i=1}^{l+u} \gamma_{ij} \mathbf{x}_i$$

$$\Sigma_j = \frac{1}{l_j} \sum_{i=1}^{l+u} \gamma_{ij} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T$$

for all j = 1,...,K

4.2 Active Learning

Active learning starts with an initial learner trained on a small number of labeled instances in the training set with supervised GMM. The learner then iteratively queries labels for one or more selected instances in the unlabeled training set, learns from the new labeled set, and then updates the learner. In the *Active Learning Adaptation* approach, the initial learner is built on the Freesound labeled data and the unlabeled training set is user-centric data.

In the context recognition on mobile phones, one can imagine an interactive online strategy that asks the user to annotate his current context when the learner confuses how to label the current situation. It is called stream-based active learning [20]. On the other hand, in pool-based active learning [20], the query decisions are made offline after collecting the entire unlabeled training set. The learner evaluates and ranks the entire unlabeled set to select to the best queries. In the auditory context recognition, one can imagine that the audio segments corresponding to the query instances are extracted and given to the user to annotate them.

We use a pool-based algorithm to query labels on the user-centric unlabeled dataset. Specifically, we use entropy [21]

as an uncertainty measure to find the most informative unlabeled instance to query a label.

$x^*_{ENT} = \arg\max_{x_i} -\sum_i P(y_i|x_i;\Theta)\ \log\ P(y_i|x_i;\Theta)$, where y_i ranges over all possible class labels.

In our work, we use a classification window of 32 ms (see Section 3.1). However, everyday context classes defined in this paper often last for at least a few minutes. Therefore, instead of asking the label for the 32 ms data segment only, we extend the labeled segment with a window of one minute around the queried instance.

5. DATASETS

For our evaluation we collect two datasets: 1) To obtain a user-centric dataset, we collect data recorded from users' smartphones; 2) For the crowd-sourced dataset we make use of the *Freesound* repository as in [19, 13].

User-centric data. We use android-based smartphones (Samsung Galaxy S2) with headset microphones for continuous sound recording. Participants were asked to record two full working days in their ordinary setting. All participants live in Zurich, Switzerland and thus, the transportation includes tram, train, bus and car. The recording application also provides an annotation tool in which user can annotate his current contexts as a ground truth. Specifically, users can indicate when a context class starts/stops happening. We do not ask the user to label fine-grained sound events, but longer lasting everyday contextual situations. In our work, we also want to support the recognition of individual and user-dependent context classes. Therefore, users can annotate different set of context classes, individual to their daily situations. Table 2 shows the list of classes provided by 7 participants and the corresponding distribution of classes in the dataset recorded by users on mobile phones. Context classes are about working, feeding, transportation and social interaction which are useful in health monitoring [12]. For each recording day, at least 9 hours of audio data were obtained for each user. As can be seen in Table 2, users spend most of the time in the office and discuss their works with colleagues. In total, about 130 hours of audio data has been collected from mobile phones for the study.

	Context Classes and Class Distribution (%)
User 1	office (83), tram (1), train (10), conversation (6)
User 2	toilet (1), office (50), restaurant (5), street (1), conversation (43)
User 3	office (37), restaurant(7), street(12), tram(2), conversation(42)
User 4	toilet (1), office (70), restaurant(2), street (4), tram (1), conversation (22)
User 5	toilet (1), office (63), restaurant (7), street (7), tram (1), train (7), conversation (14)
User 6	toilet (0.4), office(70), restaurant (8), street (4), tram (6), train (5), car (1), conversation (5.6)
User 7	toilet (0.2), office (21), restaurant (9), street (4), tram (5), train (6), car (2), bus (0.2), conversation (52.6)

Table 2: User-dependent context classes and the corresponding distribution of classes in dataset

Freesound. From the list of context classes provided by the users, we retrieve audio data for those context classes from Freesound. As a result, we download sound clips for 9 context classes from Freesound as shown in Table 3. For each class, we retrieve 30 sound clips, tagged with the label of the class, with the highest average rating given to the sounds. Besides the class label, a sound clip also has other tags that usually describe different sound events occurring in the sound clip. Table 3 shows the subset of tags in Freesound clips that we download for each context class. As can be seen, each context class contains the heterogeneity of sound events and recording conditions. For example, being in the office consists of multiple sound events such as typing, stapling, printing, etc. After manually filtering for quality, we have 163 audio clips (143 minutes) for 9 context class to train sound models. This data from Freesound is denoted as FS.

Context Class	**Tags of Freesound Clips**
Office	office, door-open, typing, locking, coffee-machine, stapler, paper-shuffling, print
Bus	bus, door-open, horn, footstep, air-brake, stop, speeding, air-pressure-release
Car	car, highway, forest, car-door, overtake, start, stop, footstep, brake, snow, rain
Train	train, rail, leaving, accelerating, wheels, door, railway, underground, passing, voice
Tram	tram, door, trolley, passing, beep, creaking, tunnel, bell, announcement, brake
Street	street, pedaling, chatter, people, music, bike, announcement, foot, bell, car, horse
Restaurant	restaurant, chat, drink, eat, pour, liquid, food, ice, dish, nibble, grill, clinking, music
Toilet	toilet, splash, water, scrub, lavatory, sink, shower, brush, urinal, flush, hand-dryer
Conversation	chat, talk, noise, bustle, phone, scream, yell, panic, male, female, English, Spanish

Table 3: The heterogeneity of sounds from freesound for each context class

6. EVALUATION

To evaluate the best use of crowd-sourced data in personalized context recognition, we address the following research questions:

1. Does the *Semi-supervised Adaptation* of a crowd-sourced model improve the recognition of user's contexts on mobile phones?
2. Can the *Active Learning Adaptation* find user data instances that the crowd-sourced data can not represent well to ask for labels and quickly achieve good performance with minimal number of label queries?
3. Does the *Active Learning Adaptation* of a crowd-sourced model perform better than active learning model based on user data only in terms of accuracy and number of label queries?

To answer these question, we compare our proposed approaches with three baseline non-adapted learning approaches. Specifically,

- *FS-Supervised*: A supervised GMM trained on Freesound data only without adaptation with user-collected data.

- *User-Supervised*: A supervised GMM trained on the user-centric annotated training data only.

- *Active Learning (AL) from User*: An active learning scheme on user-centric data only. In this baseline approach, we assume that initially each user can contribute and label randomly one minute of data ($\approx$ 1875 instances) for each context class. Totally, this labeled data takes about 1% of training data. An initial GMM classifier is trained from that labeled data, then the active learning is applied to query labels for the uncertain samples.

Figure 2 illustrates the three baseline approaches also with our proposed approaches. They are ordered increasingly in the effort of user data annotation (i.e., (1) *FS-Supervised*, (2) *Semi-supervised Adaptation*, (3) *AL Adaptation*, (4) *AL from User*, and (5) *User-Supervised*). Among them, *FS-Supervised* and *Semi-supervised* do not require any effort to label user data. It is not clear now whether *AL Adaptation* is better than *AL from User* in terms of number of label queries. We will discuss it in detail in the next section. Our goal is to find the best approach in terms of accuracy and labeling effort.

The experiments are performed based on the partitioning of the two-day recording audio data from user's mobile phone into two halves. The first fold (F1, 50% of user data for each class) is used with no labels ($F1_U$) or all labels ($F1_L$) or a small part of these labels ($F1_{1\%L}$ and $F1_{99\%U}$) in training phase. The second fold (F2, another 50% of user data for each class) is used for testing in recognition phase for all five approaches. Table 4 shows the usage of two sources of data in learning phase of five approaches.

	Data usage
FS-Supervised	FS
Semi-supervised Adaptation	FS + $F1_U$
Active Learning Adaptation	FS + $F1_U$
Active Learning from User	$F1_{1\%L}$ + $F1_{99\%U}$
User-Supervised	$F1_L$

Table 4: The usage of two sources of data in learning phase of five approaches

7. RESULTS

Table 5 gives the accuracy of five approaches (we only show the best for the active learning). Figure 3 plots the detailed performance of the active learning approaches over the first 20 label queries.

	FS-Supervised	Semi-supervised Adaptation	AL Adaptation	AL from User	User-Supervised
User 1	0.8	0.86	**0.97**	0.93	0.94
User 2	0.5	0.65	**0.94**	0.82	0.9
User 3	0.58	0.43	**0.81**	0.73	0.72
User 4	0.22	0.25	**0.93**	0.86	0.72
User 5	0.35	0.5	0.80	**0.83**	0.82
User 6	0.54	0.61	0.86	**0.87**	0.85
User 7	0.26	0.47	**0.86**	0.76	0.83

Table 5: Accuracy of learning approaches for 7 users. For active learning (AL), the best performances over 20 label queries are given

As expected, the *User-Supervised* approach gives much better results than the *FS-Supervised* approach. Supervised training on user data clearly captures user-specific environments in the model (i.e., test and training data tend to be similar for each single user) and thus, recognizes well user context in daily routines. The performance of the *FS-Supervised* model drops significantly since the crowd-sourced data hardly covers all user-specific surroundings. However, *FS-Supervised* model does not require any effort to label user data, meanwhile *User-Supervised* requires huge effort to label them. Note that the accuracy of the supervised Freesound model *FS-Supervised* is at similar range to that reported in the work by Rossi [19] which also used Freesound data to recognize users' contexts.

Semi-supervised Adaptation.

The results show that the *Semi-supervised Adaptation* approach significantly improves the performance of context recognition compared to the non-adapted *FS-Supervised* (up to 21%) for six users. Unlabeled training data from user can help adapt the crowd-sourced model to the personalized model without asking any labels for user data. Only for user 3, the *Semi-supervised Adaptation* actually decreases the performance. In this case, the contribution of unlabeled user data in the semi-supervised learning makes the model more uncertain. The result in Table 5 also shows that *Semi-supervised Adaptation* underperforms significantly the baseline *User-Supervised* approach as can be seen in Table 5. Especially from user 4, even the *Semi-supervised Adaptation* can increase the accuracy of *FS-Supervised* by 3%, the accuracies of these two approaches are much lower than that of the *User-Supervised*. Here the Freesound data does not generalize sufficiently to the data recording from that user-specific environments. The visualization of the collected data from the crowd-sourced data and user-centric data in Figure 4 supports our explanation (data dimension reduced to 2 using t-SNE stochastic neighborhood embedding method [24]). As you can see, the crowd-sourced data represents some parts of the user data only.

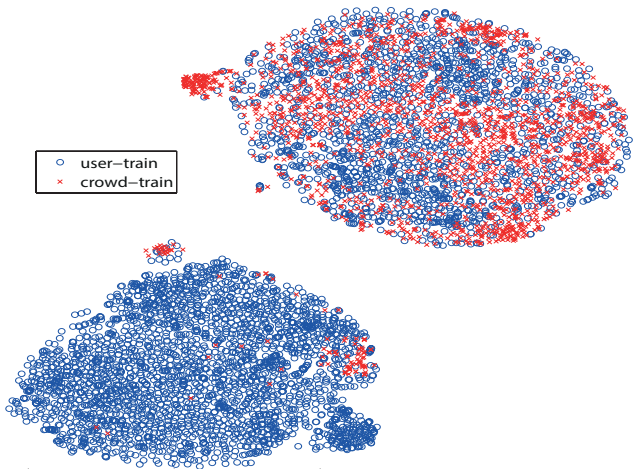

Figure 4: A visualization of the collected data from crowd-sourced Freesound and one user's mobile phone data of *train* context class

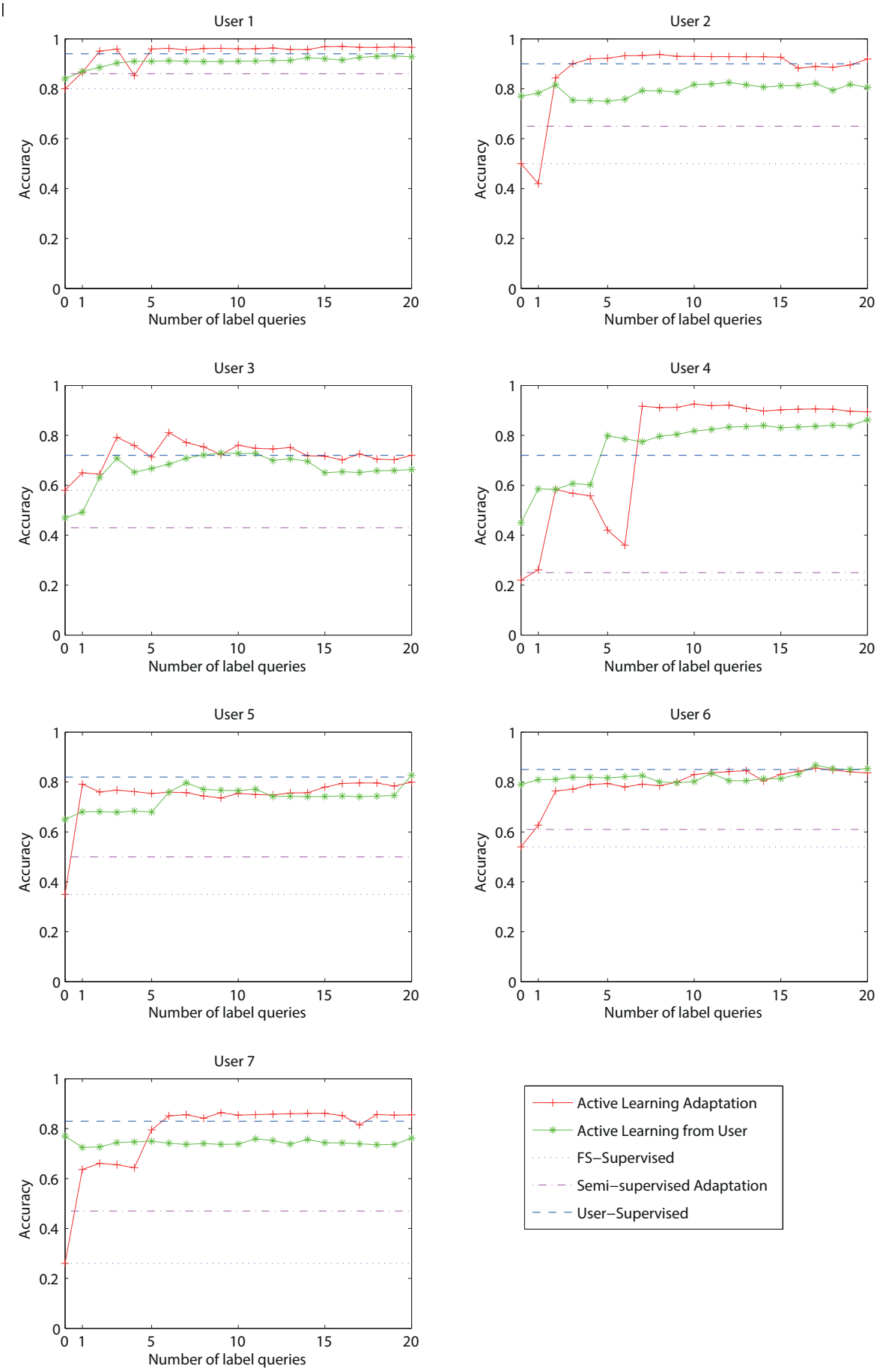

Figure 3: Performance of the active learning approaches over the number of label queries for each user

Active Learning Adaptation.

As can be seen in Figure 3, the *AL Adaptation* quickly reaches the performance of the *User-Supervised* approach and even improves further the accuracy over 20 label queries. Freesound contains rich information about the contexts and after asking a few label queries for user data instances that Freesound can not represent well, the *AL Adaptation* approach may generalize better user's context. After 20 iterative queries of labels, the *AL from User* approach also gets similar or closely results as of the *User-Supervised* learning. Thus, the active learning technique generally can achieve high accuracy using significantly fewer labeled training data (20 queries $\approx 3\%$ of the user training data). In Figure 3, we see that the *AL from User* gets better performance than the *AL Adaptation* in initial number of label queries. However, with more queries eventually, the *AL Adaptation* achieves better accuracies. We can explain this with the same reason: Freesound contains intra-class diversity, thus it may contain user's unseen contexts in the recognition phase and increase model generalization.

To compare the *AL Adaptation* and *AL from User* in terms of number of label queries, we evaluate how many label queries needed for these two approaches to reach the same performance as the *User-Supervised* approach. For the *AL from User* approach, we also count the annotation effort of user to contribute the initial labeled training set to build the initial classifier (This effort is equivalent to the total number of context class that the user has). As can be seen, the *AL Adaptation* requires much less number of queries than the *AL from User* to achieve the same accuracy as the *User-Supervised* approach. It only requires in average 5 label queries per user ($\approx 0.7\%$ of user training data) to get the good performance. Meanwhile, the *AL from User* asks for in average at least 24 label queries per user ($\approx 3.6\%$ of user training data). Moreover, the number of queries needed in *AL Adaptation* is even much less than the total number of context classes that the user has for most of the user. The Freesound data contains training instances of several classes that describe well user contexts and thus, the *AL Adaptation* does not require to ask labeled instances for those classes.

	AL Adaptation	*AL from User*
User 1	2	23
User 2	3	> 50
User 3	3	8
User 4	7	11
User 5	1	14
User 6	10	24
User 7	6	39

Table 6: Number of label queries needed in two active learning (AL) approaches to reach the performance of *User-Supervised* approach. Note that each query requests labels of one minute of data ($\approx$ 1875 data instances)

To analyze in detail which classes for the Freesound-sourced model performs poorly, we investigate the queries of the *AL Adaptation*. We want to reveal which classes are queried for and how much the obtained label improves the classification. For each user, over 20 label queries, we report how many queries performed for each context class y as well as the accumulated improvement of the context class $A(y)$ that these queries contribute to the performance. Specifically,

$A(y) = \sum_{t=1}^{20} \mathbb{1}\{t, y\}(accuracy(t) - accuracy(t-1))$, with $\mathbb{1}\{t, y\} = 1$ if an instance of class y is enquired for a label in the query at time t and accuracy(t) is the accuracy of *AL Adaptation* approach after the t-th query is asked.

Table 7 shows the accuracy improvement by *AL Adaptation* over 20 queries. All 7 users need to ask a few queries for office class. Except for user 1 for which only little improvement is achieved, the queries significantly improve the recognition performance. Meanwhile, for user 1, queries for street class are executed, which increases the performance significantly. It seems like office context contains a heterogeneous mixture of sound events and some of them are highly user-dependent. Remember that for user 4, the *FS-Supervised* yields a very bad performance (only 22% accuracy) as can be seen in Table 5 and Figure 3. However, after 2 queries of office class and 4 queries of street class, the performance is dramatically increased. There are several context classes that the Freesound-based model characterizes well in user-specific data. Thus the *AL Adaptation* does not need to query any labels from those classes. For example, for user 4, restaurant context and tram context are not enquired for labels (i.e., (0,0) in Table 7). To enforce this analysis, we show the confusion matrix of *FS-Supervised* and *AL Adaptation* for user 4 in Figure 5 . As can be seen, the *FS-Supervised* can recognize well restaurant and tram classes, and it confuses office with toilet, and street with tram. That is why the *AL Adaptation* needs to ask labels for office and street classes, but not for restaurant and tram. Therefore, it can be emphasized again that Freesound contains diverse, useful acoustic data that can be used to recognize user context.

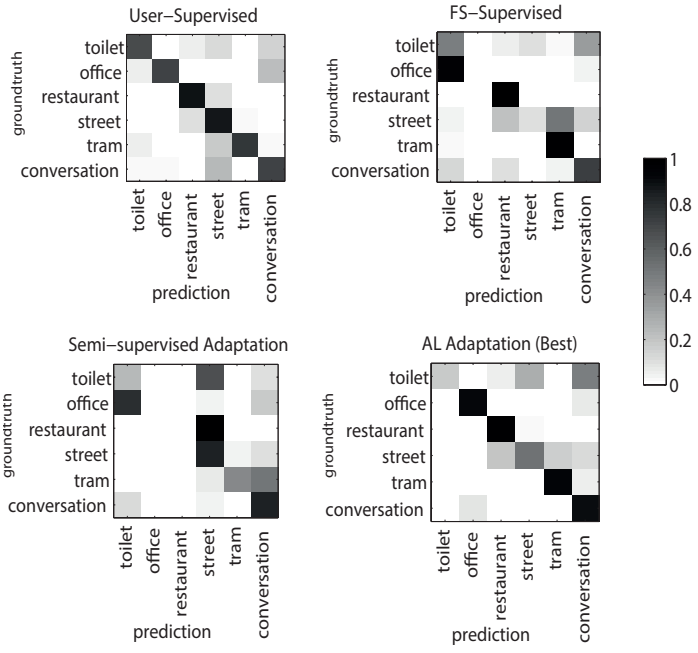

Figure 5: Confusion matrix tables of learning approaches for user 4

	User 1	User 2	User 3	User 4	User 5	User 6	User 7
office	(10, 20%)	(8, 48%)	(5, 27%)	(2, 56%)	(5, 47%)	(2, 1%)	(1, 38%)
restaurant	–	(0, 0)	(0, 0)	(0, 0)	(1, -3%)	(1, 2%)	(0, 0)
street	–	(2, -1%)	(5, -12%)	(4, 30%)	(9, 3%)	(3, 16%)	(7, 0.5%)
tram	(0, 0)	–	(0, 0)	(0, 0)	(0, 0)	(4, 7%)	(2, 1%)
train	(6, 8%)	–	–	–	(0, 0)	(6, -0.3%)	(5, 2%)
car	–	–	–	–	–	(0, 0)	(0, 0)
bus	–	–	–	–	–	–	(1, 15%)
toilet	–	(0, 0)	–	(2, 3%)	(0, 0)	(3, 3%)	(0, 0)
conversation	(4, -11%)	(10, -5%)	(10, -0.9%)	(12, -21%)	(5, -2%)	(1, 0.3%)	(4, 3%)

Table 7: The accuracy improvement by Active Learning from FS over 20 label queries. The first number in parenthesis is the total number of queries requested for the corresponding context class and the second number denotes the accumulated accuracy improvement A. The (0,0) denotes that AL from FS does not acquire a label for any instances from the corresponding context class. A dash – denotes that the user does not have the context class in his ADL.

Discussion.

Our proposed *Semi-supervised Adaptation* can improve the performance up to 21% from the *FS- Supervised*. However, *Semi-supervised Adaptation* can be very greedy in leveraging the unlabeled data, thus it can mislead and decrease the performance. One solution for this is to lower the emphasis of the unlabeled data by adding a positive weight $\lambda \leq 1$ to the semi-supervised log likelihood [14]. More importantly, it is still far from reaching the performance of the *User- Supervised* baseline approach. While crowd-sourced data indeed helps context recognition, it may not cover exact user-specific context characteristics and all user-specific situations. Thus, both data sources are probably too dissimilar to profit from the unlabeled data from the user data source. *AL Adaptation* can enquire labels for training instances of only user-specific context scenes that the Freesound data can not represent well, and the method can quickly reach the performance of *User-Supervised* approach after only a few queries. Interestingly, the *AL Adaptation* can outperform the baseline *User-Supervised* because the *AL Adaptation* can take the advantages of Freesound diversity and variability. Furthermore, *AL Adaptation* is much better than *AL from User* in terms of the number of queries needed to reach the performance of *User-Supervised*, thus the *AL Adaptation* can leverage well both sources of data: Freesound and user- centric data to get a very good accuracy, but reduce significantly user effort to annotate data.

From our evaluations, we state the lessons learned from this work. Context recognition systems which use traditional supervised learning on user training data perform the best, but it requires a huge effort to label a sufficient amount of user training data. The crowd-sourced repositories provide free labeled training data, however the performance is till far from reaching the performance of supervised learning on user-specific data. The semi-supervised learning to combine labeled crowd-sourced data and unlabeled user data is one of the cheapest ways to adapt the system without asking any effort to label user data. It can improve the performance but can not reach the best result. With a few label queries only from user training data, the active learning based on crowd-sourced data can perform a significant improvement and reach the supervised performance with only 0.7% of user data. Hereby the recognition system built on crowd-sourced data can flexibly learn a new class by first extracting training samples for that class from crowd-sourced repositories, and successively improve its performance with another few user queries by using active learning. Therefore, our best recommendation for personal context recognition is to use active learning with crowd sourced data for optimal and efficient learning. Our proposed adaptation techniques which leverages crowd-sourced data can also open a new chance to develop a scalable and open-ended context recognition system.

8. CONCLUSION AND FUTURE WORK

In this paper, we conducted experiments that combine and leverage complementary properties of two sources of data: the crowd-sourced labeled audio dataset and the user-centric audio recorded from mobile phones, to recognize user daily context. We investigated semi-supervised learning and active learning to adapt a generic model built from crowd-sourced data to a personalize context model. The semi-supervised learning can improve the recognition accuracy up to 21%, thus, the semi-supervised learning can be used to adapt user-centric data from the crowd-sourced data to build a better context recognition without asking labeling on user data. However, it still underperforms significantly the user supervised model build on user-centric data. The active learning approach that is based on the crowd-sourced labeled data can reach the performance of the supervised model quickly with surprisingly only a few label queries for the user data (in average 5 queries corresponding to 0.7% of the user training data). Furthermore, the active learning approach can even outperform the user supervised model as it leverages diversity and variability of existing crowd-sourced data. Our recommendation for personal context recognition is to use active learning with crowd sourced data for optimal and efficient learning in terms of accuracy and labeling effort. The rich availability of crowd-sourced data in terms of number of classes also open new opportunities to develop open-ended, scalable activity recognition system. In future work, we plan to analyze the influence of unlabeled data in the semi-supervised learning approach by varying their emphasis. We also plan to try stream-based active learning schemes to support interactive online strategy to get annotation from user on mobile phones. Furthermore, we also need to test the algorithms with more number of users and more context classes.

9. ACKNOWLEDGMENTS

The authors would like to thank Mirco Rossi (ETH Zurich) for useful comments. This work has been supported by the Swiss Hasler Foundation project Smart-DAYS.

10. REFERENCES

[1] V. Akkermans, F. Font, J. Funollet, B. De Jong, G. Roma, S. Togias, and X. Serra. Freesound 2: An improved platform for sharing audio clips. In *International Society for Music Information Retrieval Conference (ISMIR 2011), Late-breaking Demo Session*, Miami, Florida, USA, 2011.

[2] J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J Acoust Soc Am*, 122(2):881, 2007.

[3] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive Computing: Proc. of the 2nd Int'l Conference*, 2004.

[4] U. Blanke and B. Schiele. Daily routine recognition through activity spotting. In *Location and Context Awareness*, volume 5561 of *Lecture Notes in Computer Science*, pages 192–206. Springer Berlin Heidelberg, 2009.

[5] R. S. Bucks, D. L. Ashworth, G. K. Wilcock, and K. Siegfried. Assessment of activities of daily living in dementia: development of the bristol activities of daily living scale. *Age and ageing*, 25, Mar. 1996.

[6] S. Chu, S. Narayanan, and C.-C. J. Kuo. Environmental sound recognition with time-frequency audio features. *Trans. Audio, Speech and Lang. Proc.*, 17(6):1142–1158, Aug. 2009.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical society, Series B*, 39(1):1–38, 1977.

[8] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[9] S. Katz. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*, 31(12), Dec. 1983.

[10] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 48, Sept. 2010.

[11] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 165–178, 2009.

[12] T. H. Monk, E. Frank, J. M. Potts, and D. J. Kupfer. A simple way to measure daily lifestyle regularity. In *European Sleep Research Society*, 2002.

[13] L.-V. Nguyen-Dinh, M. Rossi, U. Blanke, and G. Tröster. Combining crowd-generated media and personal data: semi-supervised learning for context recognition. In *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*, PDM '13, 2013.

[14] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.*, 39(2-3), May 2000.

[15] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. Computational auditory scene recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, may 2002.

[16] M. Perkowitz, M. Philipose, K. Fishkin, and D. J. Patterson. Mining models of human activities from the web. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, 2004.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. In *Digital Signal Processing*, page 2000, 2000.

[18] R. Rose and D. Reynolds. Text independent speaker identification using automatic acoustic segmentation. In *International Conference on Acoustics, Speech, and Signal Processing, 1990*, pages 293–296 vol.1, 1990.

[19] M. Rossi, O. Amft, and G. Tröster. Recognizing daily life context using web-collected audio data. In *the 16th IEEE International Symposium on Wearable Computers, 2012*, June 2012.

[20] B. Settles. Active learning literature survey. Computer sciences technical report, University of Wisconsin–Madison, 2010.

[21] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[22] M. Stäger, N. Perera, and T. V. Büren. Soundbutton: Design of a low power wearable audio classification system. In *the 7th International Symposium on Wearable Computers*, 2003.

[23] M. Stikic, K. Van Laerhoven, and B. Schiele. Exploring semi-supervised and active learning for activity recognition. In *the 12th IEEE International Symposium on Wearable Computers*, pages 81–88, 2008.

[24] L. Van Der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 2008.

[25] Z. Zhang and B. Schuller. Semi-supervised learning helps in sound event classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[26] X. Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.