Combining Crowd-Generated Media and Personal Data: Semi-Supervised Learning for Context Recognition

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

ABSTRACT

The growing ubiquity of sensors in mobile phones has opened many opportunities for personal daily activity sensing. Most context recognition systems require a cumbersome preparation by collecting and manually annotating training examples. Recently, mining online crowd-generated repositories for free annotated training data has been proposed to build context models. A crowd-generated dataset can capture a large variety both in terms of class number and in intra-class diversity, but may not cover all user-specific contexts. Thus, performance is often significantly worse than that of user-centric training.

In this work, we exploit for the first time the combination of both crowd-generated audio dataset available in the web and unlabeled audio data obtained from users' mobile phones. We use a semi-supervised Gaussian mixture model to combine labeled data from the crowd-generated database and unlabeled personal recording data. Hereby we refine generic knowledge with data from the user to train a personalized model. This technique has been tested on 7 users on mobile phones with a total data of 14 days and up to 9 context classes. Preliminary results show that a semi-supervised model can improve the recognition accuracy up to 21%.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*

Keywords

Context recognition; mobile phone; semi-supervised learning; activities of daily living; crowd-generated media

1. INTRODUCTION

As mobile phones become pervasive and contain a rich set of embedded sensors, they enable new opportunities for capturing personal user context (e.g., behavior, location). Hence, many applications across a wide range of domains such as healthcare, social networks, ecommerce can make use of mobile phones to understand users' contexts and provide user-needed services [3]. Recognizing complex activities of daily living (ADL) (e.g., working in the office, having a conversation) is important in healthcare domain to monitor someone's wellbeing [2].

Most existing ADL recognition systems require a time-consuming preparation in which training data (i.e., sensor readings) is manually collected, and labeled. Recently, the idea of mining online crowd-generated sound repositories such as Freesound [1] has been exploited to extract free, extendable annotated training data for audio-based ADL recognition systems [6]. Crowd-generated sound repositories contain a large variability because sounds of the same class can be recorded from everywhere by different devices. Moreover, contributions are subjective to contributors' interpretation and preferences. Hence, a context recognition system, which is based on the crowd-generated audio dataset is only sub-optimal: it can indeed capture a variety of context classes and context variability, but still not cover exact user-specific context characteristics. As a result, the performance is often significantly worse than that of user-centric training system. Table 1 shows the trade-off between crowd-generated audio data and user-centric data recorded from user's mobile phone.

We propose to use a semi-supervised learning scheme that combines labeled crowd-generated audio data with unlabeled personal audio data recorded from user's mobile phone to recognize user daily life contexts. We conduct experiments on 7 users with a total data amount of 14 days. We compare our system to two standard supervised systems using user-trained models and crowd-generated models respectively. We show that model-refinement of the crowd-based model using unlabeled user data can significantly improve context recognition.

	Crowd-generated 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-generated data and user-centric data

[1] www.freesound.org

PDM'13, October 22, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2509352.2509396.

2. RELATED WORK

Environmental sound has been used as a rich source of information to infer person's activities and locations [1, 8, 6]. While training data is essential for recognition system, it is extremely difficult and time-consuming to obtain sufficient amounts of data with annotations that represent complex daily life situations. Consequently, most of the previous work are limited to small datasets of daily life contexts that are manually collected and labeled under controlled conditions [1, 8].

The ideas of mining the online multimedia repositories for relevant training data for activity recognition systems has been used recently by researchers [5, 6] to reduce the effort to collect and label training data as well as increase the number of available context classes. Perkowitz et al. [5] presented the web-based activity discovery using text. Rossi et al. [6] proposed to use the online crowd-generated Freesound database to obtain a heterogeneous and diverse training data to train sound models which will be exploited on mobile phones to recognize ADL.

Semi-supervised learning is a technique in machine learning that can use both labeled and unlabeled data to train a recognition system. Semi-supervised learning has been widely used in many fields such as text mining [4], sound event classification [9]. There are many variations of semi-supervised learning algorithms and a detailed survey can be found in [10].

According to the best of our knowledge, there is no previous work that investigated the use of semi-supervised learning to combine of labeled crowd-generated audio data and unlabeled user personal data which captures user's environment surroundings to improve the recognition performance. In the work by Rossi et al. [6], Freesound has been used for context recognition with supervised learning, however, they do not consider user adaptation to improve the performance. Zhang et al. [9] used semi-supervised learning to improve sound event classification. However, in their work, they used labeled and unlabeled data in the same database and did not work with personalized user context.

3. CONTEXT RECOGNITION SYSTEM

Our goal is to leverage both the abundance of labeled data from the web and of unlabeled user-centric data from mobile phones. Figure 1 shows an overview of our sound-based context recognition system. In the training phase, we collect training audio data from Freesound and user's mobile phone, and train a context recognition model with semi-supervised learning. In the recognition phase, the context recognition model will be used to infer user context from data recorded on user's mobile phone. In the following, we describe each component in our proposed system.

Freesound Repository. We consider in particular the Freesound database as a online source 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.

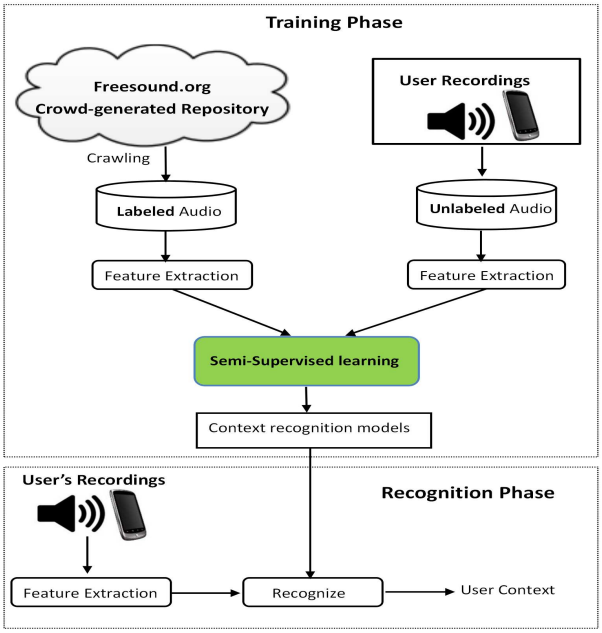

Figure 1: Architecture of our sound-based context recognition that combines Freesound data and personal audio data with semi-supervised learning

Crawling labeled audio from Freesound. In our system, we focus on normal ADL such as *dining in a restaurant* or *using toilet* as suggested in [2]. 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. We retrieve only sound clips with the highest average rating (i.e., high quality) given to the sounds provided by the Freesound community. 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. However, the techniques to filter audio clips automatically proposed by Rossi et al. [6] can be used.

User Recordings. We record continuously audio data from users' smartphones with a sampling frequency of 16 kHz and bit depth of 16 bits/sample.

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

Semi-supervised Learning. We use the semi-supervised Gaussian Mixture Model algorithm discussed in Section 4 to combine labeled and unlabeled data. The GMM models are then used to classify audio data recorded from user's mobile phone.

Classification 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 each instance is assigned to that instance. At the high level, a decision is made on the longer segments (2 seconds) by taking a class with the highest frequency in the segments as a label.

4. SEMI-SUPERVISED LEARNING

In this section, we firstly present the notation and probabilistic framework of a standard Gaussian mixture model (GMM) used in the paper. Then the semi-supervised GMM algorithm to combine labeled data (i.e., Freesound data) and unlabeled data (i.e., user-centric data) is introduced. We use the similar semi-supervised learning approach with multiple mixture components per context class as proposed in [4] for text classification.

Gaussian Mixture Model.

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., $\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log p(\mathcal{D}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \mathcal{L}$.

Semi-supervised GMM.

The goal is to find semi-supervised parameters Θ that maximize $\mathcal{L}$ to fit the labeled and unlabeled observations. The Expectation-Maximization (EM) approach is a standard procedure to find the locally optimal $\widehat{\Theta}$. 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

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-generated dataset we make use of the *Freesound* repository.

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. 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 them to label fine-grained sound events, but normal ADL. In our work, we want to support user-dependent context recognition. Therefore, users can annotate different set of context classes, subjective to their daily situations. Table 2 shows the list of classes provided by 7 participants. For each recording day, at least 9 hours of audio data were obtained for each user. In total, about 130 hours of audio data were collected from mobile phones for the study.

	Context Classes	Number of Classes
User 1	office, tram, train, conversation	4
User 2	toilet, office, restaurant, street, conversation	5
User 3	office, restaurant, street, tram, conversation	5
User 4	toilet, office, restaurant, street, tram, conversation	6
User 5	toilet, office, restaurant, street, tram, train, conversation	7
User 6	toilet, office, restaurant, street, tram, train, car, conversation	8
User 7	toilet, office, restaurant, street, tram, train, car, bus, conversation	9

Table 2: User-dependent context classes

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, in the car class, the sound clips can be recorded in different weather situations (rain, snow). After manually filtering, we have 163 audio clips (143 minutes) for 9 context class to train sound models. This data from Freesound is denoted as FS.

6. EVALUATION

We compare our proposed semi-supervised learning with two baseline approaches: (1) a supervised GMM using Freesound (FS) data only and a supervised GMM with user training data only. The experiments were 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%

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

of user data for each class) is used either without its labels ($F1_U$) or with these labels ($F1_L$). Specifically, the semi-supervised learning will train on the combination of FS and $F1_U$. The supervised user-trained approach will train on $F1_L$ data and the supervised Freesound approach will train on FS data. The second fold (F2, another 50% of user data for each class) is used for testing for all three approaches.

Results: The results are given in Table 4. As expected, the supervised user-trained learning gives the best result. Supervised training on user data captures user-specific environments accurately in the model and thus, recognizes well user context in daily routines (test and training data tend to be similar for each single user). The performance of the supervised Freesound model drops significantly since the crowd-generated data hardly covers all user-specific surroundings. The accuracy of the supervised Freesound model is similar to that reported in the work by Rossi [6] which also used Freesound data to recognize users' contexts. However, the results also show that the semi-supervised learning significantly improves the performance of context recognition compared to the Freesound supervised approach (from 3% up to 21%) for six users. However, for user 3, the semi-supervised learning actually hurts the performance. It means the contribution from unlabeled user data in the semi-supervised learning makes the model more uncertain. 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 [4]. The result from user 4 shows that even the semi-supervised learning can increase the accuracy by 3%, the accuracies of the Freesound supervised and the semi-supervised are much lower than that of user-train supervised approach. Here the Freesound data does not generalize sufficiently the data recording from that user-specific environments. In the future, we plan to use the active learning [7] to have labels for clusters which can not be represented by crowd-generated data.

7. CONCLUSION AND FUTURE WORK

In this paper, we conducted experiments that combine the crowd-generated crowd-labeled audio dataset with the user-centric audio recorded from mobile phones with semi-supervised learning to recognize user daily context. We then compared the proposed semi-supervised learning with two baseline supervised approaches which train only either on crowd-generated audio dataset or on user-centric audio data. The preliminary work showed promising results. The semi-supervised learning can improve the recognition accuracy up to 21%. Therefore, the semi-supervised learning can be used to adapt user-centric data from the crowd-generated data to build a better context recognition without asking labeling on user data. 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 apply active learning to query the label for instances which are not represented by crowd-generated data.

	User-trained, supervised ($F1_L$)	Freesound, supervised (FS)	Semi-Supervised (FS + $F1_U$)
User 1	0.94	0.8	0.86
User 2	0.9	0.5	0.65
User 3	0.72	0.58	0.43
User 4	0.72	0.22	0.25
User 5	0.82	0.35	0.5
User 6	0.85	0.54	0.61
User 7	0.83	0.26	0.47

Table 4: Accuracy for the context recognition

8. ACKNOWLEDGMENTS

This work has been supported by the Swiss Hasler Foundation project Smart-DAYS.

9. REFERENCES

[1] A. Eronen, V. Peltonen, J. Tuomi, and et al. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[2] 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.

[3] 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.

[4] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. In *Machine Learning*, 1999.

[5] 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*, 2004.

[6] 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.

[7] B. Settles. Active learning literature survey. Technical report, 2010.

[8] 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.

[9] 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.

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