Recognizing Daily Life Context using Web-Collected Audio Data

Mirco Rossi, Gerhard Tröster
Wearable Computing Lab., ETH Zurich
{mrossi,troester}@ife.ee.ethz.ch

Oliver Amft
ACTLab, Signal Processing Systems, TU Eindhoven
amft@tue.nl

Abstract

This work presents an approach to model daily life contexts from web-collected audio data. Being available in vast quantities from many different sources, audio data from the web provides heterogeneous training data to construct recognition systems. Crowd-sourced textual descriptions (tags) related to individual sound samples were used in a configurable recognition system to model 23 sound context categories. We analysed our approach using different outlier filtering techniques with dedicated recordings of all 23 categories and in a study with 230 hours of full-day recordings of 10 participants using smart phones. Depending on the outlier technique, our system achieved recognition accuracies between 51% and 80%.

1 Introduction

Sound is a rich source of information that can be used to infer person's activities and environments [3, 7]. Moreover, microphones are cheap, available in every phone, and unlike other sensors (e.g. cameras) robust to their positioning. However, identifying complex daily life situations, such as office work or commuting is challenging due to variations in the acoustic context and limited availability of example data. Consequently, most existing recognition systems suffer from incomplete modelling of the target activities and environmental contexts. While appropriate training data is essential for recognition systems, it is laborious to obtain sufficient amounts with annotations that represent these daily life situations.

In this work we investigate an approach to source sound data from the web and derive acoustic pattern models of daily life context. Our approach is inspired by the idea of crowd-sourcing: web audio data is generated by many users. It is heterogeneous, available in large quantities, and provides annotations, e.g. in the form of 'tags'. However, the web is not a source of perfectly labeled training data. Users generate web audio annotations by following personal interpretation and preferences. In some cases, even erroneous annotations could occur. Thus, web search results include also audio samples with unexpected acoustic content. We refer to these audio samples as outliers. Including outliers in training data affects the quality of the acoustic pattern model and the recognition performance.

We present an approach and system architecture to use data subsets from the open web database Freesound (http://www.freesound.org) - an audio database consisting of more than 120'000 audio samples freely annotated with tags and uploaded by around 6'000 contributors. To investigate our approach we used an example configuration of 23 sound context categories to derive a recognition system. We demonstrate that the web data can be used to discriminate daily life situations recorded from microphones of commonly available smart phones. We evaluated the system with dedicated recordings of all the 23 categories, and in a study with full-day recordings of 10 participants. We furthermore investigate different automatic outlier filtering strategies and compare them to a manually derived baseline performance.

2 Related Work

A common approach to build a recognition system is to manually collect and label training data. Most auditory scene recognition systems used this approach in the past decade. For example, Eronen et. al. [3] focused on recognition systems for environmental sounds, such as "restaurant" or "street", and Stäger et al. [7] recognized a set of activities of daily living (ADL) based on sound data. Wearable systems for sound recognition have been proposed as dedicated hardware [7] and more recently using smart phones-based solutions, e.g. Lu et al. [5]. However, many activity and environmental sound recognition solutions are yet constraint to small sets of sound contexts and well-defined recording locations. In naturalistic, real-life situations a recognition would need to cope with highly heterogeneous sounds.

The idea of mining the web for relevant training data has been used for different modalities. Perkowitz et. al. [6] presented the first method for web-based activity discovery using text. Bergamo et. al. [1] used web images to learn visual attributes of objects. In a similar direction, Checick et. al. [2] used the Freesound database to generate a content-based audio information retrieval system. So far, to our

1550-4816/12 $26.00 © 2012 IEEE
DOI 10.1109/ISWC.2012.12

knowledge Freesound had not been considered as mining source for recognising ADL-related context.

3 Concept of Web-Based Sound Modelling

Our approach is based on *Context category descriptions*, which could be provided by a user. Context category descriptions are used for *Collecting audio data* from the web. Subsequent steps of our architecture include *Extracting audio features*, *Filtering outliers*, and *Modelling context categories*. Filtering the collected audio samples for outliers is essential to derive a robust recognition system. This section details our web-based sound modelling and outlier filtering as shown in Fig. 1.

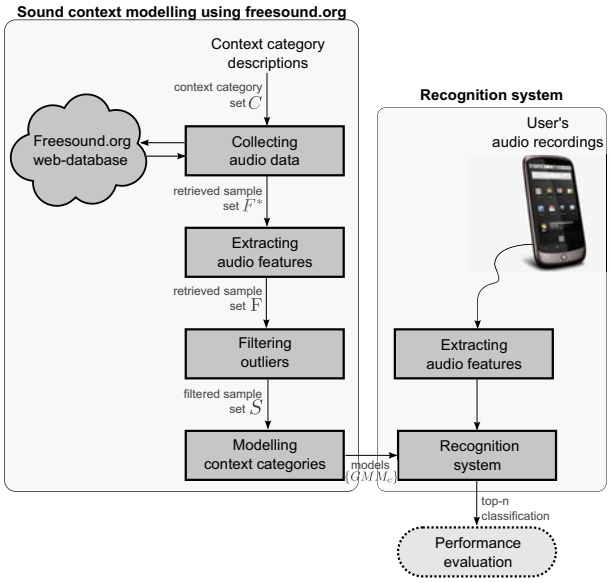

Figure 1. Overall architecture of our sound context recognition based on web-collected sound data using Freesound.

Context category descriptions provide a textual description of the set of context categories C. Each category $c_i \in C$ is described by one or more descriptive terms, which are subsequently used to retrieve sound samples from the web. In this work, we used an example configuration of 23 context categories, listed in Tab. 1. We compiled this set of categories such that a wide range of complex daily life situations were covered, including categories characterizing locations, sounds of objects, persons and animals.

Collecting audio data: According to the context category descriptions the sound samples were retrieved from the Freesound database. Sound samples having a set of tags that matches all terms in a category description of our example configuration were downloaded and labeled with the corresponding category. All the retrieved audio samples were transcoded to WAV format with a sampling frequency of $f_s = 16\,\text{kHz}$ and bit depth of $b_s = 16\,\text{bits/sa}$.

Extracting audio features: Audio features were extracted from the retrieved audio samples. We used the mel-frequency cepstral coefficients (MFCC), the most widely used audio features in audio classification. These features showed good recognition results for environmental sounds [3]. The feature vectors $\{s_f\}$ of an audio sample s were generated by extracting MFCC's (12 coefficients) on a sliding window of 32 ms length, with an overlap of 16 ms between consecutive windows. The same method was used to extract audio features from the smart phone.

Context categories
objects: *brushing teeth, bus, car, chair, coffee machine, dishwasher, phone ring, raining, shaver, sink, toilet flush, vacuum cleaner, washing machine*
locations: *beach, crowd football, forest, office, restaurant, street, railway station*
animals and persons: *bird, dog, speech*

Table 1. Context category set C. In this work, an example of 23 context categories were used. The category names are directly used as the descriptive terms.

Filtering outliers: When collecting audio from the web, outliers with regard to the targeted context category can be frequently expected. Training models with data containing outliers negatively effects the recognition performance of the system. Thus, the goal of filtering is to remove outliers from the correctly labeled data, before models are trained. In our approach to remove outliers, we assumed that correctly labeled samples of a category are more likely to sound similar than outliers. To measure sound similarity of two sound samples, $s^{(1)}$ and $s^{(2)}$, we used the Mahalanobis distance measure: $D(s^{(1)}, s^{(2)}) = (\boldsymbol{\mu_{s^{(1)}}} - \boldsymbol{\mu_{s^{(2)}}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu_{s^{(1)}}} - \boldsymbol{\mu_{s^{(2)}}})$, where $\boldsymbol{\mu_{s^{(1)}}}$ and $\boldsymbol{\mu_{s^{(2)}}}$ are the mean feature vectors of the audio samples $s^{(1)}$ and $s^{(2)}$, and $\boldsymbol{\Sigma}$ is the covariance matrix of the features across all samples. This distance measure has low computational costs and showed competitive results compared to more complex modeling schemes [4]. Based on $D(s^{(1)}, s^{(2)})$, we propose two outlier filtering methods using semi-supervised and unsupervised concepts. Both methods use an approach presented in Algorithm 1, where a filtered set of audio sample set S is created from the retrieved audio sample set F. An initial set of correct samples S_{init} is required. For **semi-supervised filtering** the initial set must be provided by the user, who selects for each category k correctly labeled samples in F. For **unsupervised filtering**, the initial set is formed by selecting for each category k samples with the smallest inter-category distance. In our evaluation we included the following two methods and used them as comparative baselines: **no filtering** in which we used all samples in F for training ($S = F$), and **manual filtering** in which we manually filtered outliers ($S = S_{hand}$).

Modelling context categories: The extracted features of the sample set S were used to train models of the 23 con-

Algorithm 1 Outlier Filtering. $F^{cat(s)}$ is the set of all samples in F belonging to the same category as sample s.

inputs: S_{init}, F
$S = S_{init}$
repeat
  for s in S **do**
    $f = \text{argmin}_{i \in F^{cat(s)}} D(i, s)$
    if $\text{argmin}_{i \in S} D(i, f)$ is s **then**
      add f to the set S
    end if
    remove f from the set F
  end for
until F is empty
output: S

text categories in our example configuration. We separately modelled the feature space of each category c with a Gaussian Mixture Model GMM_c. The number mixture components was fixed after a small-scale experiment to 16.

Recognition system: The web-trained GMM models were used to classify audio data recorded from the smart phone. The probability that an audio test sequence t belongs to the category c is calculated by: $p(t|GMM_c) = \prod_f p(t_f|GMM_c)$, where t_f is a feature vector of the test sequence t. In our evaluation we varied the length of the test sequence t between 1 and 30 s. As our approach produces a term-based description of the context, it is conceivable that several sound context could be used simultaneously to describe the situation. Thus, the recognition system generates a top-n classification by selecting n categories with the highest probabilities. Top-n classifiers with $n \in \{1, 2, 3\}$ has been evaluated.

Performance evaluation: The system's recognition performance was measured using the normalized accuracy (mean over all class-relative accuracies). We accounted the classification as correct, if the annotated context category was within the top-n categories.

4 Results

In total 4678 audio samples (114 hours of audio data) were retrieved from the Freesound database for the 23 context categories, with a mean distribution of 203 samples per category. Manually filtering outliers from the samples showed that 38 % of the samples were outliers. The system's performance was analyzed within two evaluations. Firstly, we analyzed the performance of all category models using dedicated sound data recorded for each context category. Subsequently, we evaluated the performance and system operation in a study using daily life audio recordings from smart phones of 10 participants.

4.1 Evaluation by Dedicated Recordings

The models were tested with dedicated sound data recorded for each context category using an Android phone (Google Nexus One). Sound samples of at least four different entities per sound category (e.g. four different dishwashers) had been recorded using the phone's integrated microphone. For each class we recorded 6 min of audio data. This test allowed us to assess the system's performance for the complete set of 23 context categories using self-recorded data and compare the benefit of the different outlier filtering methods. Moreover, the goal of this test was to confirm that the microphone and electronics of a commonly available smart phone suit for the recognition of daily life contexts.

Fig. 2 shows the recognition accuracy of the top-1 classifier for the different outlier filtering methods. The length of the test sequence has been varied between 1 and 30 s. As expected, increasing the length of the test sequence improved the overall recognition accuracy. Models trained with no outlier filtering performed at lowest accuracy (38 % with a 30 s test sequence). Using unsupervised and supervised filtering the accuracy increased to 46 % and 53 %, respectively. The best performance was reached using manual filtering (57 %). These performance results are comparable to studies considering similar large number of sound categories, e.g. the work of Eronen et al. [3]. In their work, the authors used an audio-based recognition system for 24 environmental categories and obtained a recognition performance of 58% using MFCC features. Their training dataset was compiled manually from dedicated recordings and included few locations only. In contrast, the web-based audio data consists of diverse field recordings aquired using different recording systems.

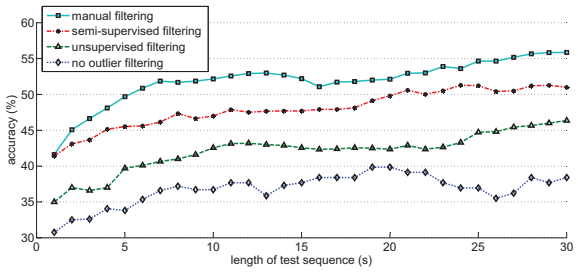

Figure 2. Performance analysis of the top-1 classification using dedicated sound recordings. The level of a random guess is at 4.35 %.

Fig. 3 shows the confusion matrix of the top-1 classifier trained on data filtered with the semi-supervised method. The category *beach*, *railway station*, and *speech* showed the best class-relative accuracies (100 %). In contrast, the category *vacuum cleaner* was not recognised by the system. For the *vaccum cleaner*, semi-supervised filtering failed to remove outliers. Some confusions could be explained by the similar context in which the sounds were recorded, e.g. *restaurants* was confused with *speech* and *bus*. All the three categories included recordings of talking people. *Brushing teeth* was confused with *shaver*, since some electronic tooth

brushers and shaving machines produced similar sounds. The top-3 classifier resulted in improved accuracies: 51 % without any outlier filtering, 69 % with unsupervised filtering, 79 % with semi-supervised filtering, and 80 % with manual filtering.

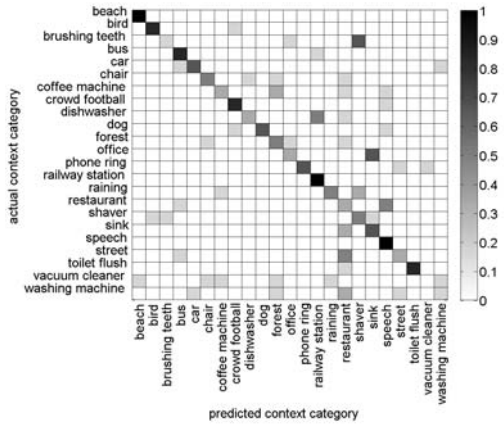

Figure 3. Confusion matrix of the top-1 classifier using semi-supervised outlier filtering. The test sequence length has been set to 30 s.

4.2 Evaluation of Daily Life Study

To investigate the web-based recognition approach in real-life data, we performed a study using smart phones for continuous environmental sound recording of 10 participants aged between 24 and 40 years. Participants were asked to record two full working days in one week. Recordings were done using the same phone model as in Sec. 4.1 with a headset microphone. During the recordings participants attached the headset to the upper body clothing between waist and collar. The recordings had been performed using our specialized Android application "AudioLogger". The application allowed us to store continuous audio data on the SD card of the smart phone. In addition, the application provided an annotation tool in which the user could select current contexts from a selection list providing all context categories shown in Tab. 1. For each recording day at least 8 hours of audio data were obtained. In total, more than 230 hours of audio data were collected in this study.

During the study, participants used only a subset of the annotations provided. Thus, we performed an analysis for the annotations used by the participants with the following context categories: *bus*, *car*, *office*, *railway station*, *restaurant*, *speech*, and *street*. However, the recognition system was maintained as previously trained for all 23 context categories. The system performance was evaluated for the top-n classifiers with $n \in \{1, 2, 3\}$. The results are showed in Fig. 4. When considering $n = 3$, a performance of 85% with manual filtering , 80% with semi-supervised filtering, 71% with unsupervised filtering, and 51% with no filtering was obtained.

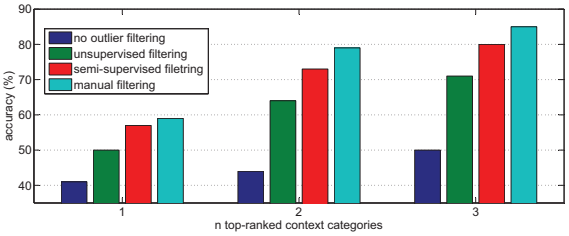

Figure 4. Recognition performance of the web-based recognition in daily life recordings. The test sequence length has been set to 30 s.

5 Conclusion

Using web-collected audio data to construct a context recognition system showed to be a promising approach. It provides opportunities to reduce the process of manually collecting training data as it is available in large quantities from the web. While our test showed that the web-collected sound samples were highly diverse, the proposed outlier filtering methods yielded a recognition accuracy increase of up to 18 %. Practical recognition rates for high-level contexts between 51% and 80% could be achieved. We expect that the presented recognition system could be implemented on a cloud server to operate in real-time. Therefore, instantaneous sound-based context recognition could be realized. Additionally, based on such a mobile system, the idea of crowd-sourcing could be extended giving the users an opportunity to share personal auditory scenes directly using their smart phone.

References

[1] L. Bergamo, Alessandro Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[2] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon. Large-scale content-based audio retrieval from text queries. In *Proceeding of the ACM international conference on Multimedia information retrieval*, pages 105–112, 2008.

[3] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Audio, Speech, and Language Processing*, 14(1):321–329, 2006.

[4] T. Heln, Marko Virtanen. Audio query by example using similarity measures between probability density functions of features. *EURASIP Journal on Audio, Speech, and Music Processing*, pages 1–12, 2010.

[5] H. Lu, W. Pan, N. Lane, T. Choudhury, and A. Campbell. SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones. In *Proc. of ACM Conference on Mobile Systems, Applications, and Services*, 2009.

[6] M. Perkowitz, M. Philipose, K. Fishkin, and D. Patterson. Mining models of human activities from the web. In *Proceedings of the international conference on World Wide Web*, pages 573–582, 2004.

[7] M. Stager, P. Lukowicz, N. Perera, G. Troester, and T. Starner. SoundButton: Design of a Low Power Wearable Audio Classi cation System. *Proc. of IEEE International Symposium on Wearable Computers*, 2003.