

WHERE AM I? SCENE RECOGNITION FOR MOBILE ROBOTS USING AUDIO FEATURES

Selina Chu[†], Shrikanth Narayanan^{†}, C.-C. Jay Kuo^{*†}, and Maja J. Matarić[†]*

Department of Computer Science[†] and Department of Electrical Engineering^{*}
University of Southern California, Los Angeles, CA 90089, USA
E-mails: {selinach, shri, cckuo}@sipi.usc.edu, mataric@usc.edu

ABSTRACT

Automatic recognition of unstructured environments is an important problem for mobile robots. We focus on using audio features to recognize different auditory environments, where they are characterized by different types of sounds. The use of audio information provides a complementary means of scene recognition that can effectively augment visual information. In particular, audio can be used toward both the analysis and characterization of the environment at a higher level of abstraction. We begin our investigation of recognizing different auditory environments with the audio information. In this paper, we utilize low-level audio features from a mobile robot and investigate using high-level features based on spectral analysis for scene characterization, and a recognition system was built to discriminate between different environments based on these audio features found.

1. INTRODUCTION

The method robotic systems navigate depends on their environment. Many robot navigation systems, especially those used indoors, employ model-based vision in a well-defined and/or highly constrained environment [1]. While these approaches can perform well as long as distinct landmarks and other visual cues are available for model matching, they lose their robustness or their utility if visual indicators are compromised or totally absent. Other approaches, such as view-based methods [2], require the system to match new incoming images against learned ones. Image processing and segmentation algorithms can be computationally expensive, especially for on-device implementation. To mitigate the system's dependency on vision alone, we propose to incorporate audio information into the scene recognition process. Using audio enables the system to capture additional, semantically rich information. Audio data can be obtained at any time, and are computationally cheaper to process than visual data. Thus, the fusion of audio and visual information can be advantageous, such as in disambiguation of environment and object types.

Many robotic applications are being utilized for navigation in unstructured environments [3, 4]. There are other tasks that require knowing the environment. For

example, Yanco [5] introduced a robotic wheelchair system that switches automatically between control modes for indoor and outdoor environments. Also, laser range-finder can track people in an outdoor environment [6]. In order to use any of these capabilities, we first have to determine the current context, e.g., the location type (outdoor environment or inside an office or hallway, etc).

Characterizing the scene or environment is the first step to choosing which modality of interaction a robot should engage in. Furthermore, environments are dynamic, and the setting might change even in the same area. With the loss of certain landmarks, a vision-based robot might not be able to recover from its displacement because it is unable to determine the environment that it is in. Knowing the scene provides a coarse and efficient way to prune out irrelevant scenarios. Even with a GPS system and a well-defined map, without clear images, it is difficult to discern different characteristics of the environment.

It is relatively easy for most people to make sense of what they hear or to discriminate where they are located in the environment on the basis of sound alone. However, this is typically not the case with a robot. Surprisingly little research has been done on audio scene analysis in robots. With increasing number of robots being built for service and social settings, it is ever more important for the robots not only to identify locations, but to comprehend and characterize their auditory features.

In this paper, we investigate using audio features to recognize different unstructured auditory environments. We begin by examining audio features, such as energy and spectral moments, gathered from a mobile robot and apply those to scene characterization.

2. BASIC AUDIO FEATURE EXTRACTION

One of the major issues in building a recognition system for multimedia data is the choice of proper signal processing features that are likely to result in effective discrimination between different auditory environments. Sounds from a general ambient environment are considered neither speech nor music, but a combination of some specific audio signals that are similar to noise. While much work has concentrated on speech and music, little research has been done on actual analysis of features for classification of environmental

sounds. One of the major goals of this work is to study the effect of various features on the efficiency of an auditory environmental recognition system.

There are many features that can be used to describe audio signals. We examined the following features in our experiments: Mel-frequency cepstrum coefficient analysis (MFCC), statistical moments from the audio signal’s spectrum (i.e. spectral centroid, spectral bandwidth, spectral asymmetry, and spectral flatness), zero-crossing rate, energy range, and frequency roll-off. The use of the term *frequency roll-off* in this paper refers to the rate at which the accumulative magnitude of the frequency response is equal to that of 95% of the total magnitude. Since the energy level varies depending on the location of the source of the sound, we do not use the mean of the energy. Instead, we only use the range and standard deviation. The feature vector contained a total of 34 features, summarized in Table 1 below.

Table 1. List of features used in classification

Feature No.	Types of Features
1-12	1 st – 12 th MFCCs
13-24	Standard Deviation of 1 st – 12 th MFCCs
25	Spectral Centroid, S_c
26	Spectral Bandwidth, S_w
27	Spectral Asymmetry, S_a
28	Spectral Flatness, S_f
29	Zero-Crossing
30	Standard Deviation of Zero-Crossing
31	Energy Range, E_r
32	Standard Deviation of Energy Range
33	Frequency Roll-off
34	Standard Deviation of Roll-off

3. ENVIRONMENTAL DATA ACQUISITION

We would like to capture actual scenarios of situations where a robot might find itself, including any environmental sounds, along with additional noise generated by the robot. To simplify the problem, we restricted the number of scenes we examined and enforced each type of environmental sound not to overlap each other. The locations we considered are recorded within and around a multipurpose engineering building on the USC campus. The diverse locations that were focused include: 1) a café area, 2) hallways where research labs are housed, 3) around and inside elevator areas, 4) lobby area, and 5) along the street on the south side of the building.

We used a Pioneer DX mobile robot from ActivMedia, running *Playerjoy* and *Playerv* [7]. The robot was manually controlled using a laptop computer. To train and test our algorithm, we collected about 3 hours of audio recordings of the five aforementioned types of environmental locations. We used an Edirol USB audio interface, along with a Sennheiser microphone mounted to the chassis of the robot. Several recordings were taken at each location, each about 10-15 minutes long, taken on multiple days and at various times. This was done to introduce a variety of sounds and to prevent biases in the recordings. The robot was deliberately driven around with its sonar sensors turned on (and

sometimes off) to resemble a more realistic situation and to include noises obtained from the onboard motors and sonar. We did not use the laser and camera because they produce little, if any, noticeable sound. Recordings were manually labeled and assigned to one of the five classes listed previously to aid the experiments described below.

Our general observations about the sounds encountered at the different locations are:

- *Hallway*: mostly quiet, with occasional doors opening/closing, distant sound from the elevators, and individuals quietly talking, some footsteps.
- *Café*: many people talking, ringing of the cash registers, moving of chairs.
- *Lobby*: footsteps with echos (different from hallways due to the type of flooring), people talking, sounds of rolling dollies from deliveries being made.
- *Elevators*: bells and alerts from the elevator, footsteps, rolling of dollies on the steel frame of elevator entrance.
- *Outside*: footsteps on concrete, traffic from buses and cars, bicycles, and occasional planes and helicopters.

We chose for this study to focus on a few simple, yet robust features, which can be extracted in a straightforward manner. Features that require many thresholds were avoided.

The audio data samples collected were mono-channel, 16 bits per sample with a sampling rate of 44 kHz and of varying lengths. The input signal was down-sampled to a 22050 Hz sampling rate. Each clip was further divided into 4-second segments. Features were calculated from a 20 msec rectangular window with 10 msec overlap. Each 4 sec segment makes up an instance for training/testing. All spectra were computed with a 512-point FFT. All data were normalized to zero mean and unit variance.

4. CLASSIFICATION METHODS

To evaluate the performance of our recognition system, we examined the following three classification methods: K-Nearest Neighbor (KNN) [8], Gaussian Mixture Models (GMM) [9], and Support Vector Machine (SVM) [10]. For KNN, we used the Euclidean distance as the distance measure and the 1-nearest neighbor queries to obtain the results. As for GMM, we set the number of mixtures for both training and testing to 5. For the SVM classifiers, we used a 2 degree polynomial as its kernel with $C=10$ and $\epsilon=1e-7$, where C is the regularization parameter and ϵ controls the width of the ϵ -insensitive zone, which is used to fit the training data, affecting the number of support vectors used. Since SVM is a two-class classifier, we use the one-against-the-rest algorithm [11] for our multi-class classification in all of the experiments.

We performed leave-one-out cross-validation on the data. Although this method is computationally expensive, it has been shown to produce almost unbiased results. The

recognition accuracy using leave-one-out cross-validation was found from calculating:

$$accuracy = \frac{\#_of_correctly_classified}{Total_#_of_dataset}$$

More than half of the data collected contained sonar and motor sounds emitted by the robot. Motor noises were found to be less noticeable than those emitted by the sonars. To determine how the sonar sounds affect the classification, we manually separated the data into two sets: A) containing sonar and B) without any sonar sounds. Classifications were performed on three sets of data: set A only, set B only, and sets A and B together. The use of set-B only data would be unrealistic for mobile robots. The accuracy using all 34 features for KNN was 90.8%, 91.2%, and 89.5% for set A, B, and A&B respectively. For the rest of the paper, all experiments are performed using set A&B.

Table 2: Summary of classification accuracy

Classifiers	Features Used	Recognition Accuracy
KNN	All 34 features	89.5%
Forward FS	1-3, 5-10, 12, 13, 16, 17, 28, 31, 33	94.3%
Backward FS	1, 2, 3, 7, 8, 9, 28, 31, and 33	94.2%
GMM	All 34 features	89.5%
Forward FS	1-10, 12-16, 20-22, 25, 26, 28,31-34	93.4%
SVM	All 34 features	95.1%
Forward FS	1-3, 5-10, 13, 15, 18, 28, 31-33	96.6%

5. EXPERIMENTAL RESULTS AND DISCUSSION

One of the problems in using a large number of features is that there are many potentially irrelevant features that could negatively impact the quality of classification. In using feature selection techniques, we can choose a smaller feature set to reduce the computational cost and running time, as well as achieve an acceptable, if not higher, recognition rate. Adding more features is not always helpful; as the feature dimension increases, data points become more sparse and some features are essentially noise. This leads to the issue of selecting an optimal subset of features from a larger set of possible features that will yield the most effective subset. The optimal solution is using an exhaustive search of all the features. This requires $2^{34}-1$, or roughly 10^{10} combinations.

Instead of performing 10^{10} computations, we use a greedy search for selecting the features. There are various ways of performing feature selection, such as forward feature selection, backward selection, branch and bound, and stochastic search, each with its advantages and disadvantages. We used forward feature selection for our experiments since it is simple and straightforward. The algorithm is given as:

Initialize selected set S = empty set
Initialize unselected set $F = \{1, \dots, M\}$
Repeat:

Evaluate performance with $S \cup f_i$ for each $f_i \in F$
 $S := S \cup f_m$ and $F := F \setminus f_m$, where f_m gives maximum improvement in performance

Stop when no significant improvement in classification or features

Using this feature selection algorithm and evaluating by picking the feature f_m that yields the maximum accuracy, we found that using 16 features enabled us to achieve a recognition accuracy of 96.6% for SVM and 94.3% for KNN. It took 25 features to achieve an accuracy of 93.4% for GMM. The results and features selected are summarized in Table 2. Figure 1 below shows a plot of various recognition accuracies with the different number of features. Using only 6 features (91.1% for KNN), we were able to surpass the accuracy of using all 34 features (89.5% for KNN).

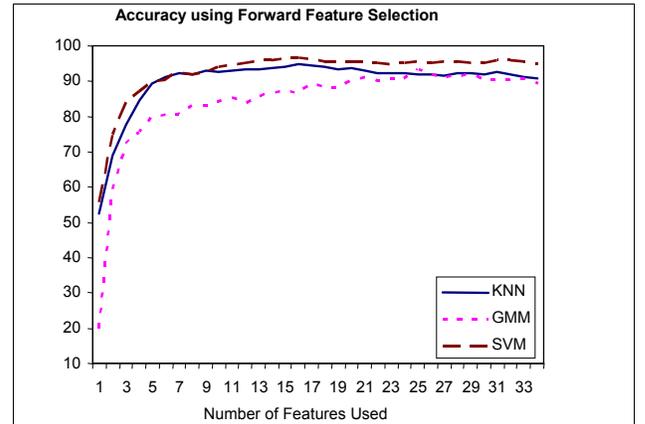


Fig. 1: The classification results with KNN, GMM, and SVM respectively. Parameters to these classifiers are given in Sec. 4

Table 3: Confusion matrix of the KNN classification using forward feature selection with 16 features, in percentage

	Street	Elevator	Café	Hallway	Lobby
Street	94.4	0	0	0	5.6
Elevator	0	90.0	1.1	7.8	1.1
Café	0	0	95.6	0	4.4
Hallway	0	0	0	100	0
Lobby	2.2	0	3.3	0	94.4

The confusion matrix in Table 3 shows the misclassified classes for the KNN classifier using 16 features. It can be seen that the worst performance was from the Elevator class and had most misclassification from Hallway. One reason for this comes from the fact that the area where the robot was driven for the Elevator class was actually part of the Hallway as well, so there was less separation between the two areas. However, Hallway gave the best performance due to its distinct characteristic of being relatively quiet most of the time. We can also observe from the same table that Lobby and Street were confused, as both contained many sharp footstep noises, but on different types of flooring. The Lobby has granite tiling, while the Street is concrete. There are footstep noises in the Hallway class as well, but the flooring for the hallways is plastic so the footsteps were less prominent than those from Lobby or Street and created less confusion. Footsteps in Café were

drowned out by other noises, such as crowds of people talking and shuffling of furnitures.

Fig. 1 shows GMM to produce the worst result when compared to KNN. One possible reason is because we refined the parameters for the case of using all features and did not re-optimize parameters when performing the forward-search feature selection experiments. A final note on the various learning algorithms studied here: unlike KNN, both GMM and SVM require a careful choice in choosing the correct parameters. There are at least two degrees of freedom for GMM and four for SVM. For GMM, the numbers of mixtures for training and another testing must be picked a priori. For SVM, one needs to decide on C , ϵ , kernel type and its bias, as explained in Section 4. In other words, even minor changes, such as number of training samples, requires fine turning of parameters. In addition, both GMM and SVM are much higher in computational complexity.

Despite the higher accuracy rate of SVM methods over KNN and GMM, SVM are very expensive to train even with five classes. The running times required for training and classification with a full feature set and no feature selection for KNN, GMM, and SVM are given as 1.1, 148.9, and 1681.8 sec, respectively. The KNN classifier works well overall, outperforming GMM and is roughly 1000 times faster than SVM.

To check for overfitting and to confirm the validity of the selected features, we performed a sensitivity analysis with respect to the forward feature selection algorithm. Since GMM and SVM require tuning of many parameters and are more complex, we restricted this experiment to just the KNN algorithm. The experiment was as follows:

- Repeat for 100 times
- Randomly pick half of the dataset
- Repeat the forward feature selection algorithm on the subset
- Record the features selected

We tallied the selected features used in each trial and picked the features that were used more than half of the time, which resulted in 11 features. With these 11 features, we performed a backward feature selection search. Similar to the idea of forward search, backward search works by using all 11 features and begins by taking out one feature at a time. Instead of picking the feature that yields the maximum recognition accuracy in the forward search, we selected the features that provided the minimum accuracy rate. The results returned 9 features, which were in turn fed back into the 1-NN classifier. We finally achieved 94.2% recognition accuracy on the entire dataset with these 9 features. As listed in Table 3, the 9 features include MFCC1-4 and 9-10, zero-crossing, std dev of zero-crossing, std dev of roll-off frequency.

6. CONCLUSIONS AND FUTURE WORK

This paper investigates techniques for developing a scene classification system using audio features. The classification system was successful in classifying the 5

classes of environment using real data obtained from a tele-operated mobile robot. We also found that using high number of features is not always beneficial to classification. In using forward feature selection, a form of greedy search, only nine of the thirty-four features were required to achieve a high recognition rate. We have also identified features that can discriminate between these 5 types of environment.

With success in using audio to discriminate between different unstructured environments, we have shown that it is feasible to build such a system. This work opens up a doorway to other challenges. Here we focused on global characterization of the environment; we need to also examine localization and the effect of various sound sources on recognition. Other issues include scaling and robustness to new environments. Our next step is to increase the number of classes of environments, as well as to investigating the combined use of audio and visual features. We then plan to implement an online version and incorporate it into a real autonomous robot for analysis.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dylan Shell for his help on the interface to the Pioneer robot and Eamonn Keogh and Wan-Chi Lee for their helpful comments and suggestions.

8. REFERENCES

- [1] DeSouza, G.N. and Kak, A.C. "Vision for mobile robot navigation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 2, pp. 237-267, 2002.
- [2] Matsumoto, Y., Inaba, M. and Inoue, H. "View-based approach to robot navigation," *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 1702-1708, 2000.
- [3] Pineau, J., Montemerlo, M., Pollack, M., Roy, N., and Thrun, S. "Towards robotic assistants in nursing homes: challenges and results," *Robotics and Autonomous Systems*, Volume 42, Issues 3-4, pages 271-281, 2003.
- [4] Thrun, S., Bennewitz, M., Burgard W., Cremers, A.B., Dellaert, F., Fox, D., Haehnel, D. Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. "MINERVA: A second generation mobile tour-guide robot," *Proc. of IEEE International Conference on Robotics and Automation*, 1999.
- [5] Yanco, H.A. "Wheesley, A Robotic Wheelchair System: Indoor Navigation and User Interface," *Lecture Notes in Artificial Intelligence: Assistive Technology and Artificial Intelligence*, Springer-Verlag, 1998
- [6] Fod, A., Howard, A., and Mataric, M. J., "Laser-Based People Tracking", *Proc. of IEEE Int. Conf. Robotics and Automation*, pages 3024-3029, 2002
- [7] <http://playerstage.sourceforge.net/>
- [8] Mitchell, T. M. "Machine Learning," Mc Graw-Hill, 1997
- [9] Moore, A. "Statistical Data Mining Tutorial on Gaussian Mixture Models," www.cs.cmu.edu/~awm/tutorials, CMU, 2004.
- [10] Scholkopf, B. and Smola, A.J. "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, 2002.
- [11] Burges, C.J. "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 1998.