AUDITORY SCENE CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

David Li, Jason Tam, Derek Toub

Cooper Union
41 Cooper Square
New York, NY 10003
samkeene@gmail.com

ABSTRACT

Audio scene classification will play an important role in context-based organization of audio data in the future. With classified and labeled audio data, it will be possible to set up a searchable database where users can retrieve audio files based on their contents. In this paper, we introduce a system to extract features from such audio scenes and identify the environments in which they were recorded. This system makes use of wavelet and Mel-frequency cepstral coefficient (MFCC) features, and classifies scenes by first classifying segments of the scene, and deciding the overall classification with a vote. The system achieves a classification accuracy of 72% for the training dataset provided for the IEEE AASP CASA challenge [1].

Index Terms—
MFCC, Treebagger, Wavelets, CASA

1. INTRODUCTION

A system for classification of audio scenes is comprised of two main parts: feature extraction, and decision making based on pattern recognition. Previous audio classification work centered mostly around speech processing, and made use of features such as MFCCs and short term cepstral coefficients. These features extract information from frequencies relevant to speech, however, they may not be the most effective for separating out different scenes. Thus, one of the challenges for the analysis of scene enviroments is choosing an effective set of features. We investigate a specific set of features in conjunction with a tree bagger classifier to successfully classify a combination of indoor and outdoor audio scenes [2].

2. FEATURE EXTRACTION

Several different features were used in order to construct a feature set that was adequate for further processing with classification techniques.

2.1. Short Term Feature Extraction

Audio scenes may be of a long enough length such that using an average of the features over the entire length is insufficient to adequately describe the scene. As such, the audio scenes of interest are initially segmented into 5 s windows with 2 s of overlap. Each 5 s scene segment will have its own set of computed features.

2.2. MFCCs

MFCCs have often been used as a feature set for human speech. There is a standard method for calculating these coefficients that allows for some variations in the spacing and shapes of windows used [3]. In this particular implementation, the short time fourier transform (STFT) was computed using Hamming windows over a 25 ms frame with 10 ms of overlap. 30 MFCCs were computed for each window. For a scene segment, the mean and variance of each MFCC were computed and used as features.

2.3. Wavelet Decomposition

The Discrete Wavelet Transform (DWT) is an alternative to the Discrete Fourier transform (DFT) that provides frequency and time resolution that more closely matches that of the human ear (compared to the DFT). More specifically, the DWT has high frequency resolution for low frequencies, and lower resolution for high frequencies. In this particular implementation, a 6 level wavelet decomposition using the Daubechies4 (D4) wavelet was used [4]. For each level, the ratio of adjacent coefficients, absolute value of the mean, variance of the wavelet coefficients over each scene segment were computed and used as features [5].

2.4. MFCCs of Wavelet Decomposition

MFCCs of a wavelet decomposition have been shown to be effective features for speaker identification systems in previous studies [6]. In this particular implementation,10 MFCCs were computed for each level of wavelet decomposition. The mean and variance of of the MFCCs are taken over the scene segment and used as features.

2.5. Other Features

Other features used were the spectral centroid and spectral flux. These measure the center of mass and rate of change of the power spectrum respectively [7].

3. CLASSIFICATION

The classification system takes in the features described in Section 2, passes them through a treebagger classifier, and decides the final class label by taking the treebagger class labels of each segment of audio as votes. Many of the tools used in this section are part of the PRT toolbox [8].

3.1. Initial Classification: Treebagger

A treebagger classifier works by using bootstrap aggregation to form a collection of decision trees. Bootstrap aggregation works by taking uniform random samples of a training set and using the ensemble average to form the final decision. In this implementation, 100 decision trees were used on a training set with the feature set as a predictors and the actual label as the response. Bagging only leads to minor improvements in error as errors in the individual models are most likely highly correlated [9].

3.2. Final Classification

Using the classifications of the initial classifier on each scene segment, a majority-wins voting system is implemented. Thus, a plurality of votes decides what the complete scene is classified as. In the event of a tie, the winner is chosen at random from the tied possibilities. The votes from the scene segments could be used as a feature set for another classifier. However, in this case, the mode has been empirically found to be effective.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

The system described above was applied to the public dataset for scene classification in the IEEE AASP CASA challenge [1]. This dataset is composed of 100 audio scenes representing 10 classes each 30 s in length. The classes represented are busy street, quiet street, supermarket, restaurant, office, park, bus, tube, tubestation, and open market. The scenes were recorded using a set of Soundman binaural microphones, model OKM II.

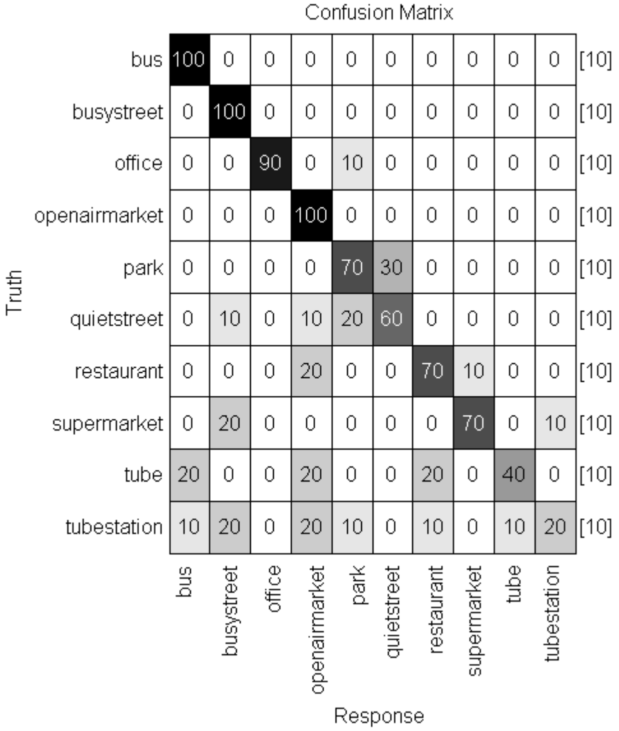
As we only have a set of training data, in order to evaluate our system, stratified K-folds cross validation is used. In stratified K-fold cross-validation, the original training set is partitioned into K subsamples with each subset being representative of the distribution of the original training set. The initial classification system is then trained using K-1 subsamples and verified using the last subsample. This procedure is repeated K times, using each subsample to verify exactly once. Hence, each member of the training set will have been ”predicted” by the classifier using other members of the training set exactly once. In this particular implementation, 10-fold cross-validation is used.

4.2. Experimental Results

The initial results are shown in Fig.1. As shown, the overall success rate is 72%. Many of the classes are classified successfully. However, classes tube and tubestation appear to have major difficulties with classification. This is likely due to the inabilty of the features used in this system to pick out unique characterstics of these scenes. Future work could focus on finding new features that show different behavior for these two scenes compared to all the others. Results for only the first stage of classification are shown in Fig.2. As shown, using scene segments resulted in improvement as the overall sucess rate is 67% without.

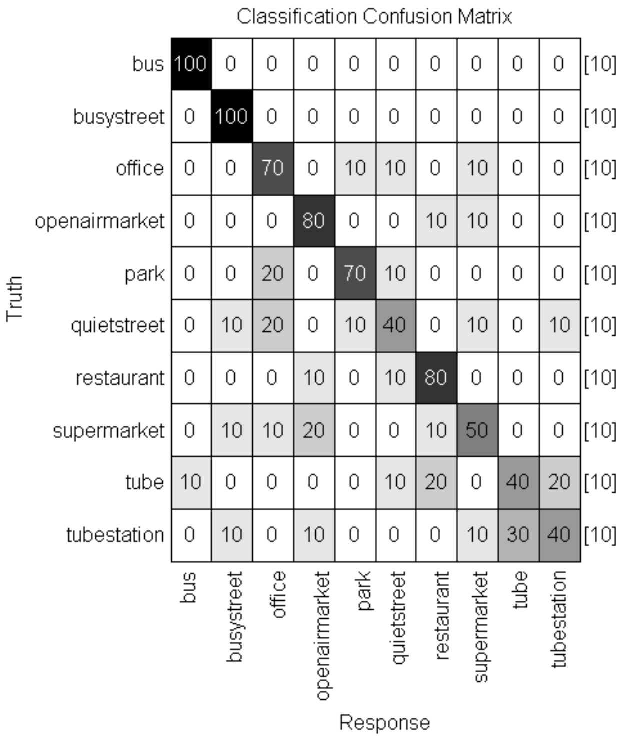
5. ACKNOWLEDGMENT

The authors would like to thank Sam Keene for constructive criticisms and suggestions.


Confusion Matrix
Truth
Response

	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation	
bus	100	0	0	0	0	0	0	0	0	0	[10]
busystreet	0	100	0	0	0	0	0	0	0	0	[10]
office	0	0	90	0	10	0	0	0	0	0	[10]
openairmarket	0	0	0	100	0	0	0	0	0	0	[10]
park	0	0	0	0	70	30	0	0	0	0	[10]
quietstreet	0	10	0	10	20	60	0	0	0	0	[10]
restaurant	0	0	0	20	0	0	70	10	0	0	[10]
supermarket	0	20	0	0	0	0	0	70	0	10	[10]
tube	20	0	0	20	0	0	20	0	40	0	[10]
tubestation	10	20	0	20	10	0	10	0	10	20	[10]

Figure 1: Classification with complete system.


Classification Confusion Matrix
Truth
Response

	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation	
bus	100	0	0	0	0	0	0	0	0	0	[10]
busystreet	0	100	0	0	0	0	0	0	0	0	[10]
office	0	0	70	0	10	10	0	10	0	0	[10]
openairmarket	0	0	0	80	0	0	10	10	0	0	[10]
park	0	0	20	0	70	10	0	0	0	0	[10]
quietstreet	0	10	20	0	10	40	0	10	0	10	[10]
restaurant	0	0	0	10	0	10	80	0	0	0	[10]
supermarket	0	10	10	20	0	0	10	50	0	0	[10]
tube	10	0	0	0	0	10	20	0	40	20	[10]
tubestation	0	10	0	10	0	0	0	10	30	40	[10]

Figure 2: Classification of partial system without segmentation and voting.

6. REFERENCES

[1] http://www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge/.

[2] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] D. P. W. Ellis, “PLP and RASTA (and MFCC, and inversion) in Matlab,” 2005, online Web Resource, Accessed Mar. 30, 2013. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[4] I. Daubechies, “The wavelet transform, time-frequency localization and signal analysis,” *Information Theory, IEEE Transactions on*, vol. 36, no. 5, pp. 961–1005, 1990.

[5] G. Tzanetakis, G. Essl, and P. Cook, “Audio analysis using the discrete wavelet transform,” in *Proc. Conf. in Acoustics and Music Theory Applications*, 2001.

[6] M. I. Abdalla and H. S. Ali, “Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov models,” *arXiv preprint arXiv:1003.5627*, 2010.

[7] T. Giannakopoulos, “Feature extraction toolbox,” 2009. [Online]. Available: http://cgi.di.uoa.gr/~tyiannak/Welcome.html

[8] P. Torrione, S. Keene, and K. Morton, *PRT: The Pattern Recognition Toolbox for MATLAB*, 2011, software available at http://newfolderconsulting.com/prt.

[9] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 4, no. 4.