Automatic Acoustic Scene Classification

Gonçalo Marques^a, Thibault Langlois^b
^{a}ISEL, Electronic Telecommunications and Computer Department, Lisbon, Portugal
^{b}FCUL, Informatics Department, Lisbon, Portugal
gmarques@deetc.isel.pt tl@di.fc.ul.pt

Abstract— **This paper presents a baseline system for automatic acoustic scene classification based on the audio signals alone. The proposed method is derived from classic, content-based, music classification approaches, and consists in a feature extraction phase followed by two dimensionality reduction steps (PCA and LDA) and a classification phase done using a k nearest-neighbours algorithm.**

This paper also reports on how our system performed in the context of the DCASE 2016 challenge [1], for the acoustic scene classification task.

Keywords: Machine Learning, Signal Processing, Music Information Retrieval, Bag of Frames.

I. Introduction

Automatic identification of sound sources in an urban environment has a huge potential in several applications related to the current panorama of intelligent cities, like monitoring systems able to recognize activities, sound environments, and create city sound maps to provide to the general public information about environmental noise, or other acoustic factors. However, a lot of research is still needed to reliably detect and recognise sound events and scenes in realistic environments where multiple sources, often distorted, are present simultaneously. This works focusses on one particular aspect of urban sound analysis: acoustic scene classification.

The system we purpose is a classical classification system in the sense that it uses typical machine-learning data transformation and classification algorithms in the decision making process. First, each audio excerpt is converted into a single feature vector which is the representation of choice for standard machine-learning methods. Then, the whole dataset is transformed via principal component analysis (PCA), an unsupervised dimensionality reduction technique, followed by a linear discriminant analysis (LDA) projection. LDA is a supervised process, and the projection tries to maximize the ratio between intra and inter class scatter, but it is not a classification method since no decision is involved. For classification, a k nearest-neighbours (k-NN) algorithm was used. The experimental configuration used in our tests is common in many audio classification works (or at least parts of it – see for *e.g.* [4], [7]) and therefore it does not bring any original contribution in terms of the algorithmic setup. In fact, our system falls under the standard “bag of frames” classifiers commonly used in music information retrieval applications (see [3], [8] and references within). Our main objective was not to bring fourth a new audio classification or feature extraction method, but rather see how a simple, non parametric algorithm performed in acoustic scene classification challenge. We used the same data partitioning and cross-validation setup provided with the database and our results are a bit better than the ones in [5] (the baseline provided with the challenge). The structure of the remainder of this report is as follows: Section II describes the data and the feature extraction process used in our experiments, Section III describes our approach to acoustic scene classification, followed by Section IV where we present our results. Section V concludes this paper.

II. Data and Feature Extraction

The dataset used in this work was created in the context of the DCASE2016 challenge [1] for the acoustic scene classification task. The dataset contains 1170, 30-seconds audio excerpts from the following acoustic scenes: Beach, Bus, Café/Restaurant , Car, City Center, Forest Path, Grocery Store, Home, Library, Metro Station, Office, Urban Park, Residential Area, Train, and Tram. The dataset is divided into four folds for cross-validation testing, and our resuls are averaged over the four test folds.

The features used are the all-purpose Mel frequency cepstral coefficients (MFCCs), a representation very popular in speech recognition (see for e.g [6]), and also widely used in content-based music information applications. The audio was divided into 23 ms segments (1024 samples at 44.1 kHz) with 50% overlap, and we used 100 Mel bands to extract 23 MFCCs plus the zero order MFCC and the frame's log-energy, plus the delta and acceleration coefficients. This means that the audio is converted into a sequence of $25 \times 3 = 75$ dimensional vectors. We used the software VoiceBox [2] to extract the features. In order to convert each audio excerpt into a single feature vector, the sequence of MFCC features is summarized using the median and logarithmic standard deviation. The median was used instead of the mean since this statistic is more robust to outliers. The log-standard deviation is given by $20 \log_{10}(\sigma_i)$ where σ_i is the standard deviation of feature i (with $i = 1, \ldots, 75$). The reason to use the log-standard deviation instead of plain standard deviation was to convert these feature values to an order of magnitude comparable the median feature values - otherwise the standard deviation values would be a few orders of magnitude lower, and during the PCA pre-processing step, this dimensions would be discarded as noise since they would not contribute in any significant way to the overall data variance. The statistics return two 75-dimensional vectors which are concatenated, so each audio excerpt is represented by a 150-dimensional feature vector.

III. Method

The proposed classification approach is divided into three main blocks: feature pre-processing via principal component analysis, feature transformation by linear discriminant analysis and finally a classification step performed by a k-nearest neighbour classifier.

Principal Component Analysis: PCA is a standard dimensionality reduction technique, where the data is decorrelated by projecting it into orthogonal directions of maximum variance. These directions, the principal components, are obtained using a eigen-decomposition of the data covariance matrix, and in our experiments we kept enough components to explain 99.9% of the total data variance. The PCA-transformed data was also whitened - each data dimension was scaled in order to have unit variance.

Linear Discriminant Analysis: LDA is commonly used as a pre-processing step for pattern classification. It is also a dimensionality reduction technique since the data is projected into $c-1$ dimensional space where c is the total number of classes ($c=15$ for this challenge).

k-Nearest Neighbours: k-NN is an instance-based learning, where class membership is assigned based on a majority vote of its neighbours. k-NN is possibly one of the simplest classification methods, and therefore it is well suited for a baseline system. We used the Euclidean distance and tested the algorithm with different number of neighbours (from 5 to 31) and chose $k=9$. The results reported in Section IV are obtained using the Euclidean distance metric, and $k=9$.

IV. Experimental Results

The results presented in this section were obtained using the following experimental setup. The PCA and the LDA projection were estimated using only the training set. In our tests, we used 4-fold cross validation and the same data partitioning provided with the dataset. The presented result pertain to the tests folds only. The average accuracy obtained was 77.4%. In Table II are the (average) accuracies per class.Table I shows the confusion matrix (obtained summing the four confusion matrices in each test fold). Each line refers to the examples of a single class; the class order is the same as the one in Table II. In the columns are the classification results. For example, for the class Beach, 60 audio excerpts were correctly classified, 9 were classified as the class Urban Park, and nine others were also misclassified.

V. Conclusion

In this work, we presented a baseline for acoustic scene classification system composed of two dimensionality reduction transformation (PCA and LDA) followed by a k-NN classification algorithm.We trained and tested our method on the DCASE 2016 acoustic scene classification dataset, and submitted it to the challenge provide by the organization.

VI. Acknowledgement

Gonçalo Marques thanks FI-Sonic for supporting his research.

TABLE I

Confusion matrix - the rows are the true classes, the columns are the classification results. The class order is the same as the one given in Table II. This matrix was obtained by the sum of the four confusion matrices - one per test fold.

60	0	0	0	**1**	**2**	0	0	**1**	0	**1**	**9**	**3**	0	**1**
0	**52**	**6**	**2**	0	0	**1**	0	**5**	0	0	0	0	**11**	**1**
0	0	**62**	0	0	**2**	**8**	0	**2**	**4**	0	0	0	0	0
0	**3**	0	**66**	0	0	0	0	**1**	0	0	0	0	**7**	**1**
1	0	0	0	**73**	0	0	0	**1**	**1**	0	0	**2**	0	0
1	0	0	0	0	**68**	0	0	0	0	**2**	**7**	0	0	0
0	0	**4**	0	0	0	**64**	0	**1**	**9**	0	0	0	0	0
2	**2**	**7**	0	0	**2**	0	**50**	**8**	0	**6**	0	0	0	**1**
0	0	**9**	0	0	0	**1**	0	**68**	0	0	0	0	0	0
0	0	0	0	0	0	**2**	0	**3**	**73**	0	0	0	0	0
0	0	0	0	0	**1**	0	**5**	0	0	**72**	0	0	0	0
2	**1**	0	0	**1**	**13**	**3**	0	**6**	0	0	**47**	**5**	0	0
2	0	**2**	0	**1**	**3**	0	0	0	0	**1**	**17**	**51**	0	**1**
0	**11**	**8**	0	0	**2**	**4**	0	0	**5**	**1**	0	0	**41**	**6**
0	**1**	0	0	0	0	**7**	0	0	0	**1**	0	0	**1**	**68**

TABLE II

Accuracy per class. Accuracy values obtained with the mean of all four test folds.

1. Beach	76.9%
2. Bus	66.7%
3. Café/Restaurant	79.5%
4. Car	84.6%
5. City Center	93.6%
6. Forest Path	87.2%
7. Grocery Store	82.1%
8. Home	64.1%
9. Library	87.2%
10. Metro Station	93.6%
11. Office	92.3%
12. Urban Park	60.3%
13. Residential Area	65.4%
14. Train	52.6%
15. Tram	87.2%

References

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[2] VOICEBOX: Speech Processing Toolbox for MATLAB (2005) by Mike Brookes.

[3] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008.

[4] Sander Dieleman and Benjamin Schrauwen. Multiscale approaches to music audio feature learning. In *14th International Society for Music Information Retrieval Conference (ISMIR-2013)*, pages 116–121, 2013.

[5] A. Mesaros, T. Heittola, and T. Virtanen. TUT database for acoustic scene classification and sound event detection. In *24th European Signal Processing*, Budapest, Hungary, 2016.

[6] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[7] Justin Salamon and Juan Pablo Bello. Unsupervised feature learning for urban sound classification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175. IEEE, 2015.

[8] Bob L Sturm. A survey of evaluation in music genre recognition. *Proc. Adaptive Multimedia Retrieval, Denmark*, 2012.