MIREX 2011: MUSIC GEREN CLASSIFICATION VIA SPARSE REPRESENTATION

Jia-Min Ren[1], Ming-Ju Wu[1], and Kaichun K. Chang[2]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
{jmren, hsbw}@mirlab.org

[2]Department of Computer Science, King's College London, London, United Kingdom
ken.chang@kcl.ac.uk

ABSTRACT

This extended abstract details our submission to the Music Information Retrieval Evaluation eXchange (MIREX) 2011 for the audio training\test task. First of all, we extract a fixed-length feature vector (composed of some timbral as well as modulation spectrum features) from each training clip. Then, by representing a fixed-length feature vector (extracted from a test clip) as a linear combination of all training feature vectors, we classify this test clip as a class with the minimal re-construction residual. This is so-called a sparse representation based classifier (SRC).

1. INTRODUCTION

In recent years, modulation spectral analysis [1] and sparse representation [2] have been attracted much attention in the field of music information retrieval. In our system, modulation spectral features such as octave-based modulation spectral contrast (OMSC) [3], modulation spectral flatness measure (MSFM) [3], and modulation spectral crest measure (MSCM) [3] are extracted from each long segment (also named *texture window*). In addition, short-time timbral features such as Mel-scale frequency cepstral coefficient (MFCC), octave-based spectral contrast (OSC), spectral flatness/crest measure, spectral centroid, spectral rolloff, spectral flux, spectral skewness, and spectral kurtosis are extracted from each short segment (also named *analysis window*). Then, we compute the mean and standard deviation along each feature dimension (see Section 2 for more details) to obtain a fixed-length feature vector for each clip. In the classification stage, we use sparse representation based classifier (SRC) [4]. Details related to SRC can be found in Section 3. It should be noted that our submission is similar to our previous work [5], except that we do not utilize a random measurement matrix (e.g. a Gaussian random matrix) to reduce the dimensionality of feature vectors.

2. FEATURE EXTRACTION

In our system, we extract short-time timbral features from "*analysis window*", and modulation spectrum features from "*texture window*". Here the length of *analysis window* and *texture window* were set to 93 ms and 10 seconds and 50% overlap was used for feature extraction. However, since the length of song clips in MIREX genre classification task is 30 seconds, we divide each clip into three 10-second segments (without overlapping).

The following describe the extracted timbral features from *analysis windows* of a segment (the number in each parenthesis is the dimensionality of extracted features).

Mel-scale Frequency Cepstral Coefficients (MFCCs) (13): represents the spectral characteristics based on Mel-frequency scaling.

Octave-based Spectral Contrast (OSC) (16): considers the spectral peak and valley in each sub-band independently, where the former corresponds to harmonic components and the latter corresponds to non-harmonic components or noise in music signals. We extracted spectral peaks and the difference between spectral peak and valley (this difference also named *spectral contrast*, reflecting the spectral contrast distribution) from eight sub-bands [5].

Spectral Flatness/Crest Measure (16): measures of the noisiness (flat, decorrelation) sinusoidality of a spectrum, where the former is computed by the ratio of the geometric mean to the arithmetic mean of the energy spectrum value in each sub-band, and the latter is computed by the ratio of the maximum value within each sub-band to the arithmetic mean of the energy spectrum value [6]. Totally eight sub-bands as set in extracting OSC features were used here.

Spectral Centroid (1): the centroid of amplitude spectrum.

Spectral Rolloff (1): the frequency bin below which 85% of the spectral distribution is concentrated.

Spectral Flux (1): the squared difference of successive amplitude spectrum.

Spectral Skewness (1): a measure (the 3rd order moment) of the symmetry of the spectral distribution.

Spectral Kurtosis (1): a measure (the 4th order moment) of the flatness of the spectral distribution.

To summarize the feature vectors extracted from each segment, the mean and standard deviation along each feature dimension are computed, resulting in a 100-dimensional feature vector for each segment.

The following describe modulation spectral features extracted from *texture window* of a segment (the number in each parenthesis denotes the dimensionality of extracted features).

Octave-based Modulation Spectral Contrast (OMSC) (16x12): this feature is extracted using long-term modulation spectral analysis [7], resulting in a two-dimensional joint acoustic frequency and modulation frequency. Here we computed modulation spectral peak and modulation spectral contrast in six sub-bands to obtain a matrix of size 16-by-12.

To capture the frequency variable (acoustic frequency), we computed mean and standard deviation for each row of this matrix. On the other hand, in order to capture time-varying information through temporal modulation (modulation frequency), we computed mean and standard deviation for each column of this matrix. After these two operations, we obtain an 88-dimensional (24+64) feature vector for each segment.

Modulation Spectral Flatness/Crest Measure (MSFM/MSCM) (8/8): these two features can be used to describe the time varying behavior of the subband energy. A detailed explanation of MSFM and MSCM can be found in [3].

In summary, totally a 204-dimensional feature vector was extracted from each segment (100-dimensional features from *analysis windows* and 104-dimensional features from *texture windows*).

3. SPARSE REPRESENTATION BASED CLASSIFIER (SRC)

Let the dimension of extracted features be m (in our case m=204), and the feature vector of the j-th clip (or segment) in the i-th genre as $v_{i,j} \in \mathbb{R}^m$. Then given sufficient training samples of the i-th genre, $A_i = \left[v_{i,1}, v_{i2}, ..., v_{i,n_i} \right] \in \mathbb{R}^{m \times n_i}$, any new sample $y \in \mathbb{R}^m$ (i.e., the extracted feature vector of a test clip) from the same genre will approximately lie on a linear subspace spanned by the i-th genre's training samples:

$$y = \sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j}, \tag{1}$$

for some scalars $\alpha_{i,j} (j = 1, 2, ..., n_i)$.

However, since the identity of a test sample is initially unknown, we can develop a *global* dictionary matrix A for the k genres by concatenating the entire training samples as $A = \left[A_1, A_2, ... A_k \right] \in \mathbb{R}^{m \times n_i k}$. Then the linear representation of y can be rewritten in terms of all n training samples $(n = n_i \times k)$ as:

$$y = Ax \in \mathbb{R}^m, \tag{2}$$

where $x = \left[0, ..., 0, \alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n_i}, 0, ..., 0 \right]^T \in \mathbb{R}^n$ is a coefficient vector whose entries are zero except those associated with the i-th genre. Note that only those entries corresponding to the genre of y are non-zero. Thus if we can solve the equation (2), then we can find the genre of y. Recent work in sparse representation [8] has shown that the sparsity of x enables us to solve the equation (2) using the l_1-norm minimization:

$$\hat{x}_1 = \arg\min \|x\|_1 \text{ subject to } Ax = y. \tag{3}$$

Once $\hat{x}_1$ has been estimated, the genre of y can be simply decided by locating the non-zero entries in $\hat{x}_1$. However, noise and modeling limitations may lead to $\hat{x}_1$ has some small non-zero entries belonging to different genres. To solve this problem, for each genre i, we define δ_i as a characteristic function: $\mathbb{R}^n \to \mathbb{R}^n$, which selects the coefficients associated with the i-th genre. Now for each genre i, we can approximate $\hat{y}_i = A\delta_i(\hat{x}_1)$, and classify y based on these approximations by assigning it to the genre with the minimum residual between y and $\hat{y}_i$:

$$\min r_i(y) = \|y - A\delta_i(\hat{x}_1)\|_2. \tag{4}$$

It should be noted that since we extracted three segments from a song clip, the final genre of a song clip is determined by a major vote of the classified genres of these three segments.

4. EVALUATION

In this extended abstract, a widely used dataset, GTZAN [9], (consisting of ten genres and 100 30-second song clips per genre) is chosen for the evaluation. However, since this dataset is not artist-filtered, a ten-fold cross-validation, which was used in most of existing approaches to evaluate their performance (see Table 1), is not a fair strategy to compare the performance with different approaches. Spe-

cifically, according to our experience, the performance (e.g., the averaged accuracy of ten-fold classification results) will be better about 5-10% if a "lucky" cross-validation split of this dataset is given. Therefore, we used **leave-one-out cross-validation** here in order to provide a *fair baseline* for researchers who want to evaluate their performance on this dataset in the future.

Reference	Cross-Validation	Accuracy
Tzanetakis and Cook [9]	Ten-fold (randomly repeated ten times)	61.0%
Panagakis *et al*. [10]	Ten-fold	78.2%
T. Li *et al*. [11]	Ten-fold	78.5%
Li and Ogihara [12]	Ten-fold	78.5%
Panagakis *et al*. [13]	Ten-fold	84.3%
MIREX 2011 Submission	**Leave-one-out**	**86.1%**
Lee *et al*. [1]	Ten-fold (randomly repeated ten times)	90.7%
Panagakis and Kotropoulos [2]	Ten-fold	93.7%
Y.-F. Huang and Y.-S. Li [14]	Ten-fold	97.2%

Table 1. Comparison of classification approaches evaluated on the GTZAN dataset.

5. REFERENCES

[1] C.-H. Lee, J.-H. Shih, K.-M. Yu, and H.-S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Feature," *IEEE Trans. Multimedia*, Vol. 11, No. 4, pp. 670–682, 2009.

[2] Y. Panagakis and C. Kotropoulos, "Music Genre Classification via Topology Preserving Non-Negative Tensor Factorization and Sparse Representations," in *Proceedings of ICASSP*, pp. 249–252, 2010.

[3] D. Jang, M. Jin, and C. D. Yoo, "Music Genre Classification Using Novel Features and A Weighted Voting Method," in *Proceedings of ICME*, pp. 1377–1380, 2008.

[4] J. Wright, A. Yang, A. Ganesh, S. Shastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. PAMI*, Vol. 31, No. 2, pp. 210–227, 2009.

[5] K. K. Chang, J.-S. Roger Jang, and C. S. Iliopoulos, "Music Genre Classification Via Compressive Sampling," in *Proceedings of ISMIR*, pp. 387–392, 2010.

[6] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO I.S.T Project Report*, 2004.

[7] T. Kinnunen, "Joint Acoustic-Modulation Frequency for Speaker Recognition," in *Proceedings of ICASSP*, pp. 14–19, 2006.

[8] D. Donoho, "For Most Large Underdetermined Systems of Linear Equations The Minimal l1-norm Solution Is Also The Sparse Solution," *Comm. On Pure and Applied Math*., Vol. 59, No. 6, pp. 797–829, 2006.

[9] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signal," *IEEE Trans. Audio and Speech Processing*, Vol. 10, No. 5, pp. 293–302, 2002.

[10] I. Panagakis, E. Benetos, and C. Kotropoulos, "Music genre classification: A multilinear approach," in *Proceedings of ISMIR*, pp. 583–588, 2008.

[11] T. Li, , M. Ogihara, and Q. Li, "A Comparative Study on Content-based Music Genre Classification," in *Proceedings of ASM SIGIR*, pp. 282–289, 2003.

[12] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, Vol. 8, No. 3, pp. 564–573, 2006.

[13] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 18, No. 3, pp. 576–588, 2010.

[14] Y.-F. Huang and Y.-S. Li, "Music Genre Classification Based on Local Feature Selection Using a Self-Adaptive Harmony Search Algorithm," *Master Thesis, Graduate School of Computer Science and Information Engineering, National Yunlin University of Science and Technology (Taiwan)*, 2011.