International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.4 (2016), pp.185-192
http://dx.doi.org/10.14257/ijsip.2016.9.4.17

Sparse Representation Classification-Based Automatic Chord Recognition with Different Pitch Class Profile Features

Zhongyang Rao[1,2], Xin Guan[1] and Jianfu Teng[1]

[1]*School of Electronic Information Engineering, Tianjin University， Tianjin, China*
[2]*School of Information Science & Electric Engineering, Shandong Jiaotong University, Jinan, China*
yaozhongyang@sohu.com

Abstract

In this paper, a machine-learning approach called Sparse Representation-based Classification (SRC) is used for automatic chord recognition in music signals. We extracted different Pitch Class Profile (PCP) features from raw audio and achieved sparse representation of classes via ℓ^1-norm minimization on feature space to recognize 24 major and minor triads. This recognition model is evaluated on MIREX'09 dataset including the Beatles corpus. Our method is compared with various methods that entered the Music Information Retrieval Evaluation eXchange (MIREX) in 2014 towards the audio chord estimation of MIREX'09 dataset in Audio Chord Estimation task of MIREX. Experimental results demonstrate that our method has good accuracy rate in recognizing maj-min chords.

Keywords*: Chord recognition, Music Information Retrieval, PCP, Sparse Representation-based Classification*

1. Introduction

In music, a chord is a set of three or more notes that is played simultaneously. Automation of chord labeling is also called chord recognition, which finds many applications such as music segmentation, cover Song identification, audio matching, music similarity identification, and audio thumb nailing[1]. Because of these reasons, automatic chord recognition has been one of the main fields of interest in musical information retrieval (MIR) in the last few years.

In chord recognition, the features used may not identical. But in most cases, the most used features is variants of the Pitch Class Profile (PCP) introduced by Fujishima (1999)[2]. PCP is also called chroma vector, which is often a 12-dimensional vector. The calculation of an audio file into a chroma representation is based either on the short-time Fourier transform (STFT) in combination with binning strategies [3] or on the constant Q transform (CQT)[4]. One of the limitations of the STFT is that it uses a fixed–length window. In music signal processing, chromagram is defined as the succession of these chroma vectors over time. And the musical content of audio musical signals can be well described with the chromagram.

The chord recognition is the chord labeling of each chord. Our chord recognition system is based on the sparse representation-based classification (SRC) [5] which has been proposed with amazing identification capability in recent years. Based on a giving 12-dimensional PCP features, SRC discriminately selects the subset that most compactly expresses the input signal and rejects all other possible but less compact representations. This classification has been applied into many applications and achieved perfect performance. This is the first time that we apply SRC into chord recognition; empirical

ISSN: 2005-4254 IJSIP

experiments demonstrate that its perfect discrimination capability compared with other classifications.

The remainder of this paper is organized as follows: Section 2 reviews previous the related work on this area; Section 3 gives a description of our construction of the feature vector; Section 4 detailedly describes our sparse approach; Section 5 gives results on a data corpus and a comparison with the other methods; Finally we will draw some conclusion and give possible developments about further work.

2. Related Work

The feature of pitch class profile (PCP) has almost without exception as the feature of the chord recognition system. In [2], Fujishima developed a real-time chord recognition system, where used discrete fourier transform (DFT) of the music audio and obtained a 12-dimensional pitch class profile, then determined chord type based on pattern matching. Lee [6] also used pattern matching based on binary chord templates, determined the 24major/minor triads. A new feature called Enhanced Pitch Class Profile (EPCP) is introduced. E. Gómez and P. Herrera [7] used a Harmonic Pitch Class Profile (HPCP) as the feature vector, which is based on Fujishima's PCP, and correlated it with a chord or key model adapted from Krumhansl's cognitive study.

Besides templates-fitting methods, it is widely used machine-learning methods such as Support Vector Machine(SVM) [6] and hidden Markov Model (HMM) [8] for this recognition process. A. Sheh and D. P. Ellis proposed a statistical learning method for chord segmentation and recognition[8]. J. P. Bello and J. Pickens also used the HMMs with the EM algorithm, but they considered the inherent musicality of audio into the models for model initialization[9].

3. Feature Vectors

First of all, the recognition system extracts a sequence of suitable feature vectors from the audio signal. In our system, the feature vectors is PCP.

Like most chord recognition systems, a chromagram or a PCP vector is used as the feature vectors. Müller and Ewert propose feature vectors 12-dimensional Quantized PCP[10] which avoids a possible frequency resolution and is sufficient to separate musical notes of low frequency comparing with others.

The calculation of feature vectors PCP can be divided into the following steps: (1)Using the constant Q transform to calculate the 36-bin chromagram; (2)Mapping spectral chromagram to a particular semitone; (3) segmenting the audio signal with beat tracking algorithm; (4)Reducing the 36-bin chromagram to 12-bin chromagram based on beat-synchronous segmentation. (5)Logarithm and normalization of 12-bin chromagram.Refer to [9] for more detailed steps on how to calculate chromagram.

(1)36-bin chromagram calculation. Using the constant Q transform, it can get $X_{cqt}(k)$ of a audio signal $x(m)$:

$$X_{cqt}(k)=\frac{1}{N_k}\sum_{m=0}^{N_k-1}x(m)w_{N_k}(m)e^{\frac{-j2\pi mQ}{N_k}} \tag{1}$$

where k is the bin position, $w_{N_k}(m)$ is the hamming window and its length $N_k = Qf_k / f_s$. And f_k is the center frequency of the k bin and f_s is the sample frequency. In this paper, the music signal is down-sampled to 11025Hz.

By adding all $X_{cqt}(k)$ that correspond to a particular frequency, then it get 36-bin chromagram of each frames. The specific formula is as follows:

$$QPCP(p)=\sum_{m=0}^{M-1}\left|X_{cqt}(p+mb)\right|,\quad p=1,2,...,36 \tag{2}$$

Where *M* is the total number of octaves and *b* is the number of bins per octave.

(2)Chromagram tuning. In the 36-bin chromagram, 3 bins represent one note in the octave. Each spectral components of 36-bin is maped to a particular semitone. The mapping formula is as follows:

$$p(k)=36*\left[\log_2(f_s/N_k*k/f_0)\right]\bmod 36 \tag{3}$$

(3)beat-synchronous segmentation. In our system, it use the beat tracking with dynamic programming method proposed by Daniel P.W. Ellis [11]. This approach has been found to work very well in in many types of music. Segmenting the audio signal with beat tracking algorithm has additional advantage that the chroma feature is a function of beat segments, rather than time.

(4)12-bin chromagram reduction. Finally, averaging the each spectral components of 36-bin in beat segments and summing them in semitones, thus the dimension of chromagram is reduced to 12 from 36. Then the chromagram of audio music can represented with these 12 dimensional vectors.

(5)Logarithm and normalization of 12-bin chromagram. $QPCP_{12}(p)$ is the 12-bin chromagram. It can get the normalized value with p-norm and logarithm. The formula is as follows:

$$QPCP_{\log}(p)=\log_{10}(C*QPCP_{12}(p)+1) \tag{4}$$

$$QPCP_{norm}(p)=QPCP_{\log}(p)/\left\|QPCP_{\log}\right\| \tag{5}$$

If it performs the Logarithm and normalization, the chromagram is called Log PCP. In step (5), if it has only normalization, it is called PCP.

4. Sparse-Based Classification

In recent years, the sparse representation become an important research focus in the field of pattern recognition, and has attracted wide attention in areas such as machine vision, machine learning, pattern recognition. The earliest in the field of signal sparse representation has been proposed[12, 13]. Its core idea is that the test sample is the linear representation of labeled training samples which the test sample belongs to. Obviously, only a few of the linear coefficient are zero, that is to say the coefficient vector is sparse.

Our chord recognition system is based on the sparse representation-based classification (SRC) [5]. Labeled samples by this algorithm can directly be used as the classifier training samples, saving lots of time and system resources. The following sections outline the method.

(1)Test Samples as a sparse linear combination of training samples.

At first, we define a matrix $D\doteq[D_1,D_2,\cdots D_k]=[u_{1,1},u_{1,2},\cdots u_{k,n_k}]\in R^{m\times n}$ by collecting n classifier training samples of all k classes, where *m* is the dimension of the feature set. For a given test sample $y\in R^m$ form subject i, can be rewritten in terms of all training samples as:

$$y=Dx_0\in R^m \tag{6}$$

where x_0 ${x_0}$ is a coefficient vector whose entries Ideally the coefficient vector $x_0=[0,\cdots,0,a_{i,1},a_{i,2},\cdots,a_{i,n_i},0,\cdots,0]^T\in R^n$ are mostly zero except the values corresponding to the *i*-th class are non-zero and other coefficient values should be 0.

As coefficient vector x_0 can identify the test sample y, it can be obtained by solving the linear equation (6).

(2)Sparse solution via ℓ^1-Minimization.

Recent development in the emerging compressed sensing theory and sparse representation reveals that if the solution x_0 sought is sparse enough, the solution to the system of equation (6) is equivalent to the following ℓ^1-minimization problem:

$$\hat{x}_1 = \arg\min \|x\|_1 \quad \text{subject to } y = Dx \tag{7}$$

(3)Classification based on sparse representation.

According to these non-zero coefficient x_1, it can quickly know the test sample belongs to the class. Actually, because of noise and model errors, some of entries with multiple object classes is small nonzero values. For each class *i*, t he given test sample *y* can be approximated as $\hat{y}_i = D\delta_i(\hat{x}_1)$, where $\delta_i : R^n \to R^n$ is the characteristic function which selects the coefficients associated with the *i*-th class. We then calculate the residual between y and $\hat{y}_i$:

$$r_i(y) = \|y - D\delta_i(\hat{x}_1)\|_2 \tag{8}$$

At last, we classify *y* based on these approximations by assigning it to the object class that minimizes the residual, as follow:

$$identity(y) = \arg\min_i r_i(y) \tag{9}$$

The resulting SRC algorithm is summarized below.

Algorithm 1: Recognition via Sparse Representation Classification (SRC)

1: Input: B is a matrix of classifier training samples, $D = [D_1, D_2, \cdots D_k] \in R^{m \times n}$ for *k* classes, a test sample $y \in R^m$.

2: Solve the following l^1-minimization problem: $\hat{x}_1 = \arg\min \|x\|_1$ subject to $y = Dx$

3: Compute the residuals $r_i(y) = \|y - D\delta_i(\hat{x}_1)\|_2$, for $i = 1, ..., k$

4: Output: $identity(y) = \arg\min_i r_i(y)$

5. Evaluation

5.1. Corpus

For evaluation, we use the MIREX'09 dataset in Audio Chord Estimation task of MIREX. The dataset consists of 12 Beatles albums (180 songs, PCM 44 100Hz, 16 bits, mono). Besides the Beatles albums, in 2009, an extra dataset was donated by Matthias Mauch which consists of 38 songs from Queen and Zweieck.

This database based been extensively used for the Audio Chord detection task at 2014. The evaluation is realized thanks to the chord annotations of the Beatles albums kindly provided by Harte and Sandler[14].

The chord dictionary used in this work is the set of 24 major and minor triads, one each for all 12 members of the chromatic scales: C Major, C minor, C# Major, C# minor ... B Major, B minor. Each triad contains 50 labeled musical fragments which select from the Beatles albums.

To evaluate the quality of an automatic transcription, a transcription is compared to ground truth created by one or more human annotators. Since 2013, MIREX typically uses chord symbol recall (CSR) to estimate how well the predicted chords match the ground truth:

$$CSR = \frac{total\ \ duration\ \ of\ \ segments\ \ where\ \ annotation\ \ equals\ \ estimation}{total\ \ duration\ \ of\ \ annotated\ \ segments} \tag{10}$$

Because pieces of music come in a wide variety of lengths, we will weight the CSR by the length of the song when computing an average for a given corpus. This final number is referred to as the weighted chord symbol recall (WCSR).

5.2. Experiment

In order to verify the impact of different dimensions of the feature space on the results, it first tests the algorithm of SRC using different samples. The results are presented in Table 1.

Table 1. Recognition Rates of SRC on the Bealts Corpus

Number of Samples	5	10	20	30	40	50
PCP[%]	57.6	60.3	61.7	63.8	66.6	67.0

We compared recognition rate of the PCP chromagram with some popular features. (1)Short Time Fourier Transform STFT chromagram features (STFTC), which is implemented by the MIR toolbox[15]. (2)Chroma DCT-reducedlog Pitch features (CRP)[10], which is the logarithmized pitch representation like MFCC. (3)Loudness based chromagram as described in [16] (denoted by LBC).

The results are presented in Table 2.

Table 2. Comparison of Recognition Rates using Different Chromagram

Chromagram	PCP	STFTC	CRP	LBC	Log PCP
Recognition Rates [%]	67.0	59.7	42.6	64.9	73.9

The recognition rates computed for this corpus confirm these observations:

(1)The performances of Log PCP features exceed the performances of classical feature PCP. More specifically speaking, the recognition rate for the Log PCP is better, 73.9%, compared to 67.0% for that of PCP.

(2)By imposing sparsity via ℓ^1-minimization, the recognition rates of all features improve gracefully as the number of samples increases from 5 to 50. The performance of PCP features gracefully increased from 55.2% to 67.0%.

From the observations 1) and 2), we can draw a conclusion that the choice of a good combination of features and samples number indeed makes some difference for SRC classification. These experimental results have verified the theoretical analyses of compressed sensing in this paper. The result is bad if the training data (feature dimension or samples number) of a single subject do not span a subspace.

5.3. Comparison with the Previous Methods

Our method is now compared to the following methods that entered MIREX 2014.

(1)KO1: Maksim Khadkevich & Maurizio Omologo [17];

(2)CM3: Chris Cannam, Matthias Mauch [18];

(3)JR2: Jean-Baptiste Rolland [19];

More details about these methods can be found from the corresponding MIREX websites - http://www.music-ir.org/mirex/wiki/MIREX_HOME.

Results of the comparison with the state-of-the-art are presented in Table 3.

Table 3 shows the chord recognition rates of these methods on the Beatles corpus. The recognition rates show that our SRC (Log PCP) method is not the highest, but slightly higher than many other methods. More specifically, the recognition rate we uses with LPCP features is 8.3% lower than the best method (KO1) in MIREX 2014. But it only needs some labeled fragments and doesn't train the temporal correlation of music.

Table 3. Comparison With the Previous Methods

		Recognition Rates
SRC methods	SRC (PCP)	67.0
	SRC (Log PCP)	73.9
MIREX 2014	Maksim Khadkevich, Maurizio Omologo (KO1)	82.2
	Chris Cannam, Matthias Mauch (CM3)	75.4
	Jean-Baptiste Rolland (JR2)	51.1

6. Conclusion

In this paper, we have presented a new machine learning model-SRC for chord recognition. In our method, the chord samples are treated as a dictionary *D* in formula (6).

Based on MIR development and combined our research, the following work is proposed. First, this paper only involved chord recognition which is a part of chord transcription task. Future work will consider adding recognition of more complex chords to our work. Second, in this work we take the effect of feature dimension into account in SRC, the results of Table 1. show that recognition rate increased along with numbers of features. We could add appropriate other features in the feature. Finally, we can use the SRC method when considering the temporal correlation of music.

Acknowledgements

This work was supported by the national Natural Science Foundation of China (Grant no. 61101225).

References

[1] M. McVicar, R. S.-Rodríguez, Y. Ni and T. De Bie, “Automatic chord estimation from audio: A review of the state of the art, Audio, Speech, and Language Processing”, IEEE/ACM Transactions on vol. 22, **(2014)**, pp. 556-575,.
[2] T. Fujishima, “Realtime chord recognition of musical sound: a system using common lisp music”, in Proc. ICMC, 1999, Beijing China , October 22-28 **(1999)**.
[3] M. A. Bartsch and G. H. Wakefield, “Audio thumbnailing of popular music using chroma-based representations. Multimedia”, IEEE Transactions on vol. 7, **(2005)**, pp. 96-104.
[4] J. C. Brown, “Calculation of a constant Q spectral transform”, The Journal of the Acoustical Society of America, vol. 89, **(1991)** , pp. 425.
[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, “Robust face recognition via sparse representation, Pattern Analysis and Machine Intelligence”, IEEE Transactions on, vol. 31, **(2009)**, pp. 210-227.
[6] K. Lee, “Automatic chord recognition from audio using enhanced pitch class profile”, Proc. of the Intern. Computer Music Conference, (ICMC), New Orleans, USA, July 23-28 **(2006)**.
[7] E. Gómez, P. Herrera, and B. Ong, “Automatic tonal analysis from music summaries for version identification.”, Audio Engineering Society Convention 121, San Francisco, CA, USA, October 5-8, **(2006)**.
[8] A. Sheh and D. P. Ellis, “Chord segmentation and recognition using EM-trained hidden Markov models”, ISMIR 2003, Maryland, USA, October 26-30, **(2003)**,
[9] J. P. Bello and J. Pickens, “A Robust Mid-Level Representation for Harmonic Content in Music Signals”, in ISMIR, **(2005)**, September 11-15, London, UK.
[10] M. Muller, S. Ewert and S. Kreuzer, Making chroma features more robust to timbre changes”, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Taipei, Taiwan 2009, April 19-24 **(2009)**.
[11] D. P. Ellis, "Beat tracking by dynamic programming," Journal of New Music Research, vol. 36, **(2007)** , pp. 51-60.
[12] E. J. Candès, “Compressive sampling”, Proceedings of the International Congress of Mathematicians, Madrid, Spain, August 22-30 **(2006)**.

[13] D. L. Donoho, "Compressed sensing, Information Theory", IEEE Transactions on vol. 52, **(2006)**, pp. 1289-1306.
[14] C. Harte, M. B. Sandler, S. A. Abdallah and E. Gómez, "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations", ISMIR, London, UK, September 11-15 **(2005)**.
[15] O. Lartillot, P. Toiviainen and T. Eerola, "A matlab toolbox for music information retrieval. Data analysis, machine learning and applications", Springer, **(2008)**, pp. 261-268.
[16] Y. Ni, M. McVicar, R. Santos-Rodriguez and T. De Bie, "An end-to-end machine learning system for harmonic analysis of music", Audio, Speech, and Language Processing", IEEE Transactions on vol. 20, **(2012)**, pp. 1771-1783.
[17] M. Khadkevich and M. Omologo, "MIREX 2014:TIME-FREQUENCY REASSIGNED FEATURES FOR AUTOMATIC CHORD RECOGNITION", Music Information Retrieval Evaluation eXchange (MIREX), Taipei, Taiwan, October 27-31 **(2014)**.
[18] C. Cannam, E. Benetos, M. Mauch, M. E. Davies, S. Dixon and C. Landone, "MIREX 2014: VAMP PLUGINS FROM THE CENTRE FOR DIGITAL MUSIC. Music Information Retrieval Evaluation eXchange (MIREX)", Taipei, Taiwan, October 27-31 **(2014)**.
[19] J.-B. Rolland, "CHORD DETECTION USING CHROMAGRAM OPTIMIZED BY EXTRACTING ADDITIONAL FEATURES", Music Information Retrieval Evaluation eXchange (MIREX), , Taipei, Taiwan, October 27-31 **(2014)**.

Authors

Zhongyang Rao received the bachelor's degree in 2001, from the Tianjin Polytechnic University, Tianjin, China, master's degree in 2004 from Tianjin University. From 2004 to 2008 he was a engineer at the National Ocean Technology Center, Tianjin, China. He moved to Shandong Jiaotong University in 2008, he works as an associate professor at Shandong Jiaotong University. His research focuses on Information acquisition and signal processing, including audio, video and image processing.

Xin Guan received the bachelor's degree in 2000, master's degree in 2003 from the Tianjin University, Tianjin, China, and the Ph.D. degree in 2008 from Tianjin University. From 1977 to 1990 he was a Lecturer at Tianjin University. From 2004 she works as a teacher at Tianjin University. Her research focuses on pattern recognition, audio and image professing.

Jianfu Teng received the bachelor's degree in 1977, master's degree in 1983 from the Tianjin University, Tianjin, China, and the Ph.D. degree in 1988 from King's College, University of London U.K.. From 1977 to 1990 he was a Lecturer at Tianjin University. From 1990 he works as a professor at Tianjin University. His research focuses on theory and design of the filter, fully integrated filter design, synthesis of active network and signal processing. He is a senior members of Chinese institute of electronics.