International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.11 (2016), pp.189-198
http://dx.doi.org/10.14257/ijmue.2016.11.11.16

Auto Chord Recognition Based on Sparse Representation Classification and Viterbi Algorithm

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

[1*]*School of Electronic Information Engineering, Tianjin University, Tianjin, China*
guanxin@tju.edu.cn
[2]*School of Information Science & Electric Engineering, Shandong Jiaotong University, Ji'nan, China*
yaozhongyang@sohu.com

Abstract

In this paper, a machine-learning approach called Sparse Representation Classification(SRC) Viterbi Algorithm is proposed for automatic chord recognition in music. We extracted Pitch Class Profile(PCP) features or Log PCP 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 MIREX'09 dataset including the Beatles corpus. Our method is also compared with various methods that entered the Music Information Retrieval Evaluation exchange (MIREX) in 2013 and 2014. Experimental results demonstrate that our method has good accuracy rate in recognizing signal chord and has fewer train data.

Keywords*: Chord Recognition; Sparse Representation Classification; Viterbi Algorithm; Log Pitch Class Profile*

1. Introduction

A musical chord can be defined as a set of notes played simultaneously. A succession of chords over time forms the harmony core in a piece of music. Hence compactly representing the overall harmonic content and structure of a song often starts with labeling every chord in it. Automation of chord labeling is also called chord recognition, which finds many applications such as music segmentation, music similarity identification, and audio thumb nailing [1-2]. For these reasons and others, automatic chord recognition has been one of the main fields of interest in musical information retrieval (MIR) in the last few years.

The features used in chord recognition may differ from a method to another but are in most cases variants of the Pitch Class Profile (PCP) introduced by Fujishima [3]. PCP is also called chroma vector, which is often a 12-dimensional vector, whose each component represents the spectral energy or salience of a semi-tone on the chromatic scale regardless of octave. The calculation of an audio recording into a chroma representation is based either on the short-time Fourier transform (STFT) in combination with binning strategies [4] or on the constant Q transform (CQT) [5]. The succession of these chroma vectors over time is often called chromagram and gives a good representation of the musical content of a piece.

The second part of the chord recognition is the chord labeling of each chord. Our chord recognition system is based on the sparse representation-based classification (SRC) [6], 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

ISSN: 1975-0080 IJMUE

representations. This classification has been applied into many applications and achieved perfect performance. If it uses the transitions between chords, SRC could easily be incorporated into Viterbi algorithm. This is the first time that we apply SRC and Viterbi algorithm into chord recognition. Experiments demonstrate that its perfect discrimination capability compared with some other classifications.

The remainder of this paper is organized as follows: Section 2 reviews previous the state-of-the-art methods on this area; Section 3 gives a description of our construction of the feature vector; Section 4 detailedly describes our sparse approach and Viterbi algorithm; Section 5 gives results on MIREX'09 dataset and a comparison with the other methods; In Section 6, we draw conclusions and directions for future work suggested.

2. Related Work

In audio chord estimation, it mainly includes the feature extraction, modelling techniques, evaluation strategies and so on. Many features used, such as non-negative least squares(NNLS) [7], chroma DCT-reduced log pitch(CRP) [8], loudness based chromagram(LBC) [9], Mel PCP(MPCP) [10]. For auto chord analysis, the most popular feature is a chromagram, also known as chroma vectors or Pitch Class Profile (PCP). In [3], Fujishima developed a real-time chord recognition system, where he derived a 12-dimensional pitch class profile from the DFT of the audio signal, and performed pattern matching using the binary chord type templates. Lee also used binary chord templates, this time for the 24 major/minor triads [11]. He introduced a new input feature called Enhanced Pitch Class Profile (EPCP) using the harmonic product spectrum. Gómez and Herrera used 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 [12].

In modelling techniques, it usually uses the templates-fitting methods [3, 13-15]. Besides templates-fitting methods, it is widely used machine-learning methods such as hidden Markov Model (HMM)[16-20] and DBNs(Dynamic Bayesian Networks)[7, 9] for this recognition process.

In our auto chord recognition method, like most of the methods, it is composed of extracted features and recognition chord process.

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 are 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[8] 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 [19] 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 gets 36-bin chromagram of each frame. 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:

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

(3)beat-synchronous segmentation. In our system, it use the beat tracking with dynamic programming method proposed by Daniel P.W. Ellis [21]. This approach has been found to work very well 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:

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

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

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.

As can be seen in Figure 1, the left picture shows a PCP of C major triad. The right one shows its Log PCP, as we can see, the strongest peaks are found at C, E, and G, since C major triad comprises three notes at C(root), E(third), and G(fifth). From the Figure 1, it can see that LPCP is clear than PCP.

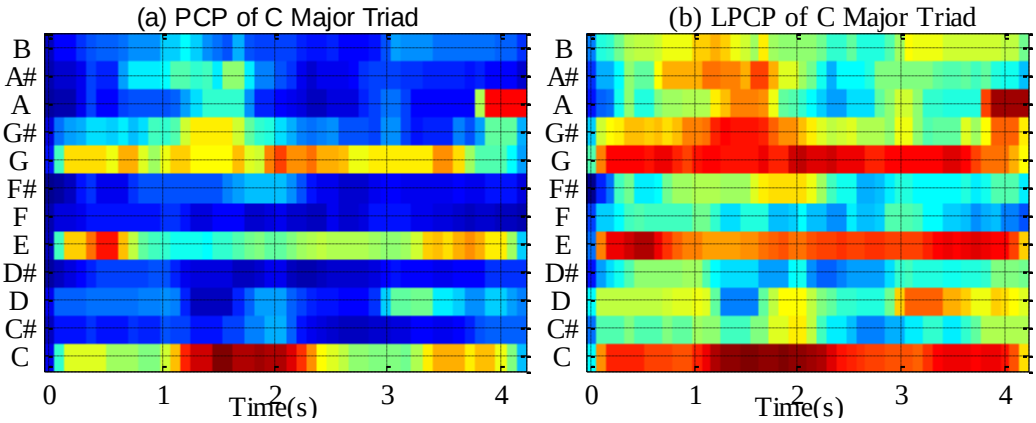

Figure 1. PCP and LPCP of C Major Triad

4. Feature Vectors

In our chord recognition method, the system includes two sections: (1) Sparse representation-based classification (SRC); (2) Viterbi algorithm. If it uses labeled musical fragments, then the system uses SRC method and only relies solely on frame-wise classification. The method doesn't need amount of training data. If it has amount of training data, the system can add Viterbi algorithm by using transitions between chords to recognize chords.

4.1 Sparse Representation-based Classification

First, if it selects matrix $W \doteq [W_1, W_2, \cdots W_K] = [c_{1,1}, c_{1,2}, \cdots c_{K,n_k}] \in \square^{M \times N}$ by collecting training PCP features of all K chord classes, where M is the dimension of the feature set and N is the Number of the samples. Chord type $i \in [1, K]$ contains n_i samples, its chromagram denoted by $[c_{i,1}, c_{i,2}, \cdots c_{i,n_i}]$. For a given PCP feature of test sample y from subject chord i, can be rewritten in terms of all training samples as:

$$y = Wx_0 \in \mathbf{R}^M \tag{6}$$

where $x_0 = [0, \cdots ..., 0, \alpha_{i,1}, \alpha_{i,2}, \cdots ..., \alpha_{i,n_i}, ..., 0]^T \in \mathbf{R}^N$ is a coefficient vector whose entries are mostly zero except the values corresponding to the i-th class.

As the entries of the vector x_0 encode the identity of the test sample y, it is tempting to attempt to obtain it by solving the linear system Eq.(6).

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 Eq.(6) is equivalent to the following ℓ^1-minimization problem:

$$\hat{x}_1 = \arg\min \|x\|_1 \ \textit{subject to}\ y = Wx \tag{7}$$

Actually, noise and modeling error may lead to small nonzero entries associated with multiple object classes. For each class i, one can approximate the given test sample y as $\hat{y}_i = W\delta_i(\hat{x}_1)$, where $\delta_i : \mathbf{R}^N \to \mathbf{R}^N$ is the characteristic function which selects the coefficients associated with the i-th class. We then calculate the residual between y and $\hat{y}_i$:

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

Finally, we classify y based on these approximations by assigning it to the object class that minimizes the residual, as follows:

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

In our system two chord types are used, *i.e.*, major and minor, and 12 for each chord type. One each for all 12 members of the chromatic scales: C Major, C minor, C# Major, C# minor... A# Major, A# minor, B Major, B minor. Each class contains 50 labeled musical fragments which select from the Beatles albums. And the number of labeled musical fragments is 1200. The training dictionary W is composed of 1200 labeled musical fragments' PCP feature. The given test sample is PCP feature of musical fragment to estimate chord.

4.2 Sparse Representation-based Classification

In SRC method, it uses the residuals $r_i(y)$ to recognize the chord. The method recognizes the chord on frame-wise classification. If it uses transitions between chords, it can improve the recognition rates of chord. Our system uses the Viterbi algorithm. Suppose the system has hidden N states, and we denote each state as $S_i, i \in [1:K]$. The observed events are $Q_t, t \in [1:T]$. The current observed events $Q = \{Q_1, Q_2 \ldots, Q_T\}, t \in [1:T]$. A_{ij} represents the probability chord S_i jump to chord S_j. At an arbitrary time point t, for each of the states S_i, a partial probability $\delta_t(S_i)$ is defined to indicate the probability of the most probable path ending at the state S_i, given the current observed events $Q_1, Q_2 \ldots, Q_t$: $\delta_t(S_i) = \max_j(\delta_{t-1}(S_j) \cdot A(S_j, S_i) \cdot P(Q_t \mid S_i))$. Here, we assume that we already know the probability $\delta_{t-1}(S_j)$ for any of the previous states S_j at time $t-1$. $P(Q_t \mid S_i)$ is the current observation probability. After having all the objective probabilities for each state at each time point, the algorithm seeks from the very end backwards to the beginning to find the most probable path of states for the given sequence of observation events $\psi_t(i) = \arg[\max_{1 \le j \le N}(\delta_{t-1}(S_j) \cdot A(S_j, S_i))]$. Where $\psi_t(i)$ indicates which state is the most optimal state at time t based on the probability computed in the first stage.

The Viterbi algorithm is as follows:

Algorithm 2: Vertibi algorithm

1: Initialization: $\delta_t(S_i) = \Pi_i P(Q_1 \mid S_i), \ \psi_t(i) = 0, \ 1 \le i \le K$.

2: Recursion: $\delta_t(S_i) = \max_{1 \le j \le N}(\delta_{t-1}(S_j) \cdot A(S_j, S_i)) \cdot P(Q_t \mid S_i), \ 2 \le t \le T$,

$\psi_t(i) = \arg[\max_{1 \le j \le N}(\delta_{t-1}(S_j) \cdot A(S_j, S_i))]$.

3: Termination: $q_T^* = \arg\max_{1 \le i \le N}[\delta_t(S_i)]$, $P^* = \max_i[\delta_t(S_i)]$.

4: Path Backtracking: $q_t^* = \psi_{t+1}(q_{t+1}^*) \ \ t = T-1, T-2 \ldots 1$.

In our method, the initialization observation probability Π_i is equal to 1/24. The observed events are PCP features y_t, where y_t is the PCP feature of tth frame. And current observation probability is $r_i(y_t)$ and replaces the $P(Q_t \mid S_i)$ in Viterbi algorithm. S_i represents the chord $i \in [1:K]$, where K is the number of chord and set to 24.

Figure 2 is the comparison of ground truth chord and estimated chord of the Beatles song Misery. In the top figure, it only uses the SRC method to recognize the chord and the bottom uses SRC and Viterbi decoding. The ground truth chord is represented in pink and the estimated chord labels are in blue. From the Figure 2 it can see that the estimation is more stable when it uses the Viterbi than without.

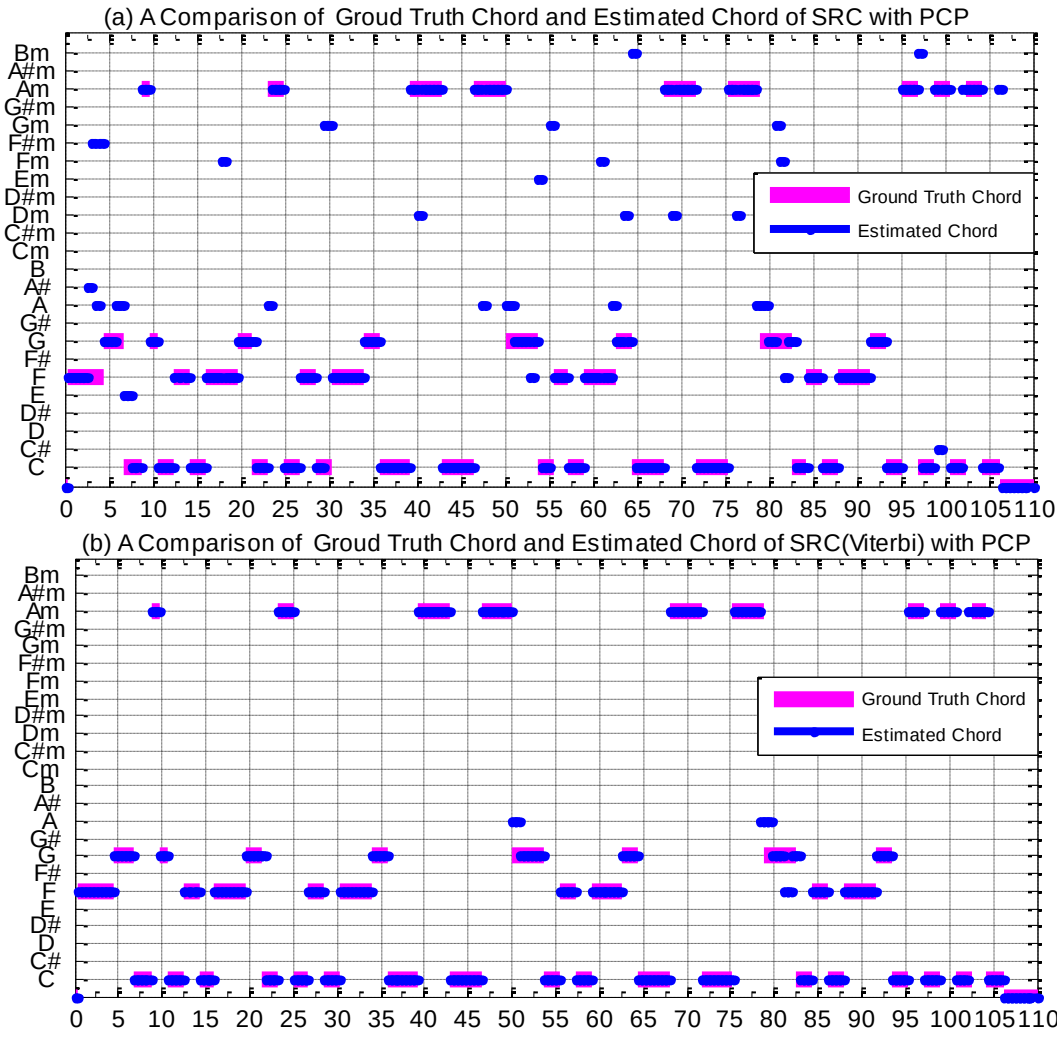

Figure 2. A Comparison of Ground Truth Chord and Estimated Chord

5. Evaluation

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.

To evaluate the quality of an automatic transcription, a transcription is compared to ground truth created by one or more human annotators. According to \citePauwels2013 the chord symbol recall (CSR) is a good metric to evaluate performances of good results.

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). In the paper, recognition rate and CSR are equivalent for a song. And recognition rate and WCSR is equivalent for a given corpus or dataset.

Our method is also compared to the following methods that entered MIREX 2013 and MIREX 2014.

MIREX 2013:

·CB4 and CB3: Taemin Cho & Juan P. Bello[22]

·KO1and KO2: Maksim Khadkevich & Maurizio Omologo[23]

·NMSD1 and NMSD2: Yizhao Ni, Matt Mcvicar, Raul Santos-Rodriguez[24]

·CF2 : Chris Cannam, Matthias Mauch, Matthew E. P. Davies[25]
·NG1 and NG2: Nikolay Glazyrin[26]
·PP3 and PP4: Johan Pauwels & Geoffroy Peeters[27]
·SB8: Nikolaas Steenbergen & John Ashley Burgoyne[28]
MIREX 2014:
·KO1: Maksim Khadkevich & Maurizio Omologo[29]
·CM3: Chris Cannam, Matthias Mauch[30]
·JR2: Jean-Baptiste Rolland[31]

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

Figure 3 presents the results with SRC method on the MIREX'09 dataset. In the Figure 3, the method SRC plus Viterbi with LPCP feature has better performance than others.

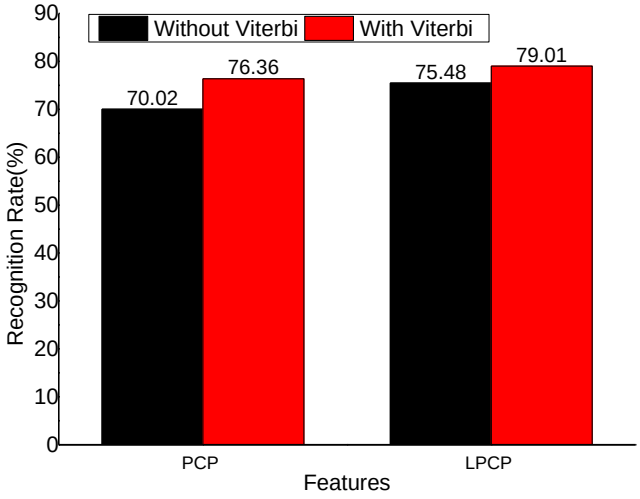

Figure 3. A Comparison of SRC' Recognition Rates with Different Features

Figure 4 presents the results obtained by these chord recognition systems on the MIREX'09 dataset. In the Figure 4, the SRC method is referred to SRC+Viterbi method. The rates of recognition show that our SRC method with LPCP features has high recognition rate than most of state of the art. More specifically, our SRC(LPCP) approach is lower 3.1 percent than the best score method (KO1) in MIREX 2014. But our SRC method has fewer train data than previous methods.

6. Conclusion

In this paper, we have presented a new machine learning model SRC+Viterbi for chord estimation. In comparison with previous work, our new approach presents good performance. The main ingredients of our new approach are calculation of PCP features, sparse representation classification and Viterbi algorithm. Our method need fewer train data and parameters than most of HMM method.

As for perspective, we envisage the following lines of work. First, this paper only involved maj-min chord estimation which is a part of chord transcription task. Future work will consider adding recognition of more complex chords to our work. This will find many applications in the field of MIR such as song identification, query by similarity or structure analysis. Second, in this work we take the effect of different features in SRC. So Improving PCP features to make them more suitable for chord recognition have a long way to go. Finally, we can use the SRC method when the audio music is a piece of fragment not a whole song.

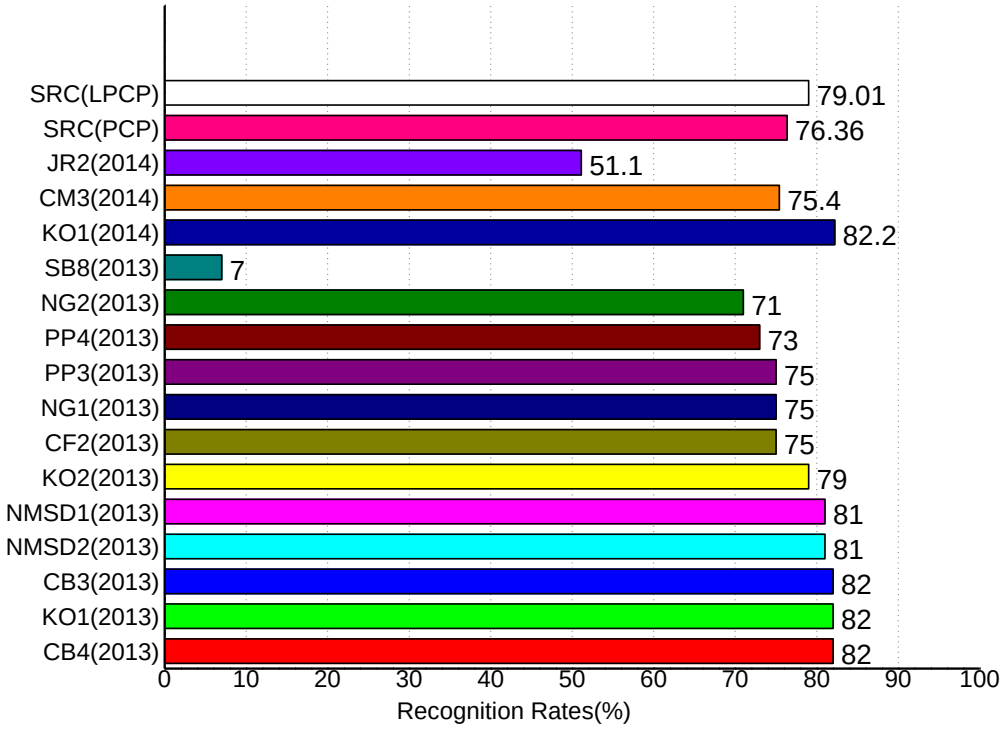

Figure 4. A Comparison of SRC' Recognition Rates With Other Methods

Acknowledgments

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

References

[1] Y. S. Wang, "Unsupervised Bayesian Musical Key and Chord Recognition", Dissertations & Theses – Gradworks, vol. 14, no. 3, **(2014)**, pp. 115-124.
[2] M. Mcvicar, R. S. Rodriguez, Y. Ni and T. D. Bie, "Automatic Chord Estimation from Audio: A Review of the State of the Art", IEEE/ACM Transactions on Audio Speech & Language Processing, vol. 22, no. 2, **(2014)**, pp. 556-575.
[3] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music", in Proceedings of the International Computer Music Conference, Beijing, China, **(1999)**.
[4] M. A. Bartsch and G. H. Wakefield, "Audio Thumbnailing of Popular Music Using Chroma-Based Representations", IEEE Transactions on Multimedia, vol. 7, no. 1, **(2005)**, pp. 96-104.
[5] J. C. Brown, "Calculation of a constant Q spectral transform", The Journal of the Acoustical Society of America, vol. 89, no. 1, **(1991)**, pp. 425-434.
[6] 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, no. 2, **(2009)**, pp. 210-227.
[7] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords", in Proc. 11th Int. Soc. Music Inf. Retrieval Conf., **(2010)**, pp. 135-140.
[8] M. Müller, S. Ewert and S. Kreuzer. "Making chroma features more robust to timbre change", in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Taipei, Taiwan, **(2009)**.
[9] 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, no. 6, **(2012)**, pp. 1771-1783.
[10] F. Wang and X. Zhang, "Research on CRFs in Music Chord Recognition Algorithm", Journal of Computers vol. 8, no. 4, **(2013)**, pp. 1017.
[11] K. Lee, "Automatic chord recognition from audio using enhanced pitch class profile", in Proc. of the International Computer Music Conference, **(2006)**.
[12] E. Gómez, P. Herrera, and B. Ong, "Automatic tonal analysis from music summaries for version identification", in Audio Engineering Society Convention, San Francisco, CA, USA, vol. 121, **(2006)**.

[13] C. Harte and M. Sandler, "Automatic chord identifcation using a quantised chromagram", in Audio Engineering Society Convention, Barcelona, Spain, vol. 118, **(2005)**.
[14] L. Oudre, Y. Grenier, and C. Févotte, "Template-based Chord Recognition: Influence of the Chord Types", in ISMIR, **(2009)**.
[15] T. Cho and J. P. Bello, "A Feature Smoothing Method for Chord Recognition Using Recurrence Plots", in Music Information Retrieval Evaluation eXchange(MIREX 2011), Miami, Florida, USA, **(2011)**.
[16] A. Sheh and D. P. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models", in ISMIR 2003, Library of Congress, Washington, D.C., USA, and Johns Hopkins University, Baltimore, Maryland, USA, **(2003)**.
[17] H. Papadopoulos and G. Peeters, "Large-Scale Study of Chord Estimation Algorithms Based on Chroma Representation and HMM", Content-Based Multimedia Indexing", 2007. CBMI '07. International Workshop on, **(2007)**, pp. 53-60.
[18] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio", Audio, Speech, and Language Processing, IEEE Transactions on, vol 16, no. 2, **(2008)**, pp. 291-301.
[19] J. P. Bello and J. Pickens, "A Robust Mid-Level Representation for Harmonic Content in Music Signals", in ISMIR, London, UK, **(2005)**.
[20] R. Scholz, E. Vincent, and F. Bimbot, "Robust modeling of musical chord sequences using probabilistic N-grams", in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, **(2009)**.
[21] D. P. Ellis, "Beat tracking by dynamic programming", Journal of New Music Research, vol. 36, no. 1, **(2007)**, pp. 51-60.
[22] T. Cho and J. P. Bello, "MIREX 2013: Large Vocabulary Chord Recognition System using Multi-band Features and a Multi-stream HMM", in Music Information Retrieval Evaluation eXchange (MIREX). Curitiba, PR, Brazil, **(2013)**.
[23] M. Khadkevich and M. Omologo, "Time-frequency reassigned features for automatic chord recognition", in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, **(2011)**.
[24] Y. Ni, M. McVicar, R. S. Rodriguez, and T. De Bie, "Harmony Progression Analyzer For Mirex 2013", in Music Information Retrieval Evaluation eXchange (MIREX), Curitiba, PR, Brazil, **(2013)**.
[25] C. Cannam, E. Benetos, M. Mauch, M.E.P. Davies, S. Dixon, C. Landone, K. Noland, and D. Stowell, "MIREX 2015: Vamp Plugins from the Centre for Digital Music", in Music Information Retrieval Evaluation eXchange (MIREX), Malaga, Spain, **(2015)**.
[26] N. Glazyrin, "Audio Chord Estimation using Chroma Reduced Spectrogram and Self-Similarity", in Music Information Retrieval Evaluation eXchange (MIREX), Curitiba, PR, Brazil, **(2013)**.
[27] J. Pauwels and G. Peeters. "The Ircamkeychord Submission for Mirex 2013", in Music Information Retrieval Evaluation eXchange (MIREX), Curitiba, PR, Brazil, **(2013)**.
[28] N. Steenbergen and J. A. Burgoyne, "MIREX 2013:Joint optimization of an hidden markov model-neural network hybrid chord estimation", in Music Information Retrieval Evaluation eXchange (MIREX), Curitiba, PR, Brazil, **(2013)**.
[29] M. Khadkevich and M. Omologo, "Time-frequency reassigned features for automatic chord recognition", in Music Information Retrieval Evaluation eXchange (MIREX), Taipei, Taiwan, **(2014)**.
[30] C. Cannam, E. Benetos, M. Mauch, M. E. Davies, S. Dixon, C. Landone, K. Noland, and D. Stowell, "MIREX 2014: Vamp Plugins from The Centre For Digital Music", in Music Information Retrieval Evaluation eXchange (MIREX), Taipei, Taiwan, **(2014)**.
[31] J. B. Rolland, "Chord detection using chromagram optimized by extracting additional features", in Music Information Retrieval Evaluation eXchange (MIREX), Taipei, Taiwan, **(2014)**.

