SUBMISSION TO MIREX 2012 AUDIO MUSIC MOOD CLASSIFICATION AND AUDIO US POP MUSIC GENRE CLASSIFICATION

Deshun Yang, Xiaoou Chen and Liqun Peng
Institute of Computer Science and Technology, Peking University
{yangdeshun,chenxiaoou,pengliqun}@pku.edu.cn

ABSTRACT

We use an audio keyword based method for audio music classification, in which music clips are converted into sequences of audio words, and keywords are selected. For classification, music clips are represented in binary keyword occurrence vectors, and SVM is used to build the classification model.

1. INTRODUCTION

Our system consists of three parts. The first is the modeling of audio words, the second keyword selection and the third the classification model. Models of audio words have been built through unsupervised methods, on a large corpus of pop music. For keyword selection, Several well performing supervised methods often used in text classification are tried. Music clips are represented by binary vectors of keyword occurrences, and the music classification model is built with SVM.

2. MODELING AUDIO WORDS

We have devised more than 10 categories of audio words, which, we expect, will describe as many as possible relevant aspects of music signal.

A word category defines its own set of audio features, and some distinctive patterns can be found in the feature space of the category. Those distinctive patterns are referred to as distinctive words. We represent a pattern with a unique vector, and we obtain the pattern vectors of a category through clustering analysis of the feature vectors of that category collected from a large corpus of real-world music examples. We use k-mean for clustering analysis.

In the following sub sections, we list some of the word categories and show what their word features are.

2.1 Words of CQ Spectrogram

Constant Q transformation[1] is performed on frames of music signals. The resulted spectrogram is first divided along the frequency axis into 7 sub bands, each corresponding to an octave, and then, each octave is cut along the time line into equal-sized units. We extract two sets of features from each unit and represent a unit with two vectors consisting of different set of features.

2.2 Words of Area of Moments

We divide the CQ spectrogram of a music clip along the time line into equal-sized rectangular units and calculate a feature vector(with 15 components) of area of moments[2]on each unit.

2.3 Words of LPC

On the consecutive frames of a music signal, we compute 13-dimension feature vectors of Linear Prediction Coding(LPC).

2.4 Words of Timbre

We extract 25-dimension vectors of spectral shape features on Constant Q spectrum.

2.5 Words of Rhythm

Two kinds of rhythm words are designed. The first kind of words is calculated on time signal. Onset sequences are first calculated on windows of a signal, and then DFT is performed on those sequences. Finally, feature vectors are extracted from the DFT results.

The other kind of rhythm words is based on CQ spectrogram. The spectrogram of a signal in a sliding window is divided into several sub bands, and onset sequences are calculated separately for each sub band. Then, all the onset sequences of the sub bands are combined to form the final onset sequence. Finally, autocorrelation coefficients are computed on the final onset sequence, and the coefficients are used as rhythm features.

2.6 Words of AR

On the frames of a time signal, we compute 12-dimension feature vectors describing the signal's attack/rest envelop.

2.7 Words of Chroma

We extract 36-dimension vectors of chroma features on Constant Q spectrums of frames of music signals.

3. AUDIO KEYWORD SELECTION

Given a set of training music examples, we extract from the examples feature vectors corresponding to each word category and, according to the pattern vectors of the category, match each vector to a single pattern(i.e. word), thus converting the music clips into sequences of audio words.

Document Frequency(DF), an unsupervised feature selection method, and Information Gain(IG), a supervised feature selection method, are combined to carry out keyword selection. On the word sequences of a set of training music examples(with class labels), the DF and IG scores of all the words are calculated and normalized to [0, 1] separately, and the sum of the normalized DF score and IG score of a word is used as the final score of the word. From about 36,000 words, we choose the 2000 words which have the largest scores as audio keywords.

4. CLASSIFICATION MODEL

4.1 Music Representation

Given a piece of music audio as a training or testing example, we extract feature vectors corresponding to each keyword category and, based on the pattern vectors of the category, match each vector to a single pattern(i.e. keyword), thus converting the music piece into sequences of selected audio keywords.

From the keyword sequences of a music, we calculate a 2000-dimension binary vector of keyword occurrences and use the vector as the final representation of the music.

4.2 Classification Algorithm

We use libSVM[3], an implementation of Support Vector Machine(SVM) algorithm for building the music classification model. The Poly kernel is used.

5. ACKNOWLEGEMENTS

This work is supported by the Natural Science Foundation of China (Multi-modal Music Emotion Recognition technology research No.61170167) & Beijing Natural Science Foundation (Multimodal Chinese song emotion recognition)

6. REFERENCES

[1] J. C. Brown, Calculation of a Constant Q Spectral Transform, *Journal of the Acoustical Society of America*, 1991.

[2] I. Fujinaga. Adaptive optical music recognition. PhD thesis, McGill University, 1997.

[3] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.