IMAGE FEATURES IN MUSIC STYLE RECOGNITION

Kamil Behúň
Master Degree Programme (1), FIT BUT
E-mail: xbehun03@stud.fit.vutbr.cz

Supervised by: Michal Hradiš
E-mail: ihradis@fit.vutbr.cz

Abstract: This work presents a novel approach to music style recognition inspired by feature extraction techniques used in image classification. To be able to utilize the image classification techniques, the 1D sound signal is transformed to its 2D representation - to a Mel-frequency spectrogram. Small local areas of the spectrum are represented by SIFT descriptors from which Bag of Words (BOW) representation of a whole signal is constructed. The BOW feature vectors are classified by Support Vector Machine classifier. The proposed approach was tested on publicly available music recognition data set GTZAN and achieving superior results over existing approaches.

Keywords: Music genre recognition, Mel-frequency spectrum, Local features, SIFT descriptors, Bag of Words, Support Vector Machine

1 INTRODUCTION

Automatic music genre recognition has potential applications in online, as well as personal music databases. It can provide music genre information where it is not available, thus support navigation and searching in such music database. Furthermore, the problem of music genre recognition is related to the task of automatically suggesting suitable songs for users based on their particular taste or personalised song rating.

The following section shortly desribes commonly used methods for music genre recognition. Section 3 describes our approach to this task and section 4 contians desription of achieved results.

2 PREVIOUS WORK

The music genre recognition is is in most cases as a pattern recognition task, which consists of two parts. These parts are feature extraction and classification.

According to Tzanetakis at al. [10], features for music recognition can be divided into three basic classes. These classes are Pitch content features, Rhythmic content features and Timbral texture features. The Pitch content features characterize audio signals in terms of energy of different frequency bands. The Rhythmic content features represent rhythmic structure of the music [10]. They can be for example computed as a 24-band psycho-acoustically modified spectrogram which reflects human loudness sensation [5]. The Timbral texture features should exhibit properties related to general timbre of the sound. They are based on a short time Fourier transform and they are calculated on short-time frames of a sound, for example Spectral Centroid, Spectral Rolloff, Mel-Frequency Cepstral Coefficients (MFCC), etc.

Additionally, many approaches that are outside the above described groups exist. For example, [7] describes a feature extraction technique called Multilinear Subspace Analysis Techniques, e.g. Multilinear principal components analysis (MPCA). These techniques are compute from tensors and their linear counterparts are NMF, SVD and PCA.

Any type of classifier can be used for music genre recognition. In the literature, the most commonly used classifier is Support Vector Machine (SVM) [5]. Other classification methods used for music genre recognition are Gaussian mixture models [10], K-nearest neighbor classifier [10], etc.

3 METHOD

As mentioned earlier, music genre recognition has two main parts - feature extraction and classification. This work presents a novel approach of feature exctraction inspired by feature extraction techniques used in image classification. Common approach in image classification [8] is to represent local parts of an image by a high-dimensional descriptor [6]. To be able to use the local feature techniques from image classification, the 1D sound signal has to be transformed to a 2D representation. We chose a Mel-frequency spectrogram. In order to be able to use the existing image local feature methods, dynamic range and contrast of the spectrograms were reduced by

$$x = \left(\frac{\log(e+1)}{maxV}\right) * 255, \tag{1}$$

where e is a value from the original Mel-frequency spectrum, $maxV$ is logarithm of the maximal value in the original Mel-frequency spectrogram and x is the resulting value in the lower dynamic range spectra. The transformation from Equation 1 assures that the resulting values are in interval $< 0,255 >$ and they correspond to how humans perceive sound intensity (perception of acoustic intensity is logarithmic).

The spectrograms were than handled as images and local features were extracted from them. As the spectrograms do not exhibit any stable and distinct areas which could be detectected by interest region detectors [6], we sampled the spectorgrams on a regular grid with cell size 8×8 pixels. SIFT descriptor was computed from the sampled small circular areas.

To create a feature vector for a whole audio recording, the extracted local features are aggregated to a Bag of Word (BOW) representation. In order to obtain the BOW representation, local features are first translated to visual words by codebook transform. We used the k-means algorithm with Euclidean distance to obtain the set of prototypes which constitute the codebook - cluster centers become the prototypes. In the experiments, we used 4096 codewords (clusters).

Experiments were performed with Support Vector Machine classifier (SVM) and Gaussian kernel

$$K(x,x') = exp\left(-\gamma \|x - x'\|_2^2\right). \tag{2}$$

Optimal value of the SVM regularization parameter C and the Gaussian kernel scale γ were estimated by grid search with 10-fold cross-validation with stratified sampling of training dataset.

4 EXPERIMENTS AND RESULTS

Experiments were done on GTAZAN genre collection [9], which contains 10 genres. Following the experimental setup used in [7, 10], stratified tenfold cross validation was employed to estimate performance in the experiments. That means, each training set contained 900 tracks (9 parts), testing set contained 100 tracks (1 part) from GTAZAN and the parts were gradually changed in the experiment.

Experiments were done for local features of size 8×8, 16×16 and 32×23 pixels to estimate the optimal size. Small sizes are aimed on detail and bigger sizes are aimed on context. The best classification accuracy (86.4 %) was achieved for 32×23 size of local features ($8 \times 8 =$ cl. a. 83.4 % and $16 \times 16 =$ cl. a. 84.2 %). As can be seen in Table 1, the proposed approach provides superior performance compared to other published results.

Approach (features + classifier)	Classification accuracy
mel-spectrogram - SIFT 32×32 + SVM (this work)	**86.4**%
non-negative MPCA + SVM [7]	84.3%
aggregate features + AdaBoost [2]	82.5%
wavelet histograms + SVM [3]	78.5%
audio and symbolic features + SMV [4]	76.8%
many features + NTF [1]	75.0%
pitch, rhytmic and timbral features + GMM [10]	61.0%

Table 1: Classification accuracies achieved by our approach and other published approaches for GTZAN Genre collection.

5 CONCLUSIONS

The concluded experiments show that the proposed feature extraction provides state-of-the-art results in music genre recognition task. Specifically, its performance is superior to previously published results on the GTZAN dataset. Overall the achieved results are very promising and we plan to evaluate it in other classification tasks which process audio data. Moreover, the field of image classification provides a wide variety of features which could all be applied to spectrograms as well, which could possibly surpass the presented approach.

REFERENCES

[1] Emmanouil Benetos and Costas Kotropoulos. A tensor-based approach for automatic music genre classification. *Proc. 16th European Signal Processing Conf.*, 2008.

[2] James Bergstra and et al. Aggregate features and ADABOOST for music classification. *Machine Learning*, 65(2-3):473–484, December 2006.

[3] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. SIGIR '03, New York, NY, USA, 2003. ACM.

[4] Thomas Lidy and et al. MIREX 2007: Combining Audio And Symbolic Descriptors For Music Classification From Audio. In *MIREX 2007.*, Vienna, Austria, 2007.

[5] Thomas Lidy and et al. On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections. *Signal Processing*, 90(4):1032–1048, 2010.

[6] Krystian Mikolajczyk and et al. A Comparison of Affine Region Detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[7] Yannis Panagakis, Constantine Kotropoulos, and G. Gonzalo Arce. Non-Negative Multilinear Principal Component Analysis of Auditory Temporal Modulations for Music Genre Classification. *IEEE Transactions On Audio Speech And Language Processing*, 18(3):576–588, 2010.

[8] Cees G M Snoek and et al. The MediaMill TRECVID 2010 Semantic Video Search Engine. In *TRECVID 2010.*

[9] George Tzanetakis. Data sets [online], [cit. 2012-03-26]. <http://marsyas.info/download/data_sets>.

[10] George Tzanetakis, Student Member, and Perry Cook. Musical Genre Classification of Audio Signals. *Audio*, 10(5):293–302, 2002.