Music Genre Classification

The Voxstudy Group
T. Brundage, G. Gliner, Z. Jin, K. Wolf
Princeton University, Princeton, NJ 08540

GENNA, KEWOLF, TOBRUND, ZJIN @PRINCETON.EDU

Abstract

Musical genre classification provides a useful tool for automatic annotation in Music Information Retrieval. The problem has been studied extensively since the late 1990s and is featured as a standard challenge in the annual Music Information Retrieval Evaluation eXchange (MIREX) competition. In this work we tackle this problem by representation learning based on data analysis and state-of-art feature aggregating and selection techniques.

1. Introduction

Musical genre classification has been a popular topic in Music Information Retrieval (MIR). Studying this problem provides insights and solutions to applications such as music search engines, recommendation systems and automatic annotation systems. The challenges associated with automatic music genre classification are twofold. First, the characteristics of a musical genre are subjective - some represent the cultural background (e.g. country, world, soul), the affordance or purpose (e.g. disco, holiday, etc.), the era (e.g. classic, 80s, etc.) or form (e.g. jazz, rock, etc.). There are also instances where a song may fit into multiple genres. For instance a song that is considered pop rock can be classified as either rock or pop. This ambiguity leads to poor distinction between genres, rendering classification a difficult task. Secondly, genre classification requires transforming an audio signal into a digital representation; compressing millions of audio samples into a vector with a limited number of dimensions may introduce significant loss of information. Alternatively, one may consider converting audio samples into a Musical Instrument Digital Interface (MIDI) and classifying based only on MIDI or jointly with audio (Cataltepe et al., 2007). The problem however is that polyphonic pitch detection and onset detection are not accurate enough to extract the MIDI representation from the audio signal. In this work, we focus on modeling a music representation that enables accurate genre classification. We develop a feature extraction, selection and classification method that achieve an accuracy above the current state-of-the-art published methods.

There are a limited number audio benchmarking collections available, several of which are freely released to the public. We use a benchmark dataset called GTZAN that consists of 1000 30-second audio clips categorized evenly into 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock). This dataset has been studied extensively in the genre classification literature (Li et al., 2003) and currently the best published methods achieve an accuracy of 93% (Panagakis & Kotropoulos, 2010). While the main-stream methods treat genre classification as either image classification (Wu & Jang, 2015) or data compression (Alam et al., 2015), we focus on feature analysis and representation learning from the musical aspect of the signal.

In the following sections, we first discuss previous work related to the genre classification problem in §2 then we follow with a discussion on feature extraction and feature selection in §3.

2. Related work

Automatic genre classification dates back to 1997 when Dannenberg et. al. applied machine learning to musical style classification for interactive performing systems (Dannenberg et al., 1997). In 2002, Tzanetakis and Cook at Princeton University set the milestone by employing a large set of musical features, including timbre features, rhythmic variations, etc. to achieve a 58% accuracy on a 10-genre classification task (Tzanetakis & Cook, 2002). Since then, automatic genre classification is a standard challenge in MIREX[1]. Many new methods have been developed in this area, including Adaboost classifiers (Bergstra et al., 2006) (82.5% accuracy on GTZAN), high-level musical features (McKay & Fujinaga, 2004) and Non-Negative Tensor Factorization (Panagakis et al., 2008) (78.2% accuracy on GTZAN). The current winning method in MIREX borrowed insights from computer vision (Costa et al., 2012; Alam et al., 2015). They treat the spectrogram as a texture image, extract visual patches such as local binary pat-

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

tern textures (LBP) and then aggregate them into a single descriptor using, for example, a histogram. The reported performance on GTZAN is 88.60% (Wu & Jang, 2015) using this approach. However, higher accuracies on GTZAN dataset exist using factorization methods such as compressive sampling (Chang et al., 2010) and non-negative matrix factorizations (Panagakis & Kotropoulos, 2010).

3. Methods

3.1. Feature Extraction

The 30-second song clips in our dataset are decomposed into frames, each representing a 20-millisecond window. Signal processing methods are applied to extract features for each frame independently, called frame-level features. Another set of algorithms are used to extract features for the entire song, which we call song-level features.

Frame-level features can be split into lower-level audio features and higher-level musical features. The audio features such as timbre and zero-crossing are extracted directly from the audio signals using spectral and Cepstral analysis. The musical features are extracted from the spectrogram representation of the signal that reflect music perfection. These features provide information related to chord progression, instrumentation and rhythmic patterns.

All features we use are depicted in Tables 1 and 2 along with their descriptions and dimensionality. Song-level features are described in Table 3. With the exception of MFCC, the features are extracted using the Matlab MIRToolbox (Lartillot & Toiviainen, 2007).

Table 1. Frame-Level Audio Features

Name (Dim)	Description
MFCC ($\mathbb{R}^{24\times n}$)	Mel-Frequency Cepstrum Coefficients A timbrel representation (character of a sound)
Chroma ($\mathbb{R}^{12\times n}$)	Captures the distribution of energy along pitches
Energy ($\mathbb{R}^{1\times n}$)	Total intensity of each frame
Zero-Crossing ($\mathbb{R}^{1\times n}$)	Indicator of the noisiness
Spectral Flux ($\mathbb{R}^{1\times n}$)	Difference between the spectrogram of each successive frame

Table 2. Frame-Level Music Features

Name (Dim)	Description
Roughness ($\mathbb{R}^{1\times n}$)	An estimation of sensory dissonance
Key Strength ($\mathbb{R}^{12\times n}$)	A score between -1 and 1 for each key candidate via a cross-correlation of the chroma
Onsets ($\mathbb{R}^{1\times n}$)	Determines whether there is an onset in the frame
HCDF ($\mathbb{R}^{1\times n}$)	Harmonic Change Detection Function The flux of the tonal centroid
Inharmonicity ($\mathbb{R}^{1\times n}$)	Estimates the amount of partials that are not multiples of the fundamental frequency

Table 3. Song-Level Features

Name	Description
Key	The estimated key for the song
Tempo	An estimate of the tempo by detecting periodicity's from onset detection

3.2. Quantization

Due to high dimensionality, frame-level features cannot be used directly for classification and need to be encoded into a global feature vector often called a descriptor. This step is called quantization in computer vision literature. Commonly used methods include the Bag-of-words model (BoW), multiple codebooks (MCB), and Gaussian Mixture Models (GMM). In this work, we use Fisher Vectors (FV) for aggregating frame-level features as it has achieved remarkable performance in recent image classification tasks (Snchez et al., 2013). FV are based on Fisher kernels that represent the distance between samples (i.e. data clouds) with respect to their generative model. Suppose we have two samples X and Y that are generated by the same random process determined by parameter set Θ. Then, the Fisher kernel is defined as

$$K_F(X,Y) = G_X^T F_\Theta^{-1} G_Y \tag{1}$$

where G_X is the score function of dataset X. The score function is defined as the gradient of the log likelihood: $G_X = \nabla_\Theta \log \mathcal{L}(X;\Theta)$ and F_Θ^{-1} is the Fisher Information matrix, defined as $\mathbb{E}_\mathbb{X}(G_X G_X^T)$. Since the Fisher Information matrix is positive semi-definite, using the Cholesky decomposition we have $F_\Theta^{-1} = \mathcal{F}^T\mathcal{F}$. A vector representing each dataset is defined as follows:

$$K_F(X,Y) = G_X^T \mathcal{F}^T \mathcal{F} G_Y = \mathcal{G}_X^T \mathcal{G}_Y \tag{2}$$

where $\mathcal{G}_X = \mathcal{F}G_X$ and $\mathcal{G}_Y = \mathcal{F}G_Y$. Replacing the score function with the likelihood, we can express a global descriptor for dataset X directly using

$$\mathcal{G}_X = \mathcal{F}\nabla_\Theta \log \mathcal{L}(X;\Theta). \tag{3}$$

The Fisher kernel is the natural kernel in the space of score functions. Since score functions measure the direction that a dataset X affects the model parameters (i.e. the gradient), the Fisher kernel represents the similarity of how

two datasets affect the model parameters. Using FV to aggregate the frame-level features, we obtain a song-level descriptor that is naturally measured by a linear kernel. In this work, we use GMM as the base-model for the Fisher vectors. The derivation and the resulting gradient formula can be found in (Snchez et al., 2013).

3.3. Encoding temporal information

Frame-level features do not encode temporal variation that may be critical to genre classification. In addition to using individual frames for quantization, we also investigate exemplar, a new way to include temporal variation in the recent speech recognition literature (Gemmeke et al., 2011). Exemplar is defined as spectrographic representations of speech spanning multiple time-frames of signal; the exemplar of frame-level features are 2k+1 consecutive frames concatenated together.

3.3.1. PCA

We apply PCA both to the matrix formed by collecting all Fisher vectors and the matrices of a single class of Fisher vectors. Given that principal components capture dimensions of greatest variance, we infer that if a subset of these principal components can yield better classification results than the full set, then some non-trivial fraction of the variance is meaningless and only adds extra noise to our signal. If however we observe a continuous drop in performance with a drop in captured variance, we can confirm that all components of each class of Fisher vectors are necessary.

References

Alam, Mohammad Rafiqul, Bennamoun, Mohammed, Togneri, Roberto, and Sohel, Ferdous. A confidence-based late fusion framework for audio-visual biometric identification. *Pattern Recognition Letters*, 52:65 – 71, 2015.

Bergstra, James, Casagrande, Norman, Erhan, Dumitru, Eck, Douglas, and Kgl, Balzs. Aggregate features and adaboost for music classification. *Machine Learning*, 65 (2-3):473–484, 2006.

Cataltepe, Zehra, Yaslan, Yusuf, and Sonmez, Abdullah. Music genre classification using midi and audio features. *EURASIP J. Appl. Signal Process.*, 2007(1):150–150, January 2007.

Chang, Kaichun K., shing Roger Jang, Jyh, and Iliopoulos, Costas S. Iliopoulos: music genre classification via compressive sampling. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR*, pp. 387–392, 2010.

Costa, Y. M. G., Oliveira, L. S., Koerich, A. L., Gouyon, F., and Martins, J. G. Music genre classification using lbp textural features. *Signal Process.*, 92(11):2723–2737, November 2012.

Dannenberg, Roger B., Thom, Belinda, and Watson, David. A machine learning approach to musical style recognition. In *Proc. International Computer Music Conference*, pp. 344–347, 1997.

Darnell, Gregory, Georgiev, Stoyan, Mukherjee, Sayan, and Engelhardt, Barbara E. Adaptive randomized dimension reduction on massive data. *arXiv preprint arXiv:1504.03183*, 2015.

Gemmeke, Jort F, Virtanen, Tuomas, and Hurmalainen, Antti. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7): 2067–2080, 2011.

Lartillot, Olivier and Toiviainen, Petri. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, 2007.

Li, Tao, Ogihara, Mitsunori, and Li, Qi. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 282–289. ACM, 2003.

McKay, Cory and Fujinaga, Ichiro. Automatic genre classification using large high-level musical feature sets. In *ISMIR*, volume 2004, pp. 525–530. Citeseer, 2004.

Panagakis, Ioannis, Benetos, Emmanouil, and Kotropoulos, Constantine. Music genre classification: A multilinear approach. In *in Proceedings of ISMIR*, pp. 583–588, 2008.

Panagakis, Y. and Kotropoulos, C. Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 249–252, March 2010.

Snchez, Jorge, Perronnin, Florent, Mensink, Thomas, and Verbeek, Jakob. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

Tzanetakis, George and Cook, Perry. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

Wu, Ming-Ju and Jang, Jyh-Shing R. Combining acoustic and multilevel visual features for music genre classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 12(1):10:1–10:17, August 2015.

4. Appendix

Glossery	
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MIDI	Musical Instrument Digital Interface
GTZAN	Benchmark Dataset
MFCC	Mel-Frequency Cepstrum Coefficients
HCDF	Harmonic Change Detection Function
GMM	Gaussian Mixture Model
FV	Fisher Vectors
PCA	Principle Component Analysis

Table 4. Glossary of acronyms