Feature Extraction for Musical Genre Classification

MUS-15

Kilian Merkelbach

July 10, 2015

Abstract

Musical genre classification is a useful tool for automatically attaching semantic information to music tracks in large online and offline music collections. Due to the vast growth of such collections and the availability of music on the internet, the manual classification of the genre of an audio track by experts is starting to get replaced by automatic systems for genre classification. Since the space of music is very high-dimensional, such systems first extract information about each track that is useful for distinguishing the genre, a so called feature. Numerous methods for feature extraction have been developed. Two fundamentally different approaches are presented and compared in this paper.

Keywords: musical genre classification, feature extraction, machine learning, mel-frequency cepstral coefficients, rhythm, digraphs, histograms

Introduction

Genres provide a broad structure in the world of music; a set of classes to sort music into. Each class exhibits special characteristics, even though the distinction between genres that are similar to each other is not always easily determined.

Many people define their taste of music through genres, and that is reflected in music collections and music platforms such as Spotify or latfm, where genres make up the topmost level of the hierarchy of music. Even similarity search in music or automatic recommendations is often based on genres.

Since genres are labels with associated characteristics of music tracks, we need a way to capture the information from a track that is similar for tracks in the same genre and different for songs not in the same genre; i.e. information about each song that properly represents its musical characteristics. Some examples for such distinguishing characteristics are the rhythm, the used frequency band, or the texture of the sound. We call those characteristics features and the computation of them feature extraction.

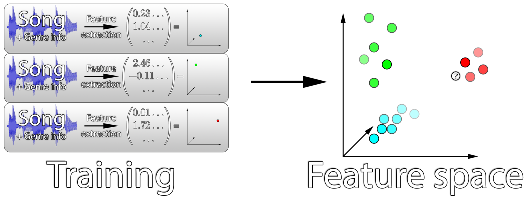

Figure 1: Pipeline for musical genre classification: The features are extracted from genre-annotated songs to populate a high-dimensional space (training). The gathered information can be used to identify the genre of an unknown song through its features. The different colors in the feature space signify the genre information that each training example is annotated with. For the new song, here marked with a question mark, a reasonable choice for a genre would be the red label.

The first step of any approach to musical genre classification is training, i.e. teaching the system the properties of the individual genres by example of music tracks (called the training set) from each genre. Only when the system has been trained on a sufficiently large training set can a reliable classification of new, before then unknown music tracks be attempted. Figure 1 shows this process schematically. It also shows another important criterion for the features. They have be chosen and adjusted in a way, such that music tracks of different genres inhabit areas of the feature space with little or no overlap. After all, the neighborhood of a point is an important aspect for finding the right genre classification of an unknown track using traditional data mining algorithms such as k-means or Gaussian mixture models (GMMs).

This paper will be limited on feature extraction, as this step is the component of a musical genre classification system where domain knowledge is necessary.

History

The research interest in automatic musical genre classification first started when manually sorting music into genres became too tiresome for human experts. While a small private music collection of the 1980s could have been sorted by hand, this changed rapidly at the end of the 1990s when the MP3 codec took off and the exchange of music via the internet became affordable in terms of bandwidth and storage. The introduction of cheap MP3 players to the market also led to a growth of private and commercial music collections. No longer was it a trivial task to divide the music tracks into multiple genres.

George Tzanetakis and Perry Cook, both then working or studying at Princeton University, were among the first researchers to write on the topic of musical genre classification with their paper *Musical Genre Classification of Audio Signals*[2] that was published in 2001. This work has been influential for the whole field and other researchers are still comparing their results to those achieved by Tzanetakis and Cook and even use the same training and testing set of music tracks to retain comparability with the paper.

Many approaches that followed build upon the findings of Tzanetakis and Cook while others utilize entirely different approaches.

Methods

For the sake of both getting a sense of the first steps of research into genre classification and examining a state of the art method, two works of research are going to be presented and compared in this paper; the first one is *Musical Genre Classification of Audio Signals*[2] by Tzanetakis and Cook, which is the first influential work in the field of musical genre classification and follows acoustic analysis approach. The second work is *Tracking the Beat: Classification of Music Genres and Synthesis of Rhythms*[1] by Debora C. Correa, Luciano da F. Costa and Jose H. Saito, which uses an entirely different approach based on weighted directed graphs to represent the rhythm of a music track.

In the following, the general idea and technical realization are presented. Finally, the results achieved with both systems are compared to one another.

Tzanetakis and Cook

The music tracks are represented in waveform to allow for the detailed analysis of the frequency band of the track. Three different kinds of features are calculated: timbral texture features (19 dimensions), rhythmic content features (6 dimensions) and pitch content features (5 dimensions), resulting in a 30-dimensional feature vector for each music track.

Features

Timbral texture features describe the way the music sounds. The timbral texture is not directly apparent from the music track but the nature of it is relatively easy to understand when one relies on intuition. For example, a Chinese gong could be said to have a smooth timbral texture while a heavily amplified electric guitar display have a rough texture. This feature category is where the usage of different instruments or singing voices will have the biggest impact.

The short time Fourier transform (STFT) is essential for the timbral texture features. It is an analysis of the frequency band of the music track in the following sense: It shows the prevalence of each specific frequency band and each point in time.

Based on the STFT, a number of frequency-related features are calculated: spectral centroid, a weighted mean of the most prevalent frequencies; spectral rolloff, the "frequency [...] below which 85% of the magnitude distribution is concentrated" [2]; spectral flux, a measure of the change in the frequency spectrum; and time domain zero crossing, a measure for the noisiness of the track.

Additionally, mel-frequency cepstral coefficients (MFCCs), features that were originally developed for automatic speech recognition, are used. MFCCs build upon STFT and are perceptually oriented, i.e. they are geared towards human audio perception. For musical genre classification, their values lie predominantly in the compact representation of the frequency spectrum. Only the first five computed coefficients are used in this approach. [2]

For feature computation, the music track to be processed is divided in partially overlapping pieces of 23ms, called the analysis window. Each of the timbral texture features previously described are only calculated for each of the analysis windows. The time frame of the analysis window was chosen to be so small such that the frequency within it is relatively stable, which aids analysis. The disadvantage of using a small window is that it is hard to make statements about the original track from such a short excerpt. To remedy that, the actual features later used for classification are computed on the 1s texture window as a running average and variance over the analysis windows contained in that texture window.

The second group of features are the rhythmic content features, for which the discrete wavelet transform (DWT) of the music track serves as the input instead of the STFT. The heart of the rhythmic content feature group is an autocorrelation function that computes for each possible time lag the alikeness of the music track signal to itself. For most tracks, the non-zero time offset at which the alikeness of the signal to itself is maximal will be the speed, since e.g. recurring signal parts such as rhythmic drums or percussion cause peaks in the autocorrelation function.

For each track, a histogram of the most dominant peaks in the autocorrelation function is created. It is called the beat histogram (BH) (Figure 2) and contains bins for beat-per-minute (bpm) values from 60 to 200, which captures most of today's music.

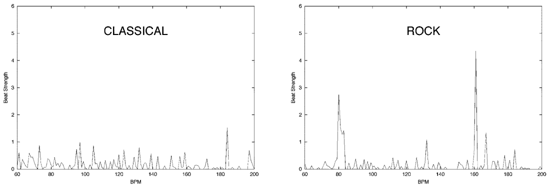

Figure 2: Beat histogram (BH) for the classical song La Mer by Claude Debussy on the left and Come Together by the Beatles on the right [2]. For each of the tracks, only an excerpt was used to calculate the BH. Since the flowing forms of classical music can even just visually be distinguished from the rhythmic nature of rock, it is clear that the BH is a useful feature for musical genre classification.

For classification, the BH is not directly used but features such as the highest peak and the sum of the whole BH are computed, resulting in a 6-dimensional feature vector for rhythmic content features.

The last set of features are pitch content features. The general structure of the computation of those features is similar to the computation of the rhythmic content features. This is due to the fact that detecting a repeating pitch can be seen as a small-scale-version of detecting a repeating rhythm. For beat detection, we use a window of about 0.5s to 1.5s, for pitch detection we use

2ms to 50ms.

Contrary to the BH, the pitch histogram (PH) is created in two different versions: folded, in which tones in different octaves are mapped to one bin; and unfolded, in which the octave of a tone is taken into consideration and the same tones in different octaves are mapped into different bins. From the two PH versions, a total of five aggregative features such as the maximum peak of the folded histogram and the total sum are calculated.

Correa, Costa and Saito

In *Tracking the Beat*, a different approach to musical genre classification is taken. As the name suggests, the focus lies on the rhythm of the music track. To receive a representation that is best suited for rhythm analysis, the tracks used for training and testing are encoded in the Music Instrument Digital Interface (MIDI) format. This format retains the full melody and rhythm informations, as opposed to waveform formats.

Features

The rhythm of a music track is composed of sequences of notes of differing lengths. Correa, Costa and Saito chose a representation which reflects that fact. They employ weighted directed graphs (digraphs) with 18 vertices corresponding to the most common note lengths and edges that show the relative prevalence of the two-note-sequence associated with them. This correctly allows for self-loops, since any note may be followed by a note of the same length.

The first step in generating a digraph for a music is extracting the note length, which is a trivial task due to the exact nature of the MIDI format. Then, the digraph is built note by note, each transition adding a small amount of weight to the respective edge in the digraph. After a digraph for each song has been created, aggregated digraphs for the different genres can be computed by taking the average weight for each edge over all songs belonging to the genre. The resulting genre digraphs serve as the input for two following procedures: calculating 15 features such as total vertex degree directly from the graph and applying principal component analysis (PCA) to the 18 × 18 matrix representing the digraph. The combined dimensionality of the result of both procedures is 52.

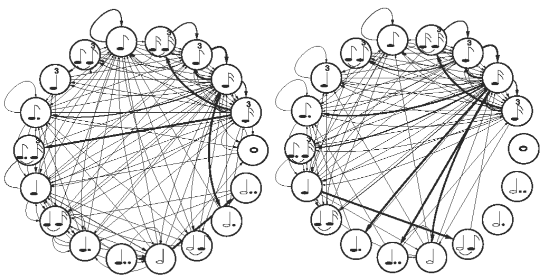

Figure 3: Digraph for rock on the left and for reggae on the right.

Results

The results obtained with both approaches are not truly comparable due to different testing settings. An attempt at a comparison that takes these differences into account will be made here.

In *Musical Genre Classification of Audio Signals*, Tzanetakis and Cook were able to classify an unknown music track correctly with an accuracy of 59% when differentiating between the ten genres classical, country, disco, hip hop, jazz, rock, blue, reggae, pop and metal. Additionally, a music/speech distinction could be made with an accuracy of 86%.[2]

Correa, Costa and Saito achieved a classification accuracy of 85.72%, working with the four genres blues, bossa-nova, reggae and rock.[1] When comparing both approaches, it is important to keep in mind that classification problems become harder as the number of alternative grows. Furthermore, the method presented in *Tracking the Beat* only deals with MIDI data, which offers much clearer input than waveform data. Since MIDI is not the format in which most music is originally created, it is probable that a track may not even be available in MIDI and has to be manually converted in order to be classified.

Conclusion

Two methods for musical genre classification have been presented. The approach from *Musical Genre Classification of Audio Signals* is based on frequency analysis, uses histograms to represent genre-distinguishing aspects of music tracks and achieves a classification accuracy of about 60%. The other approach, presented in *Tracking the Beat*, handles music stored in the discrete MIDI format and uses weighted digraphs to model the rhythms dominant in different genres, obtaining an accuracy of about 86%.

Those classification accuracies rival those achieved by humans. College students were asked to classify a song into one of ten genres and chose the right one in 70% if the cases. [2] Each song excerpt was only 3s long, but longer excerpts could not be found to improve the accuracy. Thus, automatic systems for musical genre classification are on par with human performance.

However, it must be said that in situations of many competing genres or subgenres, human music experts still exceed machine systems. This holds true especially when considering that handling more genres may also lead to more overlap between individual genres.

There are a number of ways in which the research in this topic could evolve. There could be features dedicated to and specifically developed for musical genre classification, so that speech recognition features such as MFCCs no longer need to be used. Another direction is fuzzy classification, e.g. classifying a song as 90% rock and 10% blues. This model would fit the blurred distinction between genres better than a hard classification.

References

[1] Debora C Correa, Luciano da F Costa, and Jose H Saito. Tracking the beat: Classification of music genres and synthesis of rhythms. *IWSSIP 2010 - 17th International Conference on Systems, Signals and Image Processing.*

[2] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on*, 10(5):293–302, Jul 2002.