International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 5, May - 2013

A Survey Of Mood-Based Music Classification

Sachin Dhande[1], Bhavana Tiple[2]
[1]*Department of Computer Engineering, MIT PUNE, Pune, India,*

[2]*Department of Computer Engineering, MIT PUNE, Pune, India,*

Abstract

Mood based music classification is an important application area in MIR. The purpose is to classify songs into different emotional groups like happy, sad, and angry. Despite its importance, it is rather difficult to evaluate the performance of mood classification algorithms. Acquiring emotional or mood information from music data is an important topic of music retrieval area. We have described the difference in the features and the types of classifiers used for different mood based classification systems.

Keywords*: Classification algorithm; feature extraction; feature selection; music information retrieval; music mood classification.*

IJERT

1. Introduction

Most people enjoy music in their free time. At present there is more and more music on personal computers, in music libraries, and on the Internet. Various metadata need to be created for each music piece in order to assist music organisation, music management and other music related applications such as music playlist generation and music search. Although conventional information such as the artist, the album, or the title of a musical work remains significant, these tags have restricted applicability in many music-related applications. Mood classification can be useful to events such as a DJ choosing music to control the emotional level of people on the dance floor, to a composer scoring a film, to someone preparing the soundtrack for their every day workout. Each of these situations depends extremely on the emotional content of the music.

There are very less publicly available benchmark data sets. Various researchers have used their own data sets as there is the lack of publicly available benchmark data sets. Because of this it is very difficult to compare different classification methods on an equal basis. Mood definitions are much dependent on individual feel and preferences. Psychologists developed different models such as Hevner's model and Thayer's model for mood classification. As these models are developed from subjective judgment and require general support, it is quite difficult to evaluate the performance of mood classification algorithms. Majority voting of individual opinions by discovering the mood metadata and collecting opinion polls from users, critics, and fans via collaborative filtering is one way to obtain objective ground truth information [2].

Mood based classification and genre classification has much similar features extraction and classification methods with a lot more importance on low-level spectral features. Many mood classification systems used rhythmic features but rhythmic information alone cannot yield good mood classification performance. It does play a much more important part in mood classification than genre and other classification tasks[2]. In this paper, we provide an overview of features and techniques used for mood based music classification.

2. Audio Features

Feature extraction and classifier learning are two important components of a classification system. Feature extraction tackles with the problem of how to represent the examples to be classified in terms of feature vectors or pair wise similarities. We can minimize prediction error using classifier learning by

finding mapping from the feature space to output labels.

Many audio features have been proposed in the survey music mood classification. We can categorize audio features by using different taxonomies. We can group the audio features into subcategories, namely short-term features, long-term features. The main difference is the length of local windows used for feature extraction. Short-term features like timbre features generally capture the characteristics of the audio signal in frames with 10–100 ms interval, whereas long-term features like temporal and rhythm features capture the long-term effect and interaction of the signal and are normally extracted from local windows with longer durations [2].

We can divide audio features into three groups' timbre, rhythm and pitch. Each classification tries to capture audio features from different viewpoint. We can divide audio features into two levels, low-level and mid-level features from the viewpoint of music understanding. Low-level features can be further separated into two classes of timbre and temporal features. Timbre features capture the tonal quality of sound that is related to different instrumentation, whereas temporal features capture the variation and evolution of timbre over time [2]. Various signal processing techniques like Fourier transform, spectral/cepstral analysis and autoregressive modelling are used to discover low level features. Timbre consists of features such as ZCR, SC, SR etc as shown in Table 1.Low-level features has simple procedure to obtain them and has good performance hence used mostly in music classification. However, they are not closely related to the basic properties of music as perceived by human listeners. Mid-level features include mainly three classes of features, namely rhythm, pitch, and harmony.

Low-level features are the important elements for audio classification systems. Table1 summarizes some of the familiar low-level features. Despite the good performance of low-level features for music classification, they do not capture the basic properties of music that humans perceive and appreciate. Low-level features are the most important features used for various music classification tasks.

The commonly used mid-level features in music analysis include rhythm which is frequent pattern of tension and release in music, pitch which is apparent fundamental frequency of the sound and harmony which is combination of notes simultaneously, to produce chords, and successively, to produce chord progressions. Although the above features are easily identified and acceptable by music listeners, it is not easy to define them unambiguously and extract them reliably from raw audio signals for the purpose of music analysis.

Yi Liu, Yue Gao [1], Liu D., Lu L., Zhang H.J[6] and Lie Lu, Dan Liu, Hong-Jiang Zhang[3] used intensity, timbre and rhythm features for classification and achieved good classification accuracy. Yi-Hsuan Yang, Chia-Chu Liu, and Homer H. Chen[8], Dr.M.Hemalatha, N.Sasirekha, S.Easwari, N.Nagasaranya[5] used low-level features for classification. Hongchen JIANG, Junmei BAI, Shuwu ZHANG, Bo XU[11] used sixteen kinds of audio features in SVM-based audio classifiers. These features are: Zero-Crossing Rate (ZCR), High ZCR Ratio (HZCRR), Short-Time Energy (STE), Low STE Ratio (LSTER), Root Mean Square (RMS), Silence Frame Ratio (SFR), Sub-band Energy Distribution (SED), Spectrum Flux (SF), Spectral Centriod (SC), Spectral Spread (SS), Spectral Rolloff Frequency (SRF), Sub-band Periodicity (BP), Noise Frame Ratio (NFR), Linear Spectrum Pair (LSP), Linear Predictive Cepstral Coefficients (LPCC) and Mel-frequency Cepstral Coefficients (MFCC).

2.1 Timbre features

The majority of the features listed in Table 1 are timbre features. As a basic element of music, timbre is a term describing the quality of a sound. Different timbres are produced by different types of sound sources, like different voices and musical instruments. Timbre in music and color in images are much simillar. We can define some summary features such as spectral centroid (SC), spectral rolloff (SR), spectral flux (SF), and spectral bandwidth (SB) capturing simple statistics of the spectra. Hereafter, we term the collection of these features as short time Fourier transform (STFT) features. It is possible to extract more powerful features such as MFCC, OSC, DWCH, and MPEG-7 audio descriptors like SFM, SCF, and ASE.

Mel-frequency cepstral coefficients (MFCCs) are among the most widely used acoustic features in speech and audio processing. MFCCs are essentially a low-dimensional representation of the spectrum warped according to the mel-scale, which reacts the nonlinear frequency sensitivity of the human auditory system.

2.2 Temporal features

Temporal features form another significant class of low-level features that capture the temporal evolution of the signal. Temporal features are generally constructed on top of timbre features. Some of the simplest types of temporal features are statistical moments such as mean, variance, covariance.

IJERT

Table 1 Feature for Music Classification

Low-Level Features	Timbre	spectral centroid (SC) spectral rolloff(SR) spectral flux (SF) spectral bandwidth (SB) Mel-frequency ceptrum Coefficient(MFCC) Linear Predictive ceptrum Coefficient(LPCC) Octave-based spectral contrasts(OSC)
	Temporal	Stastical moments(SM) Amplitude modulation(AM) Auto-regressive modelling(ARM)
Mid-Level Features	Rhythm	Beat histogram (BH). Beat-per-minute (BPM)
	Pitch	Pitch histogram(PH) Pitch class profile (PCP)
	Harmony	chord sequences (CS)

2.3 Rhythm Feature

Strength, regularity and tempo are three main aspects of rhythm which are closely related with people's mood response. For example, in songs with high energy and low stress, the rhythm is usually strong, steady and the tempo is fast; while in songs with low energy and high stress, music is usually slow and with no distinct rhythm pattern. In simple words we can say sad songs have a slow rhythm, whereas angry songs usually have a fast rhythm.

Rhythm is the most extensively used mid-level feature in audio-based music classification. It describes how certain patterns occur and reappear in the music. Beat and tempo (beat-per-minute, BPM) are two important indications that describe rhythmic content of the music which have been used in music classification. The auto-correlation of the time-domain envelop signal is determined. The peaks of the auto-correlation function are then identified which correspond to probable regularity of the music under analysis. The beat histogram represents the distributions of the regularities showed in the envelop signal, where rhythmic features can be obtained such as magnitudes and locations of dominant peaks and BPM. As the mood of a song is extremely correlated with rhythm, these features have good experimental performance for mood classification.

2.4 Pitch and Harmony

Pitch and harmony are also important components of music. Pitch is defined as most fundamental frequency of the sound determined by what the ear judges. However, a pitch is not equal to the fundamental frequency because the perception of pitch is completely subjective while frequency measurement is objective. Other reasons like differences in timbre, loudness, and musical context also affect pitch. A musical note played on most instruments consists of a series of harmonic-related frequency, including the fundamental frequency and partials at integer multiples, and is normally perceived as one sound with a single pitch. Hence, pitch information extraction in real audio signals is more than locating the fundamental frequency [2]. Various multi-pitch estimation algorithms have been developed to identify the top candidate pitches for each frame for frame level pitch analysis. Song level pitch feature representation like the pitch histogram (PH) can be derived and applied to classification. Along with low-level timbre features like MFCC and other spectral features Pitch histogram can be used for mood classification. It represents the distribution of candidate pitches extracted from all frames. Every histogram bin captures the occurrence frequency of the corresponding pitch. Pitch class or chroma is the important concept about pitch which defines an equivalent class of pitches. Pitch class features like pitch class profile (PCP) and harmonic pitch class profile (HPCP) have been developed and extensively used in various tasks like melody analysis and transcription. The chroma feature can be obtained

IJERT

directly by converting the spectrum values without any attempt on pitch detection. It has been earlier used in music classification in combination with MFCC and the combined feature was shown to outperform the MFCC feature alone.

Harmony entails the use of chords. Basic element of harmony is chord which involves the simultaneous combination of two or more notes. Harmony is achieved by chord progression that is a series of chords played in succession. Melody captures the horizontal information of music whereas harmony explores the vertical dimension. Various chord detection and recognition algorithms can be used to extract Chord information like chord sequences (CS) from the music audio data making. All these methods begin with pitch detection using either standard or enhanced pitch features to recognize the fundamental frequency and its partials. Then every pitch histogram feature is compared with the chord template to find out the existence of possible chords. In music mood classification the inclusion of chord features in combination with timbre and rhythm features can improve classification performance [2]. To review the choice of audio features is much dependent on the problems we deal with. Timbre features are appropriate for genre and instrument classification but not suitable for comparing the melody similarity of two songs. For mood based music classification, a large amount of works used rhythm features. In general, there is no particular set of task-independent features that can every time do better than the others.

IJERT

3. Classifiers

In standard classification, we are presented with a training data set where each example comes with a label. The purpose is to propose a classification rule that can best predict the labels for unseen data. K-nearest neighbour (K-NN) , support vector machine (SVM) and GMM classifier are most popular choices for classifiers.

3.1 K-NN Classifier

K-NN is one of the most accepted classifiers used for both general classification problems and in mood based music classification as well. K-NN uses training data directly for the classification of testing data. We can predict label of the testing instance by majority voting on the labels of the nearest instances in the training set. KNN is an example of a non-parametric classifier. If we denote *D^n=x1,x2,.....xn* a set of *n* labelled prototypes then the nearest neighbor rule for classifying an unknown vector *x* is to assign it the label of its closest neighbouring point in the set of labelled prototypes *Dn*. It is possible to show that for an unlimited number of prototypes the error rate of this classifier is never worse than twice the optimal Bayes rate. The KNN rule classifies *x* by assigning the label most regularly represented among the *k* nearest samples. Normally *k* is odd to avoid ties in voting. Various methods are used to make algorithms computation faster and storage requirements smaller as this algorithm has heavy time and space requirements.
Yi-Hsuan Yang, Chia-Chu Liu, and Homer H. Chen[8] uses fuzzy K-NN as classifier. Fuzzy k-NN classifier is a combination of fuzzy logic and k-NN classifier. It contains two steps: fuzzy labelling that computes the fuzzy vectors of the training samples and fuzzy classification that computes the fuzzy vectors of the input samples.

3.2 SVM Classifier

SVM is the high-tech binary classifier based on the large margin principle. Given labelled instances from two different classes, SVM classifier finds the optimal separating hyper plane which maximizes the distance between support vectors and the hyper plane. Instances closest to the hyper plane whose labels are most likely to be confused are the support vectors. Therefore, the SVM has better classification performance as it focuses on the difficult instances. Both K-NN and SVM are applicable to single feature vector representations as well as pair wise similarity values. In the second case, a kernel matrix is built from pair wise similarity values which can be used directly by the SVM.
Cyril Laurier, Perfecto Herrera[7] uses Support Vector Machine classifier to predict the mood cluster. They uses a set of 133 descriptors. The features are spectral, temporal, tonal but also describe loudness and danceability. The features were selected beforehand according to experiments on annotated databases. A grid search algorithm is used to optimize SVM. SVM is chosen as basic classifier by Yi Liu , Yue Gao [1] and presented 7-class mood model. Cyril Laurier Music Technology Group, Jens Grivolla Fundaci´o Barcelona Media, Perfecto Herrera Music Technology Group[4] uses both audio and lyric information for classification and SVM as classifier. Their model shows much accuracy improvement in mood based audio classification.
Ruijie Zhang, Bicheng Li, Tianqiang Peng[9] present a high-accuracy audio classification algorithm based on SVM-UBM using MFCCs as classification features. Firstly MFCCs are extracted in frame level, then a Universal Background Gaussian Mixture Model (UBM) is used to integrate these sequences of frame-

level MFCCs within a clip to form the clip-level feature, finally audio classification is done using SVM with these clip-level features. Lei Chen, S. ule G¨und¨uz, M. Tamer O¨ zsu[10] examined the appropriateness of SVM on mixed type audio classifier and comparison experiments show that the maximum of feature values in each audio clip can capture the characteristic of mixed type audio data and SVM-based classifier do better than other popular classifier such as k-NN. Hongchen JIANG, Junmei BAI, Shuwu ZHANG, Bo XU[11] employed two kinds of SVM-based classification frameworks to classify audio signals into five classes, which are pure speech, non-pure speech, music, environment sound and silence. These experiments have achieved the average 96.61% and 96.90% classification accuracy respectively.

3.3 GMM Classifier

For the GMM classifier, we fit the Gaussian mixture model over the distributions of features in each and every class. With the class conditional probability distribution, labelling of testing example can be done according to the Bayes rule

$f(x)=\arg\max P(y=k|x)$

$P(y=k|x) = P(x|y=k)P(y=k) / \sum P(x|y=k)p(y=k)$

The decision based on the maximizer of the posterior probability identifies the labels, data and the conditional probability of example for class label estimated from the training data using GMM. Prior probability specifies the proportion of label in the training data. Specifically, GMM classifier can also be used for feature set input. We can apply the product rule to calculate approximately the class conditional probability for feature sets by assuming that timbre features in each class are independent and identically distributed. Dan Liu , Lie Lu[6] and Lie Lu[3] uses GMM as a classifier. Yi Liu , Yue Gao [1]presented 7-class mood model and Compares the result of GMM with SVM. George Tzanetakis [12] explains GMM classifier and the EM algorithm.

3.4 Other Classifiers

Various other classifiers have also been used for different music classification tasks, including logistic regression, artificial neural networks (ANN), decision trees, linear discriminant analysis (LDA), nearest centroid (NC) and sparse representation-based classifier (SRC). Convolutional neural network (CNN) is a simplification of the standard neural network model by taking convolutions over the segments of the input signal can directly handle feature set classification. Hence, such model can be used for audio classification based on sequence of timbre features like raw MFCC features. Dr.M.Hemalatha, N.Sasirekha, S.Easwari, N.Nagasaranya [5]proposed a model for audio clustering and classification technique by using neural networks for classifying the data.

4. Feature Learning and Classifier Learning

The reason of feature learning is to automatically select and extract features for improving the classification performance over general audio features. Feature learning is very much related to classifier learning. In selection, features are directly selected based on some feature selection rules from a large number of candidate input features. Both feature selection and extraction can be done in supervised or unsupervised fashion. In the supervised setting, labelled data are used to help out the selection or extraction of useful features that best distinguish between different labels. One possible approach for feature selection is to learn a front-end classifier like logistic regressor, which can be trained efficiently, and rank the attributes based on the classifier weights[2]. The lowest ranked feature attributes are then leftover in training the final classifier. It is possible to perform linear feature extraction by learning a transformation matrix to project higher dimensional feature vectors to a lower dimensional subspace that preserves most of the discriminated information. This can be achieved by a variety of metric learning algorithms that are useful for feature learning in music classification.

In unsupervised feature extraction methods process input features without making use of the label information. Principal component analysis (PCA) is a standard method for unsupervised feature extraction. Principal Component Analysis (PCA) can be used for dimensionality reduction.

Feature combination from different sources is an efficient way to improve the performance of mood based music classification systems. We can combine feature in some way for music classification if multiple features are available. One of the simple way to feature combination is to concatenate all features into a single feature vector, for combining timbre with beat and pitch features. Feature combination can be incorporated with classifier learning. Multiple kernels learning (MKL) is one such framework developed mainly for SVM classifiers [13]. The use of MKL is to learn an optimal linear combination of features for SVM classification. MKL has recently been applied to music classification and found to do better than any of the single feature types. As an choice to feature combination, we can also perform decision-level fusion

IJERT

to combine multiple decisions from different classifiers. There are lots of ways to perform decision level fusion some of them are majority voting, sum rule which takes the average of decision values returned by individual classifiers, etc.

A more common framework is established by the technique of stacked generalization (SG)[14], which provides a cascaded framework for classification by stacking classifiers on top of classifiers. In the stacked generalization framework, classifiers at the first level are trained on individual features and those classifiers at the second level are trained by using the decision values returned by level-1 classifiers as new features. Hence, Stacked Generalisation finds the fusion rule through supervised learning. The option of classifiers used for SG is quite flexible. Usually SVMs are used within SG for optimized performance. Another vital class of feature combination methods is based on ensemble methods for classification. One such example is AdaBoost with decision trees (AdaBoost. DT)[15], which combines decision tree classifiers with the boosting framework. Every decision tree classifier is trained on a single type of feature.

5. Conclusion

The survey has provided current discussion of audio features used for mood based music classification. Survey describe the difference in the features and the types of classifiers used for different mood based classification systems also states how much accuracy can be achieved with particular classifier. If multiple features are available, we can combine those features in some way for music classification. Feature combination from different sources is an effective way to improve the performance of mood based music classification systems.

6. References

[1] Yi Liu, Yue Gao. "Acquiring Mood Information from Songs in Large Music Database", 2009 Fifth International Joint Conference on INC, IMS and IDC.

[2] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang, "A Survey of Audio-Based Music Classification and Annotation", IEEE Transactions on Multimedia, Vol. 13, No. 2, April 2011.

[3] Lie Lu, Dan Liu, Hong-Jiang Zhang," Automatic Mood Detection and Tracking of Music Audio Signals" IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 1, January 2006.

[4] Cyril Laurier , Jens Grivolla Fundaci´o , Perfecto Herrera," Multimodal Music Mood Classification using Audio and Lyrics",International Conference on Machine Learning and Applications San Diego, California (USA) December 2008.

[5] Dr.M.Hemalatha, N.Sasirekha, S.Easwari, N.Nagasaranya," An Empirical Model for Clustering and Classification of Instrumental Music using Machine Learning Technique", 2010 IEEE International Conference on Computational Intelligence and Computing Research.

[6] Liu D., Lu L., Zhang H.J.," Automatic Mood Detection from Acoustic Music Data", Proc. of the 4th Int. Conf, Music Information Retrieval (ISMIR'03), Washington, DC, USA, October 2003:

[7] Cyril Laurier, Perfecto Herrera, "Audio Music Mood Classification Using Support Vector Machine." International Society for Music Information Research Conference (ISMIR).

[8] Yi-Hsuan Yang, Chia-Chu Liu, and Homer H. Chen," Music Emotion Classification: A Fuzzy Approach", Proceedings of the 14th annual ACM international conference on Multimedia.

[9] Ruijie Zhang, Bicheng Li, Tianqiang Peng, "Audio Classification Based on SVM -UBM", ICSP2008 Proceedings.

[10] Lei Chen, S. ule G¨und¨uz, M. Tamer O¨ zsu, "MIXED TYPE AUDIO CLASSIFICATION WITH SUPPORT VECTOR MACHINE" In Multimedia and Expo, 2006 IEEE International Conference on (July 2006).

[11] Hongchen JIANG, Junmei BAI, Shuwu ZHANG, Bo XU, "SVM-based Audio Scene Classification", Proceeding ofNLP-KE'05.

[12] GEORGE TZANETAKIS," Manipulation, analysis and retrieval systems for audio signals", 2002 Doctoral Dissertation.

[13] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the Kernel matrix with semidefinite programming," *J. Mach. Learn.Res.*,vol.5,pp.27–72,2004.

[14] D. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259,1992.

[15] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl, "Aggregate features and ada boost for music classification," *Mach. Learn.*, vol. 65, no. 2–3, pp. 473–484, 2006.

IJERT