FACTORS IN AUTOMATIC MUSICAL GENRE CLASSIFICATION OF AUDIO SIGNALS

Tao Li

Computer Science Department
University of Rochester
P.O Box 270226
Rochester, NY, 14627-0226, USA
taoli@cs.rochester.edu

George Tzanetakis

Computer Science Department
Carnegie Mellon University
Forbes Avenue 5000
Pittsburgh, PA 15221, USA
gtzan@cs.cmu.edu

ABSTRACT

Automatic musical genre classification is an important tool for organizing the large collections of music that are becoming available to the average user. In addition it provides a structured way of evaluating musical content features that doesn't require extensive user studies. This paper provides a detailed comparative analysis of various factors affecting automatic classification performance such as choice of features and classifiers. Using recent machine learning techniques such as Support Vector Machines we improve on previously published results using identical data collections and features.

1. INTRODUCTION

Improvements in audio compression together with increases in hard disk capacity and network bandwidth have made possible the creation of large personal music collections. Digital music distribution is already popular in peer-to-peer file sharing environments and the exchange of music files consumes the majority of internet bandwidth. The recording industry although reluctantly is slowly embracing these new technologies while trying to retain copyright control. Once copyright protection can be enforced using techniques such as audio fingerprinting [1] it is very likely that all of recorded music will be available digitally. This scenario is very likely to happen in the near future. The problems and challenges of organizing these vast amounts of musical information for searching and browsing is the topic of the emerging research area of Music Information Retrieval (MIR) (two good recent overviews of MIR are [2, 3]).

The automatic analysis of music stored in audio format is one of the important topics of MIR. The majority of such audio analysis techniques make use of numerical features that attempt to capture information about musical content. Musical genres are categorical labels created and used by humans in order to structure the vast universe of music. Although the boundaries that separate them are fuzzy and there is significant overlap, members of a particular genre share characteristics related to the instrumentation, rhythmic structure and pitch content of the music. Therefore automatic musical genre classification provides a good way to evaluate numerical features that attempt to capture musical content. Such features form the basis of any type of audio analysis and retrieval work. In addition automatic genre classification provides an automatic way to structure and organize the large number of music files available digitally on the Web. An excellent recent overview of representing musical genre in digital music distribution is [4] which covers manual annotation, automatic methods and usage-based methods such as collaborative filtering.

In this paper, we provide comparisons of various factors that affect automatic musical genre classification performance. The majority of existing literature in automatic musical genre classification makes use of traditional statistical pattern recognition classifiers such as Gaussian Mixture Models and K-Nearest Neighbors [5]. We investigate the effect of using more recent powerful classification methods such as Support Vector Machines [6] and report significant improvements in classification accuracy compared to results previously reported in the literature using identical features and data collections. The obtained results are also comparable to the human genre classification results reported in [7].

Previous work in the area of automatic musical genre classification includes: features computed based on wavelet analysis and simple classifiers [8], visual texture features of spectograms for classification [9], and a specialized architecture called "Explicit Time Modeling Neural Networks" for genre discrimination [10]. A comparison of audio features with features extracted from analysis of cultural meta-data such as download usage patterns is presented in [11]. A more detailed study of automatic musical genre classification is presented in [12]. In this work, three different sets of features for representing timbral texture, rhythmic content and pitch content are proposed. For the experiments described in Section 4 the data collections and features of [12] are used and are briefly described in Section 2. In all these papers, there is little comparative evaluation of different feature combinations and classifiers which is the main goal of this paper.

2. MUSIC CONTENT FEATURES

2.1. Timbral Texture

The features used to represent timbral texture are based on standard features proposed for music-speech discrimination and speech recognition. They consist of a set of 4 features computed based on the Short Time Fourier Transfrom (STFT) magnitude spectrum such as the Spectral Centroid (defined as the first moment of the magnitude spectrum) as well as the first 5 Mel-Frequency Cepstral Coefficients (MFCC) [13]. These features are computed using an analysis window of 20 milliseconds. Means and variances of the features over a larger texture window (1 second) with a hop size of 20 milliseconds are computed resulting in a set of 18 features. An additional feature (the percentage of low energy frames over the texture window) results in a timbral texture feature vector of 19 dimensions. These features are described in more detail in [12].

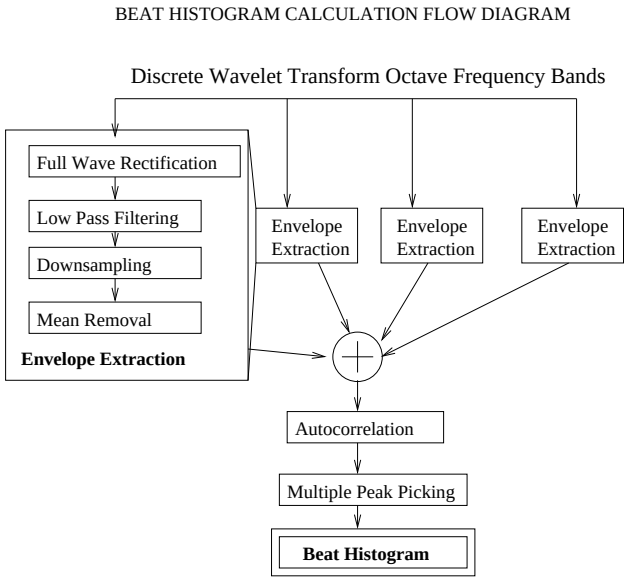

Figure 1: Beat Histogram Calculation Diagram

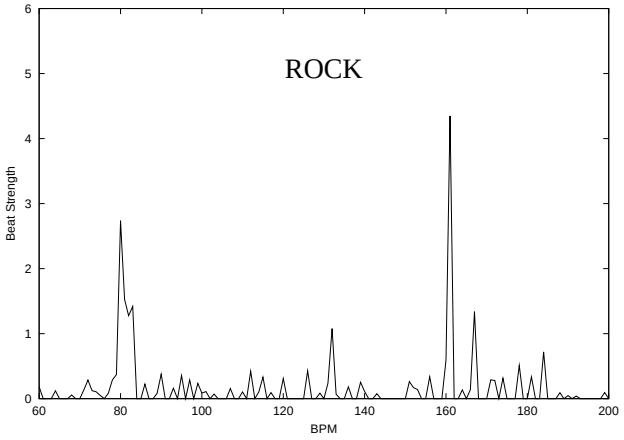

Figure 2: Beat Histogram

2.2. Rhythmic Content Features

The basis of representing rhythmic content is the calculation of a Beat Histogram (BH) that shows the distribution of various beat periodicities of the signal. For example a piece with tempo 60 Beats-per-Minute (BPM) would exhibit BH peaks at 60 and 120 BPM. The BH is calculated using periodicity detection in multiple octave channels that are computed using a Discrete Wavelet Transform. Figure 1 shows a schematic diagram of the this calculation. In [12], six numerical features that attempt to summarize the BH are computed and used for classification. In addition in this paper, the use of the full BH for classification was explored. Figure 2 shows a BH for a piece of Rock music (notice the peaks at 80 BPM (main tempo) and 160 BPM).

2.3. Pitch Content Features

Similarly, the pitch content features are based on accumulating the results of multiple pitch detection [14] in a Pitch Histogram (PH). The histogram provides informations about the pitch class and pitch probability distribution across the file. (pitch class refers to folding the pitch to the range of one octave - for example A4=440Hz and A5=880Hz map to the same pitch class A). The PH attempts to capture information such as jazz pieces have on average more chord changes than pieces of country music. Five numerical features that summarize the PH are proposed in [12] and used for classification. In addition, in this paper, the use of the full PH for classification is explored.

3. CLASSIFIERS

3.1. Support Vector Machines

Support vector machines (SVMs) [6] have shown superb performance at binary classification tasks and handle large dimensional feature vectors better than other classification methods. Basically, a Support Vector Machine aims at searching for a hyperplane that separates the positive data points and the negative data points with maximum margin. To extend SVMs for multi-class classification, we use pairwise comparison approaches and multi-class objective functions approaches. In pairwise comparison, a classifier is trained for each possible pair of classes. For K classes, this results in $(K-1)K/2$ binary classifiers. Given a new instance, the multi-class classification is then executed by evaluating all $(K-1)K/2$ individual classifiers and assigning the instance to the class which gets the highest number of votes. The idea of multi-class objective function is to directly modify the objective function of support vector machine (SVM) in such a way that it simultaneously allows the computation of a multi-class classifier. For pairwise comparison method, our SVM implementation is based on the LIBSVM [15], a library for support vector classification and regression. For multi-objective functions, our implementation is based on multi-category Proximal Support Vector Machines(MPSVM) [16]. For experiments involving SVMs, we test them with linear, polynomial and radius-based kernels and the results reported are the best among these trials.

3.2. Linear Discriminant Analysis (LDA)

Discriminant analysis approaches are well known to learn discriminative feature transformations in statistical pattern recognition literature and has been successfully used in many classification tasks [5]. The basic idea of LDA is to find a linear transformation that best discriminate among classes and the classification is then performed in the transformed space based on some metric such as Euclidean Distances etc. Fisher discriminant analysis finds discriminative feature transform as eigenvectors of matrix $T = \hat{\Sigma}_w^{-1}\hat{\Sigma}_b$ where $\hat{\Sigma}_w$ is the intra-class covariance matrix and $\hat{\Sigma}_b$ is the inter-class covariance matrix. Basically T captures both compactness of each class and separations between classes and hence eigenvectors corresponding to largest eigenvalues of T would constitute a discriminative feature transform. In our experiments, we use Fisher's Linear discriminant Analysis.

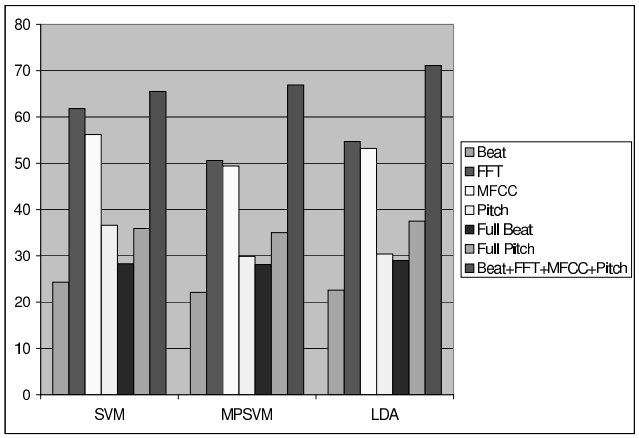

Figure 3: Accuracy comparison of different features.

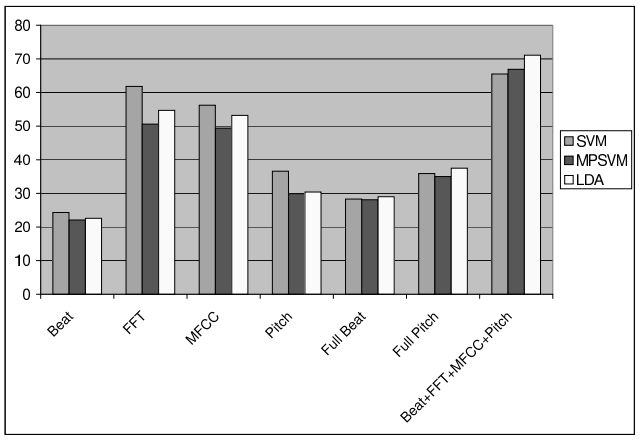

Figure 4: Accuracy comparison of different methods

4. EXPERIMENTAL RESULTS

For the conducted experiments the following feature subsets and their combinations were used (the numbers in parentheses indicate the dimensionality): FFT (9), MFCC (10), Beat (6 BH-based), Pitch (5 PH-based), Full Beat (300), Full Pitch (130). Each genre was represented by 100 sound files resulting in 10 * 100 = 1000 feature vectors. (the genres were: classical, country, disco, hiphop, jazz, rock, blues, reggae, pop, heavy metal). The data collection was the same as the one used in [12]. The features were calculated using MARSYAS, a free software framework for audio analysis (`http://marsyas.sourceforge.net`).

Table 1 compares the classification accuracy of various subsets and their combinations using three classifiers: Pair-wise Support Vector Machines (SVM), Multi-category Proximal Support Vector Machine (MPSVM) and Linear Discriminant Analysis (LDA). Figure 3 shows graphically some of the comparisons between different subsets of features. Figure 4 compares different classifiers. These results were obtained using 10-fold cross-validation (the labeled data is split randomly into 90% training data and 10% testing data 100 times and the accuracy of each run was averaged).

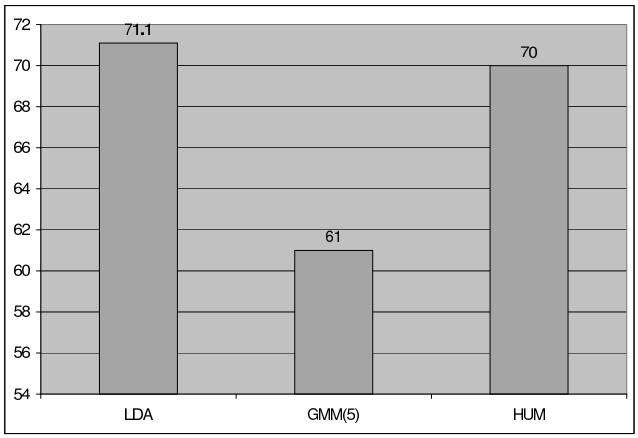

Figure 5: Comparison with previously reported results for automatic and human genre classification

5. CONCLUSIONS AND FUTURE WORK

From the experimental comparison it can be seen that the relative importance of feature subsets is (in order of decreasing classification accuracy): FFT, MFCC, Pitch, and Beat. This result indicates that Beat features, which incidentally are the most time consuming to calculate, are not as important to classification as the other feature sets. Another interesting result, is that using the Full Beat Histogram and Full Pitch Histogram doesn't seem to improve significantly the classification accuracy. This implies that the published Beat and Pitch features of low dimensionality capture most of the classification information contained in the full histograms. The best classification result is given by Linear Discriminant Analysis on the full feature set (FFT+MFCC+Pitch+Beat). For other feature combinations, there seem to be no consistent winners. Moreover, the accuracy results of the three methods do not differ by much.

The results of this paper improve significantly the classification accuracy reported in [12]. The numbers are directly comparable as the same data collection and feature set was used in both works. The best accuracy (71% for the LDA classifier) is significantly better than the (61%) reported in [12] for a Gaussian Mixture Model (GMM) classifiers with 5 components. In addition our result is indirectly comparable to the 70% human musical genre classification accuracy reported in [7]. Although direct comparisons of these results is not possible due to different data collections, it is clear that automatic performance is not far away from human performance. In addition the results of [7] indicate that genre classification is a hard problem with fuzzy boundaries not only for machines but also for humans. Figure 5 displays these results which indicate that more powerful machine learning classifiers can have significant impact in classification performance.

In the future, we are planning to investigate the performance of our classification methods to larger data collections with more genres. In addition to timbre, rhythm and pitch content, two other sources of information that could be useful are melody and singer voice. Although results from melodic analysis and singer identification probably will not be very good they might still provide enough information for genre classification. Finally we are exploring hierarchical as well as real-time musical genre classification.

Features \ Methods	SVM	MPSVM	LDA
Full Beat	28.3(3.36)	28.1(4.75)	29.0(3.68)
Full Pitch	35.9(2.51)	35.0(3.74)	37.5(2.68)
Full Beat + Pitch	35.9(2.51)	35.0(3.74)	36.6(2.84)
Full Beat + other	64.4(5.60)	63.8(5.12)	68.4(5.56)
Full Pitch + other	63.5(5.28)	65.6(3.20)	69.7(4.35)
Full Beat + Pitch + other	60.5(5.15)	62.5(5.68)	60.3(4.52)
Full Beat + pitch + other -Beat	60.3(5.66)	61.9(5.63)	60.5(5.52)
Full Beat + pitch + other -pitch	60.8(4.51)	61.1(4.82)	60.0(4.40)
Full Beat + pitch + other -pitch-beat	60.2(4.31)	61.1(5.67)	60.2(5.13)
Beat+FFT+MFCC+Pitch	65.5(4.88)	66.9(5.74)	71.1(7.27)
Beat+FFT+MFCC	65.8(4.18)	64.7(6.49)	69.6(8.29)
Beat+FFT+Pitch	55.8(3.74)	56.0(4.67)	60.3(6.27)
Beat+MFCC+Pitch	59.9(3.67)	57.8(3.82)	61.0(5.49)
FFT+MFCC+Pitch	67.2(4.80)	65.7(5.21)	67.9(7.78)
Beat+FFT	53.0(4.11)	50.8(5.16)	55.5(7.75)
Beat+MFCC	53.3(4.69)	53.5(4.45)	55.5(4.47)
Beat+Pitch	36.8(3.29)	35.6(4.27)	36.9(4.58)
FFT+MFCC	69.1(5.30)	64.1(5.76)	68.4(7.49)
FFT+Pitch	59.4(4.58)	56.1(5.82)	59.2(6.75)
MFCC+Pitch	55.9(5.57)	53.3(2.95)	56.9(5.02)
Beat	24.3(2.50)	22.1(3.04)	22.6(2.63)
FFT	61.8(4.18)	50.6(5.76)	54.7(8.03)
MFCC	56.2(4.64)	49.4(2.27)	53.2(3.22)
Pitch	36.6(2.95)	29.9(3.76)	30.4(3.53)

Table 1: Accuracy(standard deviation) table of various methods on various feature sets. SVM denotes pairwise SVM, Full Beat and Pitch are the full histograms and "other" refers to the 30-vector feature set (FFT + MFCC + BEAT + PITCH).

6. REFERENCES

[1] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 107–115.

[2] J. Futrelle and S. J. Downie, "Interdisciplinary Communities and Research Issues in Music Infomration Retrieval," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 215–221.

[3] F. Pachet, "Content Management for Electronic Music Distribution: The Real Issues," *Communications of the ACM*, vol. 46, no. 4, Apr. 2003.

[4] J. J. Aucouturier and F. Pachet, "Musical Genre: a Survey," *Journal of New Music Research*, vol. 32, no. 1, 2003.

[5] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed. New York: Academic Press, 1990.

[6] V. N. Vapnik, *Statistical Learning Theory*. Wiley, New York, 1998.

[7] D. Perrot and R. Gjerdigen, "Scanning the dial: An exploration of factors in identification of musical style," in *Proc. Society for Music Perception and Cognition*, 1999, p. 88, (abstract).

[8] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of Audio Signals using Statistical Features on Time and Wavelet Transform Domains," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

[9] H. Deshpande, R. Singh, and U. Nam, "Classification of Musical Signals in the Visual Domain," in *Proc. COST G-G Conf. on Digital Audio Effects (DAFX)*, Limerick, Ireland, Dec. 2001.

[10] H. Soltau, T. Schultz, M. Westphal, and A. Waibel, "Rcognition of Music Types," in *Proc. Int. Con.f on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 1998, pp. 1137–1140.

[11] B. Whitman and P. Smaragdis, "Combining Musical and Cultural Features for Intelligent Style Detection," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 47–52.

[12] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[13] S. Davis and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *IEEE Transcactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[14] T. Tolonen and M. Karjalainen, "A Computationally Efficient Multipitch Analysis Model," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, Nov. 2000.

[15] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[16] G. Fung and O. Mangasarian, "Proximal support vector machine classifiers," in *Knowledge Discovery and Data Mining*, 2001, pp. 77–86.