A Survey of Audio-Based Music Classification and Annotation

Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang


presenter: Yin-Tzu Lin (阿孜孜^.^) 2011/08
Types of Music Representation

• Music Notation
  – Scores
  – Like text with formatting

• Time-stamped events
  – E.g. Midi
  – Like unformatted text

• Audio
  – E.g. CD, MP3
  – Like speech

*Image from: [http://en.wikipedia.org/wiki/Graphic_notation](http://en.wikipedia.org/wiki/Graphic_notation) Inspired by Prof. Shigeki Sagayama’s talk and Donald Byrd’s slide*
Intra-Song Info Retrieval

Composition
Arrangement
Music Theory
Learning

Symbolic

Score Transcription
MIDI Conversion
Melody Extraction
Structural Segmentation
Key Detection
Chord Detection
Rhythm Pattern
Tempo/Beat Extraction
Onset Detection

probabilistic inverse problem

Modified speed
Modified timbre
Modified pitch
Separation

Audio

Accompaniment
Performer
Synthesize
Inter-Song Info Retrieval

- **Generic-similar**
  - Music Classification
    - Genre, Artist, Mood, Emotion…
  - Tag Classification (Music Annotation)
  - Recommendation

- **Specific-similar**
  - Query by Singing/Humming
  - Cover Song Identification
  - Score Following
Classification Tasks

- Genre Classification
- Mood Classification
- Artist Identification
- Instrument Recognition
- Music Annotation
Paper Outline

• Audio Features
  – Low-level features
  – Middle-level features
  – Song-level feature representations

• Classifiers Learning

• Classification Task

• Future Research Issues
Audio Features

Top-Level labels

- **Genre**
  - Pop/rock
  - Classical...
- **Mood**
  - Happy, Sad
  - Angry...
- **Instrument**
  - Piano, Violin
  - Guitar, Drum..
- **More**
  - Artist? Style?
  - Similar Song?

Mid-Level features

- **Rhythm**
  - BH, BPM
- **Pitch**
  - PH/PCP, EPCP
- **Harmony**
  - CP, CH

Low-Level features

- **Timbre**
  - ZCR, SC, SR, SF..
  - MFCC, DWCH...
- **Temporal**
  - SM, ARM
  - FP, AM

Short-Term

Long-Term
Low-level Features

10~100ms

Ex: Mel-scale, bark scale, octave
Short-Time Fourier Transform

Time Domain

(a): $f$

(b): $2f$

(c): $(a)+(b)$

(d): $(a) (b)$

Frequency Domain

150
100
50
0
0
20
40
60
80
100
120
0
50
100
150
200
250
0
50
100
150
200
250
0
50
100
150
200
250
0
50
100
150
200
250
Short-Time Fourier Transform(2)

Time Domain

Frequency Domain

Cut into overlapping frames
Low-level Features

10~100ms

Ex: Mel-scale, bark scale, octave
Bark scale

Image from: http://www.ofai.at/~elias.pampalk/ma/documentation.html
Low-level Features

10~100ms

Ex: Mel-scale, bark scale, octave
Timbre

• Timbre’s Characteristics
  – A sound’s timbre is differentiate by the ratio of the fundamental frequency & the harmonics that constitute it.

Image from: http://www.ied.edu.hk/has/phys/sound/index.htm
Timbre Features

• Spectral Based
  – Spectral centroid/rolloff/flux….

• Sub-band Based
  – MFCC, Fourier Cepstrum Coefficient
    • Measure the frequency of frequencies.

• Stereo Panning Spectrum Features
Issues of timbre features

• Fixed-window
• Subtle differences in filter bank range affects the classification performance
• Usually discard phase information
• Usually discard Stereo information
Low-level Features

10~100ms

Input Audio Signal

F1 → Spec1

F K → SpecK

F N → SpecN

FFT/DWT

Subband Decomposition

Subband 1

Subband m

Timbre Feature 1

Subband 1

Subband m

Timbre Feature l

Subband 1

Subband m

Timbre Feature n

Temporal Integration

Temporal Feature 1

Temporal Feature i

Temporal Feature p

Ex: Mel-scale, bark scale, octave
Temporal Features

• The statistical moment (mean, variance,...) of timbre feature (in larger local texture window, few seconds)
  – MuVar, MuCor
  – Be treated as multivariate time series

• Apply STFT on local window
  – Fluctuation pattern(FP), Rhythmic pattern
Fluctuation Pattern

![Graph showing frequency transform over time and frequency axes.](image)
Middle Level Features

• Rhythm 節奏
  – Recurring pattern of tension and release in music

• Pitch 音高
  – Perceived fundamental frequency of the sound

• Harmony 和聲
  – Combination of notes simultaneously, to produce chords, and successively, to produce chord progressions
Rhythm Features

• Beat/Tempo 速度
  – Beat per minute (BPM)

• Beat Histogram (BH)
  – Find the peaks of auto-correlation of the time domain envelope signal
  – Construct histogram of Dominant peaks

• Good performance for Mood Classification

Image from: http://en.wikipedia.org/wiki/Envelope_detector
Pitch Features

• Pitch ≠ Fundamental Frequency
  – Pitch is subjective
  – (Fundamental freq+harmonic series) perceived as a pitch

• Pitch Histogram
• Pitch Class Profiles (Chroma)
• Harmonic Pitch Class Profiles
Pitch Class Profile (Chroma)

Harmony Features

• Chord Progression
• Chord Detection
  – Use the previous pitch features to match with existing chord template

• Usage
  – Not popular in standard music classification works
  – Most used in Cover Song Detection
Choice of Audio Features

• Timbre
  – Suitable for genre, instrument classification
  – Not for melody similarity

• Rhythm
  – Most mood classification used rhythm features

• Pitch/Harmony
  – Not popular in standard classification
  – Suitable for Song similarity, cover song
Song-level feature Representations

waveform

Feature extraction

Feature vectors

Distribution
(Single Gaussian Model, GMM, Kmeans)

One Vector
(Mean, median, codebook model…)

27
Paper Outline

• Audio Features
• Classifiers Learning
  – Classifiers for Music Classification
  – Classifiers for Music Annotation
  – Feature Learning
  – Feature Combination and Classifier Fusion

• Classification Task
• Future Research Issues
Classifier for Music Classification

- K-nearest neighbor (KNN)
- Support vector machine (SVM)
- Gaussian Mixture Model (GMM)
  \[ f(x) = \arg \max_j P(y = j | x) \]
  \[ P(y = j | x) = \frac{P(x | y = j)P(y = j)}{\sum_j P(x | y = j)P(y = j)} \]
- Convolutional Neural Network (CNN)
Classification vs. Annotation

**Genre Classification**

- Song 1 -> Disco
- Song 2 -> Metal
- ... Song N -> Classical

**Single-Label**

**Music Annotation**

- Song 1 -> Disco, Happy, Fast
- Song 2 -> Metal, Angry, Drum
- ... Song N -> Classical, Sad, Violin

**Multi-Label**
Classifier for Music Annotation

- Multiple binary classifier
- Multi-Label Learning version of KNN, SVM
- (Language Model/ Text-IR)
Feature Learning

• (Metric Learning)
  – Find a projection of feature that with higher accuracy
  – Not just feature selection

• Supervised
  – Linear discriminant analysis (LDA)

• Unsupervised
  – Principle Component Analysis (PCA)
  – Non-negative matrix factorization (NMF)
Feature Combination and Classifier Fusion

• Early Fusion
  – Concatenate feature vectors
  – Integrate with classifier learning
    • Multiple kernel learning (MKL)
      – Learn best linear combination of features for SVM classifier

• Late Fusion
  – Majority voting
  – Stacked generalization (SG)
    • Stacking classifiers on top of classifiers
    • Classifier at 2\textsuperscript{nd} level use 1\textsuperscript{st} level prediction results as feature
  – AdaBoost (tree classifier)
Paper Outline

• Audio Features
• Classifiers Learning

• Classification Task
  – Genre Classification
  – Mood Classification
  – Artist Identification
  – Instrument Recognition
  – Music Annotation

• Future Research Issues
Genre Classification Benchmark

Datasets

• GTZAN1000
  – [http://marsyas.info/download/data_sets](http://marsyas.info/download/data_sets)

• ISMIR 2004

• Dortmund dataset
# Genre Classification

<table>
<thead>
<tr>
<th>Reference</th>
<th>Features</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]⁺</td>
<td>{STFT+MFCC}×MuVar+beat+pitch</td>
<td>K-NN</td>
<td>60</td>
</tr>
<tr>
<td>[1]⁺</td>
<td>{STFT+MFCC}×MuVar+beat+pitch</td>
<td>GMM</td>
<td>61</td>
</tr>
<tr>
<td>*[79]</td>
<td>{MFCC}×FP</td>
<td>SVM</td>
<td>77.7 ± 2.8</td>
</tr>
<tr>
<td>*[113]⁺</td>
<td>{MFCC}×GMM</td>
<td>SVM</td>
<td>70.6 ± 3.0</td>
</tr>
<tr>
<td>*[113]⁺</td>
<td>{MFCC}×GMM</td>
<td>SVM</td>
<td>70.4 ± 3.1</td>
</tr>
<tr>
<td>[12]⁺</td>
<td>{STFT+MFCC}×MuVar+beat+pitch</td>
<td>SVM</td>
<td>72 ± 5.1</td>
</tr>
<tr>
<td>[12]⁺</td>
<td>{STFT+MFCC}×MuVar</td>
<td>SVM</td>
<td>71.8 ± 4.8</td>
</tr>
<tr>
<td>[12]⁺</td>
<td>DWCH+STFT+MFCC×MuVar</td>
<td>SVM</td>
<td>78.5 ± 4.1</td>
</tr>
<tr>
<td>*[35]</td>
<td>MFCC×MuCov</td>
<td>SVM</td>
<td>78.6 ± 2.4</td>
</tr>
<tr>
<td>[84]</td>
<td>STFT+MFCC×MuVar²</td>
<td>SVM</td>
<td>79.8</td>
</tr>
<tr>
<td>[9]</td>
<td>STFT+FFT+MFCC+LPC</td>
<td>AdaBoost.DT</td>
<td>82.4</td>
</tr>
<tr>
<td>[16]</td>
<td>CR×NTF</td>
<td>SVM</td>
<td>78.2 ± 3.8</td>
</tr>
<tr>
<td>[18]</td>
<td>{MFCC+ASE+OSC}×FP×LDA</td>
<td>NC</td>
<td>90.6 ± 3.1</td>
</tr>
<tr>
<td>[19]</td>
<td>CR×NTF</td>
<td>SRC</td>
<td>92.4 ± 2.0</td>
</tr>
<tr>
<td>[123]</td>
<td>{MFCC+ASE+OSC}×{MuCov,FP}+beat+chord</td>
<td>SVM (MKL)</td>
<td>90.4</td>
</tr>
<tr>
<td>[123]</td>
<td>{MFCC+ASE+OSC}×{MuCov,FP}+beat+chord</td>
<td>SVM (SG)</td>
<td>90.9</td>
</tr>
</tbody>
</table>

:+ both  
+x : sequence  
* : their implementation  
Use GTZAN dataset

1. MFCC不錯  
2. Pitch/beat看不出好壞  
3. SRC: good classifier,多Feature Combine也不差
Mood Classification

• Difficult to evaluate
  – Lack of publicly available benchmark datasets
  – Difficulty in obtaining the groundtruth
    • Sol: majority vote, collaborative filtering
    • → but performance of mood classification is still influenced by data creation and evaluation process

• Specialty
  – Low-level features (spectral xxx)
  – Rhythm features (effectiveness is debating)
  – Articulation features (only used in mood, smoothness of note transition)
    – Happy/sad → smooth, slow, angry → not smooth, fast

• Naturally Multi-label Learning Problem
Artist Identification

• Subtasks
  – Artist identification (style)
  – Singer recognition (voice)
  – Composer recognition (style)

• MFCC + low order statistics performs well for Artist id and Composer recog

• Vocal/Non-vocal segmentation
  – Most in singer recognition
  – MFCC or LPCC + HMM

• Album Effect
  – Song in the same album too similar to produce overestimate accuracy
Instrument Recognition

• Done at segment level
  – Solo / Polyphonic

• Problem
  – Huge number of combinations of instruments

• Methods
  – Hierarchical Clustering
  – Viewed as multi-label learning (open question)
  – Source Separation (open question)
Music Annotation

- Convert music retrieval to text retrieval
- CAL500 dataset
- Evaluation (view as tag ranking)
  - Precision at 10 of predicted tags
  - Area under ROC (AUC)
- Correlation between tags (apply SG)

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Classifier</th>
<th>Precision-At-10</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MixHier [54]</td>
<td>dMFCC</td>
<td>GMM</td>
<td>0.265 ± 0.007</td>
<td>0.710 ± 0.004</td>
</tr>
<tr>
<td>Autotag [58]</td>
<td>dMFCC</td>
<td>AdaBoost</td>
<td>0.281</td>
<td>0.678</td>
</tr>
<tr>
<td>CBA [65]</td>
<td>dMFCC</td>
<td>CBA</td>
<td>0.286</td>
<td>0.765</td>
</tr>
<tr>
<td>AudioSVM [126]a</td>
<td>dMFCC×MuVar²</td>
<td>SVM</td>
<td>N/A</td>
<td>0.78</td>
</tr>
<tr>
<td>AffinitySVM [126]b</td>
<td>dMFCC×MuVar²</td>
<td>SVM+SG</td>
<td>N/A</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Paper Outline

• Audio Features
• Classifiers Learning
• Classification Task
• Future Research Issues
  – Large-scale content based music classification with few label data
  – Music mining from multiple sources
  – Learning music similarity retrieval
  – Perceptual features for Music Classification
Large-scale Classification with Few Label Data

• Current: thousands of songs

• Scalability Challenges
  – Time Complexity
    • Feature extraction is time consuming
  – Space Complexity
  – Ground Truth Gathering
    • Especially for mood classification task

• Possible Solution
  – Semi-supervised learning
  – Online learning
Music Mining from Multiple Sources

• Social Tags
  – Collect from sites like last.fm
  – Social tags do not equate to ground truth

• Collaborative filtering
  – Correlation between songs in user’s playlist
    • Eg. Song A list by 甲乙丙, song B listen by 乙丙
      \[ \text{sim} = \frac{<(1,1,1),(0,1,1)>}{||(1,1,1)|||(0,1,1)||} \]

• Problem
  – Need test song’s title, artist to gather the above info
  – Possible solution
    • Recursive classifier learning (Use predicted label)
Learning Music Similarity Retrieval

• Previous Retrieval System
  – Predominantly on Timbre similarity
  – Some application focus on melodic/harmonic similarity
    • Cover song detection, Query by humming

• Problem
  – We need “different similarity” for different task
  – Standard similarity retrieval is unsupervised

• Similarity Retrieval based on Learned Similarity
  – Relevance feedback (依照user feedback修改結果)
  – Active learning (每次查完的結果都加進去train)
Perceptual features for Music Classification

• Previously, Low-level feature dominates
  – High-specific, identify exact content
  – Fingerprint, near duplicates

• Middle level feature
  – Models of music
    • Rhythm, pitch, harmony
    • Combine with low-level feature  better results
      – Hard to obtain middle-level feature reliably
  – Models of auditory perception and cognition
    • Cortical representation inspired by auditory model
    • Sparse coding model
    • Convolutional neural network
Conclusion

• Review recent development in music classification and annotation
• Discuss issues and open problems

• There is still much room for music classification
  – Human can identify genre in 10~100 ms
  – There is gap between human and auto performance
THANK YOU 😊