Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

7-2010

Effective Music Tagging through Advanced Statistical Modeling

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Meng WANG
Microsoft Research Asia, Beijing

Shuicheng YAN
National University of Singapore

Hwee Hwa PANG
Singapore Management University, hhpang@smu.edu.sg

Xian-Sheng HUA
Microsoft Research Asia, Beijing

DOI: https://doi.org/10.1145/1835449.1835555

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

Part of the Databases and Information Systems Commons, and the Numerical Analysis and Scientific Computing Commons

Citation

SHEN, Jialie; WANG, Meng; YAN, Shuicheng; PANG, Hwee Hwa; and HUA, Xian-Sheng. Effective Music Tagging through Advanced Statistical Modeling. (2010). *SIGIR 2010: Proceedings: 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Geneva, Switzerland, July 19-23, 2010*. 635-642. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/515

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Effective Music Tagging through Advanced Statistical Modeling

Jialie Shen[†] Meng Wang[‡] Shuicheng Yan[⋆] HweeHwa Pang[†] Xiansheng Hua[‡]
[†] School of Information Systems, Singapore Management University, Singapore
{jlshen, hhpang}@smu.edu.sg

[‡] Microsoft Research Asia, Beijing, China
{mengwang, xhua}@microsoft.com

[⋆] Department of ECE, National University of Singapore, Singapore
eleyans@nus.edu.sg

ABSTRACT

Music information retrieval (MIR) holds great promise as a technology for managing large music archives. One of the key components of MIR that has been actively researched into is music tagging. While significant progress has been achieved, most of the existing systems still adopt a simple classification approach, and apply machine learning classifiers directly on low level acoustic features. Consequently, they suffer the shortcomings of (1) poor accuracy, (2) lack of comprehensive evaluation results and the associated analysis based on large scale datasets, and (3) incomplete content representation, arising from the lack of multimodal and temporal information integration.

In this paper, we introduce a novel system called MMTagger that effectively integrates both multimodal and temporal information in the representation of music signal. The carefully designed multilayer architecture of the proposed classification framework seamlessly combines Multiple Gaussian Mixture Models (GMMs) and Support Vector Machine (SVM) into a single framework. The structure preserves more discriminative information, leading to more accurate and robust tagging. Experiment results obtained with two large music collections highlight the various advantages of our multilayer framework over state of the art techniques.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval, Search process; I.2.m [**Computing Methodologies**]: Artificial Intelligence; H.5.5 [**Sound and Music Computing**]: Systems

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Music Information Retrieval, Tagging, Browsing, Search

1. INTRODUCTION

Music is an unique art form created by human to represent emotion, cultural background, social context and time. To facilitate music information retrieval (MIR) from large music collections, it is necessary to annotate the music documents with comprehensive textual information [2, 16, 24]. Existing music tagging approaches generally perform musical feature extraction, followed by applying machine learning methods to model the relationship between text labels and music features. The approaches hinge on two interrelated issues: (1) the extraction of high quality acoustic features to represent multiple music characteristics, and (2) the judicious application of statistical model(s) for classification and tagging. The effectiveness of the approaches should be demonstrated through a proper evaluation process involving large test collections and appropriate benchmarking metrics.

There has been a long history of using low level acoustic features extracted from audio objects as content descriptors [13, 3, 10, 9]. Unfortunately, how to effectively derive high-level semantic concepts (such as genre and mood) from the physical features still remains an extremely difficult problem. There are several reasons for this. First, there is a gulf between high level concepts and low level acoustic characteristics, as evident by the mismatch in semantic similarity in the search results produced by systems that rely solely on low level features [5]. Second, the content of music is rich and complex, spanning a wide range of features like timbral texture, harmony, rhythm structure and pitch [21, 27, 12, 19, 20, 30]. It is thus imperative to employ a content representation that captures these features comprehensively, and to determine which features to use for what purpose. In view of the challenges, it is not surprising that existing music tagging systems that adopt a simple approach of applying machine learning classifiers directly on low level acoustic features do not deliver good performance.

In this work, we propose a framework called MMTagger (Multifeature based Music Tagger) that combines advanced feature extraction techniques and high level semantic concept modeling for effective annotation of music documents. The basic idea for the proposed scheme is to model music information (text based description) with hierarchical structure and relationship between tags and concepts. It tries to map sound documents to a representation in the so-called *latent musical concept space*, where relevance between documents and tags can be more accurately modeled than in the acoustic feature space. The MMTagger's architecture comprises three interconnected functionality layers. The technical design of the first layer aims at not only providing high quality feature combination but also to incorporate temporal information. The latter is motivated by the observation that music documents belonging to the same category generally share certain temporal patterns. The second layer of the proposed system is for discriminative musical concept modeling, and is intended to bridge the 'semantic gap' between the low level music features in the first layer, and the music tags in the third layer. Here, we utilize multiple Gaussian Mixture Models (GMMs) to represent different concepts [15, 7]. Since a semantic concept could be relevant to many different keywords, the third layer contains multiple support vector machines (SVM), each trained to derive the likelihood score of a tag from its association strength with the various music concepts. We have conducted a comprehensive experiment study with two large test collections. The results indicate that our solution achieves substantial performance improvement in accuracy and robustness in annotating music documents.

The rest of the article is structured as below: Section 2 gives a brief overview of related work in the area of music tagging, including their assumptions and limitations. In Section 3, we provide details on our proposed architecture and introduce the structure of each system component module and its learning algorithms. Section 4 reports on our experiment configuration while Sections 5 presents empirical evaluation results. Finally, our conclusions and directions for future research are summarized in Section 6.

2. RELATED WORK

Automated music tagging is an important research problem with numerous applications such as music search and music recommendation. This area has received considerable attention and many related techniques have been developed in recent years. Among the earliest of such systems, Whitman and Rifkin [29, 28] proposed a novel Regularized Least-Squares Classification (RLSC) based approach. The goal is to derive non-linear relationship between text captions and acoustic features in an efficient way. For performance evaluation, 255 songs from 51 performers are separated into training and testing sets with roughly equal size. Using the SVM classifier, accuracy achieved ranges from 0.0% to 38.9% depending on the terms used for evaluation process. In [23], Turnbull et al. applied a supervised multiclass naïve Bayes model to estimate relationship between musical sound and words. The features considered by this system can be classified into two categories - textual features and audio features (e.g., dMFCC and auditory filter-bank temporal envelope features). The test collection contains totally 2,131 songs and their song reviews. Using dMFCC feature, precision and recall rates achieved by the system is 0.072 and 0.119 for the annotation task. To facilitate effective music retrieval with semantic description, Turnbull et al. developed a music labeling scheme based on the supervised multi-class labeling model (SML) [26, 25][1]. In this approach, sound documents are modeled as a GMM distribution over a set of predefined terms (corpus). The distance between the multinomial distributions of keyword query and a music feature can be estimated with the Kullback-Leibler (KL) divergence for the purpose of ranking search results. The acoustic feature considered in this system is MFCC. Using the CAL500 dataset, they achieved a nice performance improvement in retrieval and annotation accuracy. More recently, Duan et al. designed an interesting approach for collective annotation of music data [6]. It assumes that there are certain levels of correlation between different tags. Studying the relationship is useful for improving annotation performance. They employed two different statistical models - GMMs and Conditional Random Field to exploit the label correlation. Experiment results demonstrate a small but consistent performance gain. In addition, Bertin-Mahieux et al. proposed Autotagger system using advanced ensemble learning schemes to combine discriminative power of different classifiers [8, 4]. Those schemes include AdaBoost and FilterBoost. Acoustic features considered by the scheme include 20 MFCC Coefficients, 176 autocorrelation coefficients, and 85 spectrogram coefficients. Experiment results based on the CAL500 dataset and another large test collection demonstrate that Autotagger performs better than MSML. It is currently the most advanced technique for music tagging.

Symbols	Definitions
C	Total number of high level music concepts
s	Notation of music segment s
f	Notation of feature f
F	Total number of acoustic features extracted
t	Notation of tag t
T	Total number of tags
G^c	GMMs for music concept c
w_k	Weight of the kth Gaussian component
μ_k	Mean of the kth Gaussian component
Σ_k	Covariance matrix of the kth Gaussian component
K	Number of mixture components in GMMs
V	Vocabulary of test collection
$\|V\|$	Size of vocabulary
A	Annotation length
M	Transformation matrix
ms_s	Music segment s
ts_s	Starting time of music segment s
te_s	End time of music segment s
p_t	Probability for music tag t
λ	Likelihood vector generated by DCML
$\mathbf{r}$	Tag relevance vector generated by TRL

Table 1: Summary of symbols and definitions

3. A TAGGING FRAMEWORK WITH MULTILAYER STRUCTURE

This section presents a novel scheme to facilitate effective automated tagging over large music collections. As illustrated in Figure 1, the architecture of our system consists of three functionality layers: music preprocessing layer (MPL) for music sequence segmentation and acoustic feature extraction, discriminative concept modeling layer (DCML) and SVM based tag refinement layer (TRL). Similar to an Artificial Neural Network [7], each layer is fully connected with each other. The first layer MPL aims to extract four different features including timbral feature, spectral feature, rhythm feature and melody feature. Using those features, a set of statistical models based on GMMs are constructed to characterize high level musical concepts in database, one GMMs per concept (e.g., genre, singer, mood). Those concepts can be treated as the most fundamental component of

[1] In this paper, we use MSML to denote this system.

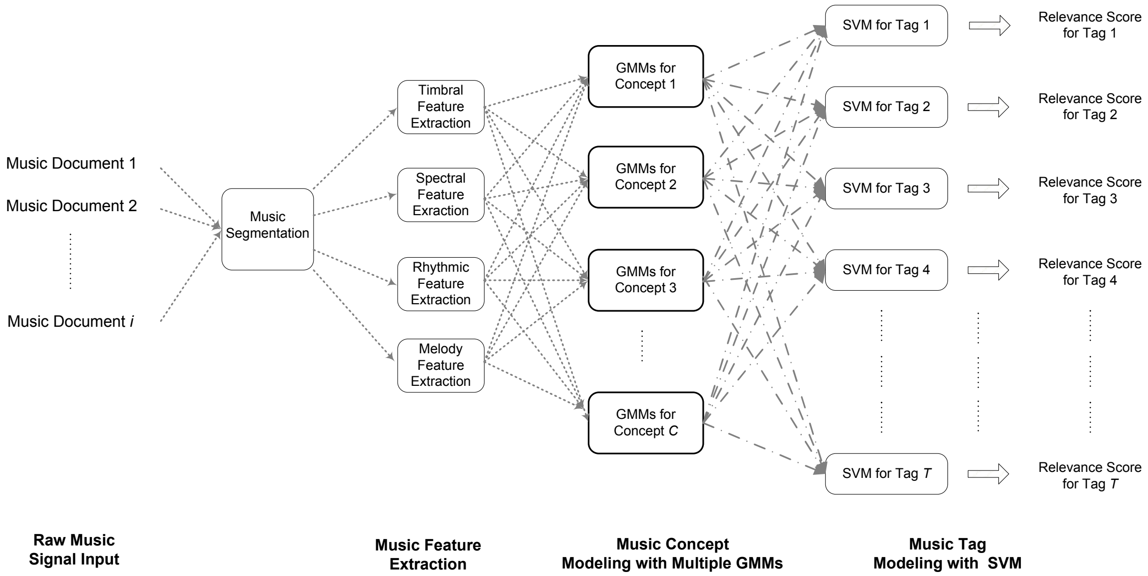
Music Document 1
Music Document 2
Music Document *i*
Music Segmentation
Timbral Feature Extraction
Spectral Feature Exraction
Rhythmic Feature Extraction
Melody Feature Extraction
GMMs for Concept 1
GMMs for Concept 2
GMMs for Concept 3
GMMs for Concept *C*
SVM for Tag 1
SVM for Tag 2
SVM for Tag 3
SVM for Tag 4
SVM for Tag *T*
Relevance Score for Tag 1
Relevance Score for Tag 2
Relevance Score for Tag 3
Relevance Score for Tag 4
Relevance Score for Tag *T*
Raw Music Signal Input
Music Feature Extraction
Music Concept Modeling with Multiple GMMs
Music Tag Modeling with SVM

Figure 1: Architecture of MMTagger scheme.

latent musical concept space. The output of this layer is a set of likelihood scores, which serve as input to TRL. The TRL contains a collection of SVM based tag classifier, each is trained to generate relevance score for a tag. Based on the relevance scores, we can finally rank each tag and pick the top N tags to annotate the input music. The following sections elaborate on each of the layers and give a full description of the associated algorithms.

3.1 Music Preprocessing Layer

The function of this layer is to preprocess raw audio signal and compute music features. The related process comprises two steps: music segmentation and feature extraction. When an audio signal is received, it is first partitioned into several short fixed length time-frames. For this study, we set the length of each frame to be 0.5 second. Distinguished from previous tagging schemes, we introduce a temporal descriptor in the music content representation. Its main advantage is better content description capability through combining both acoustic information and temporal information. For a segment with starting time ts_s and end time te_s, the corresponding temporal musical descriptor is defined as,

$$td_f(m_i, ms_s) = extra_f(m_i, ms_s) \qquad (1)$$

where $td_f(m_i, ms_s)$ denotes the feature f calculated from the segment $ms_s = (ts_s, te_s)$ of music file m_i and $extra_f$ is an extraction function for feature f. Each music document is treated as a composite of different kinds of acoustic feature vectors with temporal information. The motivation derives from the observation that discriminative characteristics are often embodied within local temporal acoustic features. Thus the proposed temporal based feature enjoys greatest potential to provide more comprehensive summarization for the purpose of classification. The MMTagger system considers four different kinds of music features:

- **Timbral features (TF)** characterize the timbral property of music objects. We apply short time Fourier transform in the calculation. The timbral features computed include *Mel-Frequency Cepstral Coefficients* (MFCCs) [13], *Spectral Centroid*, *Rolloff*, *Flux*, *Low-Energy feature* [27], and *Spectral Contrast* [14]. The total dimensionality of timbral features calculated is 20.
- **Spectral features (SF)** characterize the spectral composition of music signal. In our implementation. each spectral feature vector contains *Auto-regressive (AR) features*; *Spectral Asymmetry, Kurtosis, Flatness, Crest Factors, Slope, Decrease, Variation*; *Frequency Derivative of Constant-Q Coefficients*; and *Octave Band Signal Intensities* [14]. The total dimensionality of these feature vectors is 20.
- **Rhythmic features (RF)** summarize the patterns of a music object over a certain duration. The rhythmic features calculated in this study include: *Beat Histogram* [27]; *Rhythm Strength*, *Regularity* and *Average Tempo* [14]. The total dimensionality is 12.
- **Melody features (MF)** describe the pitch content and its duration in a music document. The *Pitch Histogram* proposed in [27] is used as melody features in our proposed framework. The total dimensionality of this group of features is 48.

Accordingly, the final content representation includes four different groups of musical features (local content information) and time information (temporal information). Total dimensionality of the feature set considered in this study is 100.

3.2 Discriminative Concept Modeling Layer

For the second layer of the proposed MMTagger system, multiple GMMs are trained to statistically model the relationship between each high level concept and various acoustic features. Those high level concepts constitute the *latent musical concept space*. Each high level music concept corresponds to one GMMs. GMMs is among the most widely applied statistical analysis methods due to its flexibility of representing different kinds of distributions. However, gaining an accurate estimation of the distribution of the music features associated with to a high level concept goes beyond a straightforward application of the GMM method. While the distance between two music concepts can be estimated via the KL divergence between their GMMs, accurate result cannot be expected in general. There are two main reasons:

- Due to the limited number of learning examples (music clips) for a certain music concept, it is very hard to estimate the parameters of a GMM robustly and accurately.
- The KL divergence between GMMs does not take the concept label information into account and consequently can result in poor discriminative ability.

To solve those problems, we develop a two-step adaptation approach to construct the GMMs based on adaptive learning [1]. It includes generative adaptation and discriminative music concept adaptation. Generative adaptation has been widely explored in many tasks such as speaker identification and image categorization [11, 18]. It tries to use all learning samples in the training process and then obtain the Universal Background Model (UBM), which is the GMMs optimized by the principle of Maximum a Posteriori (MAP). In the second step, discriminative music concept adaptation is designed to adjust the mean vectors of each GMMs to achieve the targets of 1) keeping the music documents belonging the same semantic concept closer, and 2) separating out those with different labels. In this way, the obtained GMMs exhibit better classification capability.

3.3 Generative Adaptation

The UBM obtained in the initial phase of training the GMMs is denoted as,

$$G = P(x|\theta) = \sum_{k=1}^{K} w_k N(x;\mu_k,\Sigma_k) \qquad (2)$$

where w_k, μ_k and Σ_k are the weight, mean and covariance matrix of the kth Gaussian component, respectively. x is input feature vector. K is the total number of Gaussian components and the probabilistic density is calculated as a weighted combination of K Gaussian densities,

$$N(x;\mu_k,\Sigma_k) = \frac{e^{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)}}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}. \qquad (3)$$

The parameters of UBM are estimated using the traditional EM algorithm. In the E-step, the posterior probability is calculated via

$$Pr(k|x_i) = \frac{w_k N(x;\mu_k,\Sigma_k)}{\sum\limits_{k=1}^{K} w_k N(x;\mu_k,\Sigma_k)}, \qquad (4)$$

where $n_k = \sum\limits_{k=1}^{K} Pr(k|x_i)$,and the M-step updates the mean vectors via

$$\hat{\mu}_k = \frac{1}{n_k + r}\sum_{i=1}^{n} Pr(k|x_i)x_i + \frac{r}{n_k + r}\mu_k \qquad (5)$$

When EM iteration stops, the resulting GMMs is the Universal Background Model (UBM).

3.4 Discriminative Concept Adaptation

The purpose of discriminative concept adaptation is to estimate the GMMs parameter belonging to a certain music concept from a UBM. After it, a series of GMMs $\{G^1, ..., G^C\} = \{P^1(x|\theta^1), P^2(x|\theta^2), ..., P^C(x|\theta^C)\}$ are constructed with each GMMs approximating distribution over the audio feature space of a high level concept. A special transformation matrix M on the mean vectors of the GMMs is designed to enhance classification capability further. Since each high level music concept is represented by a GMMs, the distance between two concepts can be measured using the KL divergence between their GMMs. However, since the KL divergence of GMMs is not analytically tractable, we use the upper bound of the divergence for calculating distance. It can be proved that

$$\begin{aligned} D(G^a||G^b) &\leq \sum_{k=1}^{K} w_k D(N(x;\mu_k^a,\Sigma_k)||N(x;\mu_k^b,\Sigma_k)) \\ &\approx \frac{1}{2}\sum_{k=1}^{K} w_k(\mu_k^a - \mu_k^b)^T\Sigma_k^{-1}(\mu_k^a - \mu_k^b) \qquad (6) \end{aligned}$$

where μ_k^a and μ_k^b respectively denote the mean of the kth component from music concept a and b. The current model cannot yet achieve optimal classification performance because inter-class and intra-class distances are not taken into account. To improve its effectiveness, Neighborhood Component Analysis (NCA) is performed to derive transformation matrix M and apply it on $D(G^a||G^b)$. NCA is a learning method, which constructs a distance metric optimizing leave-one-out (LOO) performance based on training data. An infinitesimal change in A may change the neighbor graph and thus lift LOO discrimination power by a finite amount. NCA adopts a more well behaved measure of nearest neighbor performance by introducing a differentiable cost function based on stochastic neighbor assignment in the transformed space. Each point i in multidimensional feature space selects another point j as its neighbor with certain probability p_{ij}, and inherits its class label from the selected point. p_{ij} is defined as

$$p_{ij} = \frac{e^{(-D_{ij})}}{\sum_{k\neq j} D_{jk}} \qquad (7)$$

where $D_{ij} = ||Mx_i - Mx_j||^2$. The objective is to maximize the expected number of samples correctly classified. Thus, we have,

$$f(M) = \sum_i \sum_{j\in C_i} \frac{exp(-D_{ij})}{\sum_{k\neq j} D_{jk}} \qquad (8)$$

where C_i denotes the set of samples in the same class as i-th sample. After carrying out differentiation with respect to the transformation matrix M, we can obtain,

$$\frac{\partial f}{\partial M} = -\sum_i \sum_{j\in C_i} p_{ij}(q_{ij} - \sum_k p_{ik}q_{ik}) \qquad (9)$$

where

$$q_{ik} = \sum_{k=1}^{K} w_k \Sigma_k^{-1} M(\mu_k^a - \mu_k^b)(\mu_k^a - \mu_k^b) \quad (10)$$

The optimization problem above can be easily solved with a gradient descent process. After the training process, the GMMs for each high level music concept estimates the likelihood score λ_c and this score is used to quantify the distance between raw feature input and concept label. At the same time, the output of DCML is a vector λ that models probabilistic relationship between different music concepts and audio input, where $\lambda = [\lambda_1, \lambda_2, ..., \lambda_C]$. It serves as input for tag refinement layer, the last layer in the MMTagger framework.

3.5 Tag Refinement Layer

The SVM based computational nodes constitute the Tag Refinement Layer (TRL) of the MMTagger framework. Each of them is designed to estimate the probability of a particular tag based on input from DCML. In the current implementation of MMTagger system, SVM is selected as a tag classifier due to its effectiveness [22].

Since traditional SVM can be used only for binary classification, the method proposed by Hastie and Tibshirani [17] is employed to derive numeric value for the probability that an unknown sample belongs to a certain tag. Its key idea is to adopt Gaussian distribution to model tag-conditional densities $p_t(\lambda|y=1)$ and $p_t(\lambda|y=-1)$. Using Bayes' rule, the relevance score (posterior probability) r_t for each given tag t can be computed via:

$$r_t = P_t(y=1|\lambda) = \frac{p_t(\lambda|y=1)P_t(y=1)}{\sum_{i=-1,1} p_t(\lambda|y=i)P_t(y=i)} \quad (11)$$

where $\lambda = [\lambda_1, \lambda_2, ..., \lambda_C]$ is a vector generated by DCML. It contains a set of likelihood values describing the probability that an input music belongs to music concept c, where $c = 1, ..., C$. Using equation 11, a set of relevance scores $\mathbf{r} = [r_1, r_2, ..., r_T]$ are obtained for ranking tags. Eventually the top k tags are selected as annotation for input music, with value of k being predefined by the user.

4. EXPERIMENTAL CONFIGURATION

This section introduces the experiment configuration for our performance evaluation. We report details on two test datasets, performance metrics, competitors and evaluation methodology. All tagging methods evaluated in this study have been fully implemented and tested on a Pentium (R) D, 3.20GHz, 1.98 GB RAM PC running the Windows XP operating system.

4.1 Test Collections

Test collections play a very important role in empirical study. Two test collections are used in this evaluation. The first one (TS1) is the Computer Audition Lab 500-Song (CAL 500) data set developed by the CAL group [25, 26]. This collection contains 500 modern western music documents performed by 500 different artists. Altogether, there are 174 tags categorized into six different semantic groups including instrumentation, vocal characteristics, genre, emotions, solo and usage terms. For this dataset, we use those six groups as high level musical concepts to train our statistical model.

Since the size of the CAL500 data set is relatively small, we developed the second test collection called TS2. It contains 4000 popular music items downloaded from Youtube. They are performed by 110 different singers including 55 females and 55 males. The music documents are converted to 22050Hz, 16-bit, mono audio files. 12 amateur musicians, who are familiar with various music taxonomy and concepts, were hired to create ground truth about this collection. The ground truth information was generated by attaching a tag to a music item if at least three people assigned the tag to the song. In the case that an agreement on tag assignment between different respondents can not be reached, a similar resolution used in generating CAL500 is applied. At the end of the process, we obtain totally 250 tags, belonging to 8 different categories. They are instrumentation, emotions, country, time, genre, vocal characteristics, solo and usage terms. Consequently, our evaluation with TS2, involves 8 high level concepts. The size of the vocabulary $|V|$ is 174.

4.2 Evaluation Metrics and Methodology

Textual information generated by music tagging systems can be used for many different MIR applications. To validate the performance of different tagging schemes, we select two MIR tasks:

- Task 1 - music annotation: for a given song track issued by the user, determine a proper set of tags. In this study, the size of tag set is 10.
- Task 2 - music search: for a given tag selected from the vocabulary, search for relevant song tracks in the test collection.

Here, three different evaluation metrics are used for music annotation. They are mean per-tag precision and recall, and the F-score. Based on the methodology used by Turnbull et al. [26], the top 10 tags generated by the models are used for comparison and thus annotation length A is 10. Per-tag recall and per-tag precision is formally defined as

$$Precision = \frac{|t_{TP}|}{|t_{GT}|} \qquad Recall = \frac{|t_{TP}|}{|t_A|} \quad (12)$$

where $|t_{GT}|$ is the number of songs annotated with the tags in the human-generated "ground truth" annotation and t_{TP} is the number of songs annotated correctly with the tags. Based on per-tag precision and per-tag recall, the F-score is defined as

$$F - score \quad = \quad 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

To measure the performance of different approaches for music search, the mean average precision (MeanAP) and the area under the receiver operating characteristic curve (AROC) are adopted as assessment metrics. Given a query tag, the focus of MeanAP is on finding the most relevant songs, while AROC emphasizes whether relevant songs are ranked higher than irrelevant ones. We also apply α-fold cross validation to ensure the stability and robustness of the empirical results. α is predefined to be 5.

In this study, we compare the performance of our system against two state-of-the-art approaches including Autotagger [8, 4] and MSML [26, 25]. Acoustic feature considered by MSML is Mel-frequency cepstral coefficient (MFCC). Autotagger is evaluated based on three feature sets including MFCC delta, afeats and bfeats [2] For MMTagger, we consider five low level feature configurations (timber features denoted by TF, rhythm features denoted by RF, spectral features denoted by SF, melody features denoted by MF and timber features+rhythm features+spectral features+melody features denoted by ALL.). Autotagger(MFCC delta), Autotagger(afeats) and Autotagger(bfeats) denote Autotagger with MFCC delta, afeats and bfeats respectively. MMTagger(TF), MMTagger(SF), MMTagger(MF), MMTagger(RF), MMTagger(ALL) denote our proposed model with timbral features, spectral features, rhythmic features, melody features and the combination of all four musical features.

5. EXPERIMENT RESULTS

This section presents an experiment study to evaluate the competing techniques on the task of music annotation, retrieval as well as music annotation in noisy environment.

5.1 Result Analysis for Music Annotation

We report a comparative study of the various tagging systems on music annotation processing. Table 2 and 3 summarize the empirical results for three systems with various configurations on the two test collections. The size of tag set is set to 10. Here, MMTagger is tested with five different feature settings is tested. The bottom four rows of both tables present the accuracy of our proposed system with just one acoustic feature. In comparison to MMTagger(ALL), they suffer from lower accuracy. In fact, the multiple feature combination achieves significant effectiveness gain ranging from 5% to 15%. The empirical results points clearly to the importance of combining features intelligently to tagging effectiveness. The experimental results also demonstrate that the MMTagger significantly outperforms the existing approaches. For example, in Table 2, comparing to Autotagger(bfeats), MMTagger(ALL) improves the precision from 0.291 to 0.351 for the CAL500 dataset, and from 0.268 to 0.327 for the TS2 collection. Similar observations can be made on the other two evaluation metrics over different test collections. We thus conclude that MMTagger emerges as the most effective music tagging scheme.

Model	Precision	Recall	F-Score
MSML	0.144	0.064	0.089
Autotagger(MFCC delta)	0.281	0.131	0.179
Autotagger(afeats)	0.266	0.094	0.139
Autotagger(bfeats)	0.291	0.153	0.205
MMTagger(ALL)	0.351	0.291	0.314
MMTagger(TF)	0.256	0.141	0.176
MMTagger(SF)	0.241	0.137	0.165
MMTagger(MF)	0.226	0.131	0.149
MMTagger(RF)	0.289	0.150	0.171

Table 2: Tagging accuracy on test collection CAL500(TS1).

Model	Precision	Recall	F-Score
MSML	0.121	0.043	0.072
Autotagger(MFCC delta)	0.257	0.102	0.101
Autotagger(afeats)	0.239	0.073	0.117
Autotagger(bfeats)	0.268	0.139	0.186
MMTagger(ALL)	0.327	0.241	0.284
MMTagger(TF)	0.231	0.117	0.144
MMTagger(SF)	0.220	0.116	0.139
MMTagger(MF)	0.207	0.103	0.125
MMTagger(RF)	0.262	0.125	0.156

Table 3: Tagging accuracy on test collection TS2.

5.2 Result Analysis for Music Retrieval

With the wide availability of large music collections, accurate music search is mandatory to achieve usability. This section presents empirical results to compare the accuracy of music retrieval facilitated by our proposed scheme and the two competitors. Experimental methodology is that given a keyword query kw_q in vocabulary V, a test set of songs are ranked. The metrics MeanAP and MeanAROC of each ranking are calculated for performance comparison. Tables 4 and 5 summarize the experiment results with CAL500(TS1) and TS2. Clearly, the proposed MMTagger(ALL) significantly outperforms the other approaches.

In particular, the results shows that relative to Autotagger, MMTagger enjoys at least 10% MeanAP increase on both test collections. Although a nice improvement over Autotagger can be observed, we find that there is more significant gain over MSML. Averagely around 21% lift in term of accuracy can be found for the two datasets. In addition, we can summarize that for our proposed system, a proper integration of multiple music features can bring substantial improvement for search and annotation effectiveness. This observation corroborates other researchers' finding that accurate MIR can not be achieved with a single type of music feature and development of effective acoustic feature combination scheme plays important role for tagging system's performance enhancement.

Model	MeanAP	MeanAROC
MSML	0.231	0.503
Autotagger(MFCC delta)	0.305	0.678
Autotagger(afeats)	0.323	0.622
Autotagger(bfeats)	0.340	0.662
MMTagger(ALL)	0.410	0.782
MMTagger(TF)	0.282	0.496
MMTagger(SF)	0.275	0.489
MMTagger(MF)	0.286	0.501
MMTagger(RF)	0.288	0.508

Table 4: Music retrieval accuracy on test collection CAL500(TS1).

[2] Detail information about those feature sets can be found in [26].

Model	MAP	MeanAROC
MSML	0.204	0.461
Autotagger(MFCC delta)	0.267	0.613
Autotagger(afeats)	0.289	0.597
Autotagger(bfeats)	0.301	0.626
MMTagger(ALL)	0.385	0.753
MMTagger(TF)	0.251	0.449
MMTagger(SF)	0.257	0.452
MMTagger(MF)	0.249	0.467
MMTagger(RF)	0.242	0.472

Table 5: Music retrieval accuracy on test collection TS2.

5.3 Result Analysis for Noise Robustness

Modern MIR systems often need to work robustly in presence of ambient noise (e.g. raw music signal recorded from live concerts or other outdoor environments). However, existing schemes might not perform effectively when handling noisy audio input. Thus, it is important to evaluate the robustness of different music annotation schemes against music sources containing different kinds of audio distortion. In this work, we study how different types of noise in the query music affect annotation accuracy. We use the TS1 (CAL500) as evaluation data and apply the same set of test music as those used in the music annotation experiment. Before the test, various kinds of audio distortion are injected into each query music item. The distortion cases include 50% volume amplification, 50% volume deamplification, 10 second cropping, 35dB SNR mean background noise and 35db SNR white background noise[3].

Tables 6-10 illustrate the noise robustness performance of MMTagger and its competitors for the different distortion cases. In general, the results show that when the music input is polluted by a certain kind of noise, annotation accuracy of all the systems suffers. Comparing to the other approaches, MMTagger demonstrates high resilience and stable performance. Specifically, MMTagger using single kind of acoustic feature suffers greater performance degradation than MMTagger with all four acoustic features. Moreover, crosschecking the results from this section and Section 5.1 reveals that MMTagger demonstrates more stable performance and enjoys less accuracy degradation than Autotagger and MSML under noisy circumstances. For example, MMTagger's precision decreases about 7% when annotating inputs with 50% volume amplification. Whereas, Autotagger and MSML suffer about 15% and 17% drop on average. We thus conclude that MMTagger is robust to different kinds of noise.

6. CONCLUSION

As a key enabling technology for music information retrieval, tagging has received a lot of research attentions in recent years. However, the performance of existing systems is far from satisfactory. In this paper, we describe a new music annotation scheme based on advanced feature extraction and multilayer structure. We have applied our method to two large test collections. Theoretical analysis and empirical results indicate that our approach achieves substantially

[3] $SNR_{dB} = 10log_{10}\frac{S}{N}$ is the equation used to calculate the signal-to-noise ratio, where S denotes the signal power, and N denotes the noise power in dB

Model	Precision	Recall	F-Score
MSML	0.121	0.053	0.075
Autotagger(MFCC delta)	0.241	0.112	0.154
Autotagger(afeats)	0.226	0.084	0.113
Autotagger(bfeats)	0.252	0.130	0.167
MMTagger(ALL)	0.325	0.271	0.291
MMTagger(TF)	0.206	0.121	0.133
MMTagger(SF)	0.201	0.107	0.127
MMTagger(MF)	0.216	0.101	0.119
MMTagger(RF)	0.253	0.117	0.143

Table 6: Tagging accuracy in noisy environment on test collection CAL500(TS1). Noise type - 50% volume amplification.

Model	Precision	Recall	F-Score
MSML	0.132	0.059	0.071
Autotagger(MFCC delta)	0.239	0.117	0.167
Autotagger(afeats)	0.231	0.088	0.119
Autotagger(bfeats)	0.248	0.125	0.171
MMTagger(ALL)	0.321	0.281	0.295
MMTagger(TF)	0.212	0.128	0.131
MMTagger(SF)	0.209	0.111	0.119
MMTagger(MF)	0.214	0.108	0.121
MMTagger(RF)	0.249	0.118	0.145

Table 7: Tagging accuracy in noisy environment on test collection CAL500(TS1). Noise type - 50% volume deamplification.

Model	Precision	Recall	F-Score
MSML	0.127	0.051	0.080
Autotagger(MFCC delta)	0.237	0.120	0.150
Autotagger(afeats)	0.230	0.085	0.121
Autotagger(bfeats)	0.251	0.131	0.162
MMTagger(ALL)	0.330	0.281	0.211
MMTagger(TF)	0.211	0.128	0.135
MMTagger(SF)	0.211	0.116	0.120
MMTagger(MF)	0.209	0.111	0.121
MMTagger(RF)	0.249	0.113	0.145

Table 8: Tagging accuracy in noisy environment on test collection CAL500(TS1). Noise type - 10 second cropping.

Model	Precision	Recall	F-Score
MSML	0.117	0.064	0.081
Autotagger(MFCC delta)	0.240	0.110	0.160
Autotagger(afeats)	0.221	0.081	0.109
Autotagger(bfeats)	0.259	0.133	0.171
MMTagger(ALL)	0.332	0.271	0.291
MMTagger(TF)	0.201	0.124	0.130
MMTagger(SF)	0.208	0.106	0.129
MMTagger(MF)	0.212	0.105	0.111
MMTagger(RF)	0.252	0.112	0.146

Table 9: Tagging accuracy in noisy environment on test collection CAL500(TS1). Noise type - 35dB SNR mean background noise

Model	Precision	Recall	F-Score
MSML	0.115	0.061	0.083
Autotagger(MFCC delta)	0.242	0.108	0.156
Autotagger(afeats)	0.219	0.082	0.107
Autotagger(bfeats)	0.256	0.130	0.176
MMTagger(ALL)	0.328	0.268	0.287
MMTagger(TF)	0.204	0.127	0.127
MMTagger(SF)	0.207	0.111	0.125
MMTagger(MF)	0.209	0.112	0.113
MMTagger(RF)	0.245	0.109	0.140

Table 10: Tagging accuracy in noisy environment on test collection CAL500(TS1). Noise type - 35dB SNR white background noise

higher accuracy in tagging music documents comparing to existing techniques. Moreover, our method demonstrates superior robustness against different kinds of audio distortion.

This work can be extended in several directions: At this stage, we have only tested our method on audio data. It would be very interesting to apply the method to data from other application domains (e.g. image and video retrieval) and investigate its performance characteristics. In addition, we plan to integrate more acoustic features into our framework. A natural question arises as to what kinds of feature combination is best in terms of effectiveness and robustness enhancement. Finally, designing a robust and effective evaluation methodology is also very important for further investigation and fair performance comparison.

7. ACKNOWLEDGEMENTS

Shuichang Yan is partially supported by AcRF Tier-1 Grant of R-263-000-464-112, Singapore.

8. REFERENCES

[1] E. Alpaydin. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] M. Bartsch and G. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[4] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2), 2008.

[5] C. Dorai and S. Venkatesh. Bridging the semantic gap with computational media aesthetics. *IEEE Multimedia*, 10(2), 2003.

[6] Z. Duan, L. Lu, and C. Zhang. Collective annotation of music from multiple semantic categories. In *Proc. of ISMIR*, 2008.

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[8] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Proc. of NIPS*, 2007.

[9] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transaction on Speech and Audio Processing*, 2:578–589, 1994.

[10] N. Hu, R. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 185–188, 2003.

[11] C. Lee, C. Lin, and B. Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Transactions on Signal Processing*, 39(4), 1991.

[12] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. of ACM SIGIR Conference*, 2003.

[13] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. of the ISMIR*, 2000.

[14] L. Lu, D. Liu, and H. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Acoust., Speech, Signal*, 2006.

[15] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[16] N. Orio. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1), 2006.

[17] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 2000.

[18] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, Z.-J. Zha, and H.-J. Zhang. A joint appearance-spatial distance for kernel-based image categorization. In *Proc. of CVPR*, 2008.

[19] J. Shen, B. Cui, J. Shepherd, and K.-L. Tan. Towards efficient automated singer identification in large music databases. In *Proc. of ACM SIGIR Conference*, pages 59–66, 2006.

[20] J. Shen, J. Shepherd, B. Cui, and K.-L. Tan. A novel framework for efficient automated singer identification in large music databases. *ACM Trans. Inf. Syst.*, 27(3), 2009.

[21] J. Shen, J. Shepherd, and A. H. H. Ngu. Towards effective content-based music retrieval with multiple acoustic feature combination. *IEEE Transactions on Multimedia*, 8(6):1179–1189, 2006.

[22] S. Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proc. of ICML*, 2008.

[23] D. Turnbull, L. Barrington, and G. Lanckriet. Modeling music and words using a multi-class naïve bayes approach. In *Proc. of ISMIR*, 2006.

[24] D. Turnbull, L. Barrington, G. R. G. Lanckriet, and M. Yazdani. Combining audio content and social context for semantic music discovery. In *Proc. of ACM SIGIR Conference*, pages 387–394, 2009.

[25] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the CAL500 data set. In *Proc. of ACM SIGIR Conference*, 2007.

[26] D. Turnbull, L. Barrington, D. Torres, and G. R. G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech & Language Processing*, 16(2), 2008.

[27] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 2002.

[28] B. Whitman. *Learning the meaning of music*. PhD thesis, Massachusetts Institute of Technology, 2005.

[29] B. Whitman and R. M. Rifkin. Musical query-by-description as a multiclass learning problem. In *Proc. of IEEE Workshop on Multimedia Signal Processing*, 2002.

[30] B. Zhang, J. Shen, Q. Xiang, and Y. Wang. Compositemap: a novel framework for music similarity measure. In *Proc. of ACM SIGIR*, pages 403–410, 2009.