6-2015

EMIF: Towards a scalable and effective indexing framework for large scale music retrieval

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

Tao MEI

Dacheng TAO

Xuelong LI

Yong RUI

Published in ICMR '15: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval: Shanghai, June 23-26, 2015, pp. 543-546.
http://doi.org/10.1145/2671188.2749346

EMIF: Towards a Scalable and Effective Indexing Framework for Large Scale Music Retrieval

Jialie Shen[†], Tao Mei[§], Dacheng Tao[⋆], Xuelong Li[‡], and Yong Rui[§]
[†]Singapore Management University, Singapore
[§]Microsoft Research Asia, Beijing, China
[⋆]University of Technology, Sydney, Australia
[‡]Xi′an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi′an, China
jlshen@smu.edu.sg; {tmei, yongrui}@microsoft.com; dacheng.tao@gmail.com; xuelongli@ieee.org

ABSTRACT

This article presents a novel indexing framework called EMIF (Effective Music Indexing Framework) to facilitate scalable and accurate content based music retrieval. EMIF system architecture is designed based on a "classification-and-indexing" principle and consists of two main functionality layers: 1) a novel semantic-sensitive classification to identify input music's category and 2) multiple indexing structures - one local indexing structure corresponds to one semantic category. EMIF's layered architecture not only enables superior search accuracy but also reduces query response time significantly. To evaluate the system, a set of comprehensive experimental studies have been carried out using large test collection and EMIF demonstrates promising performance over state-of-the-art approaches.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]; H.5.5 [**Sound and Music Computing**]: Systems

Keywords

Content based Music Retrieval; Indexing Structure; Retrieval

1. INTRODUCTION

The last decade has witnessed a fast growth in digital music from various application domains [9]. To support effective management of such a large volume of music data, how to improve search efficiency becomes very important issue [3]. Particularly, there has been growing interest in studying indexing techniques for content based music retrieval (CBMR). The typical examples include the M-tree [15], LSH [16, 17], CM*F [11], QUC-Tree [12] and so on. Although the approaches have demonstrated promising performance in certain applications, there are still many open questions and unsolved research issues. First, we believe that while main research focus for developing multi-dimensional indexing scheme is effeciency improvement, how to enhance search accuracy has been becoming more and more important. Unfortunately, much less attention has been paid on it in the previous work. Further, most of existing indexing schemes for CBMR is developed based on "feature transformation" paradigm. In general, it has two nested issues: (1) how to compute small but effective signature to represent complex music contents based on low level acoustic features, and (2) how to design advanced architecture to facilitate fast and accurate search.

Motivated by the analysis given above, this paper presents a novel indexing framework called EMIF [1] to facilitate effective and efficient music information retrieval. Distinguished from previous approaches, architecture of our technique is designed based on a "Classify-and-Indexing" principle and uses a layered structure including two basic components - Classification Module and Indexing Module. This innovation enables superior search efficiency and effectiveness. To achieve good music classification accuracy, multiple acoustic feature based music class profiling model is proposed to characterize different music categories. Together with Logistic regression based likelihood value estimation and combination scheme, it can enhance categorization effectiveness and thus improve the overall retrieval accuracy greatly. Meanwhile, a novel deep learning based music signature generation scheme is proposed to compute compact and comprehensive music descriptor. It can effectively fuse various kinds of acoustic features for the purpose of indexing and search by leveraging a wide range of mature multidimensional indexing techniques.

2. SYSTEM ARCHITECTURE OF EMIF

EMIF has two main components: classification module and indexing module. When it receives input music, classification module will identify music category. Then top k query processing is carried out using the corresponding local indexing tree in the indexing module. In following, we give a detailed introduction on system architecture of EMIF.

2.1 Classification Module

2.1.1 Multifeature based Statistical Class Modelling

To characterize each high level semantic category (or class), we develop a Music Category Modelling Module (MCMM) to construct a multiple feature based statistical scheme. Each MCMM in EMIF corresponds to one music category in the database. Each MCMM is made up of two parts: 1) feature extraction, and 2) a set of LDMMs (Linear Discriminative Mixture Model) built for statistical modelling of music class based on various kinds of acoustic features. A LDMM is a stochastic model combining the advantages of LDA and Gaussian Mixture Model (GMM). The novelty

[1]**EMIF** stands for **E**ffective **M**usic **I**ndexing **F**ramework

of LDMM is its great capability and flexibility for effective feature modelling. In MCMM, each LDMM corresponds to one acoustic feature type.

Feature extraction aims to calculate a numerical summarisation of music documents. EMIF applies the partition-based approach to extract multiple local features. When receiving input music, EMIF firstly segments it into small blocks and then different kinds of acoustic features are extracted from each block as basic content representation. Specifically, the features considered in this study include timbre, rhythm and pitch. The extraction process can be denoted as $\boldsymbol{V}_f = Extract_f(mo) = [\boldsymbol{v}_{1f}, \boldsymbol{v}_{2f}, ..., \boldsymbol{v}_{Bf}]$, where $\boldsymbol{V}_f$ is the set of vectors for a feature f extracted from the B blocks of the input music object mo. For our system, GMM is used as a statistical processor to model feature distributions for the particular semantic concepts. Based on each kind of feature, a GMM based category model can be trained separately for the task of class identification. In this study, we consider three different acoustic features including timbre feature, rhythm feature [14] and pitch feature [13].

In order to gain effective music category identification, EMIF constructs a statistical model for each class using multiple features. To achieve this, the individual feature of the music is extracted, and then, individual profiling model for one class is constructed using each feature. In EMIF, category profiling captures statistical properties of different features using Linear Discriminative Mixture Model (LDMM), which is a novel classification scheme combining the advantages of both LDA and GMMs. The main advantage of LDA over other linear subspace methods is to generate a discriminative feature space to maximize the ratio of between-class scatter against within-class scatter (Fisher's criterion) [2, 5]. In the LDMM for each acoustic feature, LDA is used as feature extraction that provides a linear transformation of raw features (n dimension) to m dimensional subspace (m dimension, $m < n$). Consequently, the samples belonging to the same category are closer and the samples from different categories are far apart. At the same time, since LDA can significantly reduce the dimensionality of raw feature, LDMM's training and classification will be much faster. With GMM, the probability of class c can be modeled as a random variable drawn from a probability distribution for a particular feature f after LDA transformation. Given a parameter set Θ_f^s based on feature f, the probability distribution is present as a mixture of multivariate component densities: $P_f^c(\boldsymbol{V}_f|\Theta_f^s) = \prod_{b=1}^{B}\{\sum_{j=1}^{J} w_{fj}^c p_f^c(\boldsymbol{v}_{bf}|\, \boldsymbol{\mu}_{fj}^c, \boldsymbol{\Sigma}_{fj}^c)\}$. The Gaussian density is used as the multivariate component in this study, according to GMM $\Theta_f^s = \{w_{fj}^c, \boldsymbol{\mu}_{fj}^c, \boldsymbol{\Sigma}_{fj}^c|\ where\ 1 < j < J\}$, where w_{fj}^c, $\boldsymbol{\mu}_{fj}^c$ and $\boldsymbol{\Sigma}_{fj}^c$ denote, respectively, mixture weights, mean vectors and covariance matrices. Also, $p_f^c(\boldsymbol{v}_{bf}|\, \boldsymbol{\mu}_{fj}^c, \boldsymbol{\Sigma}_{fj}^c)$ is the probability of a class label c based on feature f extracted from segment b. Given feature vector $\boldsymbol{v}_{bf}$, it can be easily calculated using the Gaussian density function and associated parameters $\{\boldsymbol{\mu}_{fj}^c, \boldsymbol{\Sigma}_{fj}^c\}$.

To effectively estimate model parameters, EMIF applies EM algorithm [1]. The EM is an iterative method to estimate and optimize some unknown parameters based on given data set. Since EMIF considers multiple features, the overall training procedure will be repeated multiple times, once for each feature. After the training process of system is completed, the likelihood value generated based on feature f for input feature vector $\boldsymbol{V}_f$ can be given as below,

$$\begin{aligned} l_f^c &= log(P_f^c(\boldsymbol{V}_f|\Theta_f^s)) \\ &= \sum_{b=1}^{B} log(\{\sum_{j=1}^{J} w_{fj}^c p_f^c(\boldsymbol{v}_{bf}|\, \boldsymbol{\mu}_{fj}^c, \boldsymbol{\Sigma}_{fj}^c)\}) \end{aligned} \quad (1)$$

An overall likelihood value can be derived based on various features for category c, denoted as $L^c = C^c(\vec{l^c}, \vec{W^c})$, where $\vec{l^c} = \{l_1^c, l_2^c, ..., l_F^c\}$ and $\vec{W^c} = \{W_1^c, W_2^c, ..., W_F^c\}$ include all the combination weights and scores from the category profiling model for class c. C^c is likelihood value combination function. When classifying input music, L^c can be used to measure the universal similarity distance between and input music and a class label c. Thus, how to compute combination weights is very important and in fact, the simplest way to determine combination weights would be to assign all combination weights same value regardless of their scores. However, the key disadvantage of the approach is the inability to take into account different features' effects on music classification.

2.1.2 Fusion Weight Estimation

How to fuse likelihood value generated based on different features is important to music classification performance in classification module of EMIF. To gain a comprehensive statistical model for each music category, EMIF applies the Logistic function as a combination scheme C^c to estimate an overall likelihood score between input song and class c [6, 7]. Basic idea of the approach is same to the one applied in HSI singer identification system [10]. By using Logistic function, overall likelihood value L^c can be scaled to [0,1] and formulated as below,

$$L^c = C^c(\vec{l^c}, \vec{W^c}) = \frac{1}{1 + exp(-y_c \sum_{f=1}^{F} W_f^c\, l_f^c)} \quad (2)$$

where $y_c = 1$ when c is input music's category, $y_c = -1$ otherwise, W_f^c denotes the fusion weight to compute the overall likelihood value generated based on feature f for class c. By using Equation 3, we can compute the likelihood value for the learning samples

$$\prod_{m=1}^{M} \frac{1}{1 + exp(-y_c \sum_{f=1}^{F} W_f^c\, l_f^c)}, \quad (3)$$

where M is the total number of learning examples. To train the model, we apply the learning algorithm introduced in [10], whose goal is to gain the overall likelihood value maximization.

2.2 Indexing Module

Indexing module contains multiple local indexing structures, one per music class. Each local indexing structures has two major components: music signature generation scheme based deep learning and multidimensional indexing scheme.

2.2.1 Deep Music Signature Generation

In EMIF, a deep learning based music signature generation scheme (DMSG) is developed to combine various low level acoustic features extracted from different segments into Deep Music Signature (DMS) - a set of linear vectors. Its physical representation can be $DMS = \{dms_1, dms_2, ..., dms_B\}$, where B is number of blocks in the music. Then, linear similarity functions (e.g., Euclidean distance) can be applied to calculate the similarity between two music documents based on their DMSs. DMSC is deep neural network architecture based on Stacked Denoising Autoencoder (SDA) and principal components analysis (PCA).

- PCA is used to preprocess raw input features from different blocks via linear transformation and speed up learning of SDA.
- SDA is adopted to pretrain neural networks for each block with unlabelled data.

- For each block of input music documents, the parameters of SDA are optimized via stochastic gradient descent.

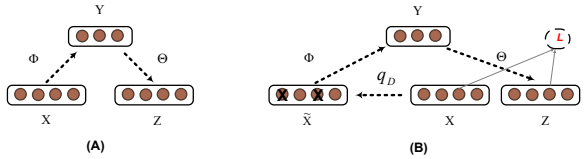

Figure 1: (A) Autoencoder (B) Denoising Autoencoder

Both Denoising Autoencoder (DAE) and Stacked Denoising Autoencoder (SDA) are developed based on Autoencoder (AE). They consists of two key components - encoder and decoder. Encoder transforms an input X into hidden representation y and decoder maps it back to a reconstructed d dimensional vector z. Figure 1 (A) illustrates basic idea of AE. In EMIF, the hidden layer is encoded by a nonlinear one-layer neural network and the mapping can be $Y = \Phi(X)$. The reconstruction from hidden representation Y can be computed using $Z = \Theta(Y)$. Various kinds of distributional assumptions can be applied on the input given the code. Also, different loss functions can be used to measure reconstruction errors on the output side. In EMIF, since we assume the distribution $dist(X|Z)$ is Gaussian, squared error loss can be $L_{sqe}(X, Z) = ||X - Z||^2$. In real world, the reconstruction criterion alone may not be always able to guarantee the generation of effective raw data representation. It might easily lead to the undesirable result - "simply copy the input". Thus, DAE is proposed to avoid this phenomenon by taking different strategy - training neural network locally to denoise noisy versions of initial inputs. The part (B) of Figure 1 visualizes the basic idea of DAE. It is done by firstly constructing X's corrupted version $\widetilde{X}$ via a stochastic mapping $\widetilde{X} - q_D(\widetilde{X}|X)$. q_D is a function to corrupt X and the corrupted input $\widetilde{X}$ is then mapped to a hidden representation $Y = \Phi_1(X)$, where Y is then used to reconstruct the initial version of X by $Z = \Theta(Y)$. The reconstruction error $L(X, Z)$ in stead of $L(\widetilde{X}, Z)$ is minimized in DAE. During the training, each round one training example X is given, a different version of corrupted X is generated based on function q_D.

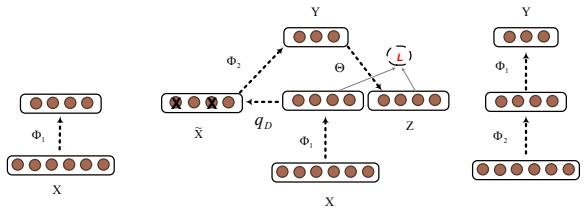

Figure 2: Stacking Denoising Autoencoder

In EMIF, SDA is applied to build deep learning architecture for computing DMS as basic component, one for each music block. We initialize the deep neural network using the same strategy which stacking RBMs in deep belief networks apply. Figure 2 illustrates the procedure to gain multilayer DAE. Firstly, the corrupted input is only used for training each layer at very beginning. This is very important to learn effective features. Right after the mapping function Φ has been learnt successfully, it can be applied to process uncorrupted inputs. Then, to train the neurons in the next layer, corrupted training examples will be used as inputs. After a set of encoders are trained and stacked as SDAs, outputs from top layer will serve as music content representation - DMS and inputs to existing multidimensional indexing structure (e.g., R-Tree or Hybrid Tree) for effective and efficient music search.

2.3 Music Query Processing

EMIF system construction includes two key steps. At the initial step, various kinds of acoustic features are extracted from input music. Then the parameters of the multiple feature based statistical modelling module for each category can be estimated using EM procedure. After classification module is trained, we will construct C indexing structures based on DMSG in conjuncture with indexing method, one per music category. In this study, we apply Hybrid tree due to its efficiency[4].

After system training and configuration, EMIF firstly identifies the category of input music using classification module. Once that class label has been determined, the top k query process can be carried out based on the corresponding local indexing structure.

3. EMPIRICAL STUDY

This section presents an experimental study to evaluate the proposed method and its competitive schemes. The dataset (Dataset I) used for the empirical study contains 5000 music items covering ten genres with 500 songs each genre. This dataset is very similar to the test collection used in [8, 14]. The test query we consider is to find music that has similar genre from database constructed using Dataset I. We compare EMIF with three different approaches including CM*F+Hybrid Tree (CM*F+HT), DWCH+Hybrid tree (DWCH+HT) and MARSYAS+Hybrid tree (MARS+HT). Our results demonstrate the superiority of EMIF over different approaches, including improvements on retrieval accuracy and efficiency in terms of the query response time.

3.1 Effectiveness Comparison

In the first experiment, we report a comparative study on the retrieval effectiveness of the CM*F+HT, DWCH+HT, MARS+HT, and EMIF. Tables 1 illustrates the query precision in terms of two different measurements: P@10 and MAP. Specifically, in each test, we randomly select query examples from the database and there is no overlap between query sets and training sets. It can be clearly seen that EMIF achieves significant improvement on query accuracies for all cases. In particular, the EMIF method improves the query effectiveness, on average, 10.2% for P@10 and by 12.2% for MAP. These results indicate that EMIF, whose structure integrates classification scheme and multiple indexing structures into single framework, is more effective than other approaches for CBMR task. This superior effectiveness is due to the multiple layer structure combining classification model and a Logistic regression based likelihood score fusion scheme. Furthermore, the indexing structure supporting query process on the music from individual category lead to a much smaller searching space and faster retrieval.

Query Methods	Query Accuracy	
	P@10	MAP
EMIF	0.617	0.511
CM*F+HT	0.438	0.385
DWCH+HT	0.372	0.302
MARS+HT	0.297	0.275

Table 1: Query accuracy comparison of EMIF and other approaches for music retrieval

Size of Result Set	Query Response Time(Sec)			
	EMIF	CM*F+HT	DWCH+HT	MARS+HT
5	**0.065**	0.073	0.086	0.095
10	**0.089**	0.109	0.172	0.209
15	**0.097**	0.205	0.259	0.287
20	**0.132**	0.258	0.332	0.385

Table 2: Query response time comparison of EMIF and other approaches

3.2 Efficiency Comparison

Query response time is another important aspect for system performance evaluation, especially when scale of the music dataset becomes large. Comparing to traditional indexing structure, although the statistical concept model in EMIF lifts the accuracy significantly, they might introduce query cost overhead. Thus, the second set of experiments evaluate and compare query efficiency of the EMIF and other competitors. The test was run with 1000 query examples randomly selected from the music dataset. Table 2 shows the query response time of different methods with various size of the result sets. From the experimental results summarised in the table, we can see that EMIF achieves great saving in terms of query speed against the other approaches for all sizes of result sets. MARS+HT performs worst among six different approaches tests. EMIF achieves the fastest response time and compared to other approaches, performance gain is very significant. We believe that the main reason behind efficiency enhancement is EMIF's layered structure facilitating retrieval processing based on index structure built for single music class. It suggests a more compact searching space (smaller indexing structure). Consequently we can observe significant reduction in query response time.

4. CONCLUSION

In this research, we propose and evaluate an intelligent indexing framework called EMIF based on the "classify-and-indexing" design principle. To achieve accurate music classification, an independent LDMM-based profiling model for individual music category is constructed using multiple features to generate likelihood score. Moreover, EMIF's layered architecture enables more compact indexing structure for each music category and consequently achieve a significant reduction on query execution time. To validate our approach, a series of comprehensive experimental studies have been carried out over large scale test collection and the results reveal the various advantages of EMIF over the existing state-of-the-art indexing methods.

Acknowledgement

Jialie Shen is supported by Microsoft Research Asia Grant (FY14-RES-OPP-048) - "My Mobile Music: Towards Cloud based Intelligent Music Recommendation on the Move".

5. REFERENCES

[1] N. L. A. Dempster and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1977.

[2] C. M. Bishop. *Neural Network for Pattern Recognition*. Oxford University Press, 2000.

[3] C. Böhm, S. Berchtold, and D. A. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3), 2001.

[4] K. Chakrabarti and S. Mehrotra. The hybrid tree: An index structure for high dimensional feature spaces. In *Proc. of ICDE Conference (ICDE'99)*, 1999.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, 2000.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, 2001.

[7] M. I. Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. Technical Report 9503, Massachusetts Institute of Technology, 1995.

[8] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. of ACM SIGIR Conference (SIGIR'03)*, 2003.

[9] F. Pachet. Content management for electronic music distribution. *Communications of the ACM*, 46(4), 2003.

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

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

[12] J. Shen, D. Tao, and X. Li. Quc-tree: Integrating query context information for efficient music retrieval. *IEEE Transactions on Multimedia*, 11(2):313–323, 2009.

[13] T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8(6):708–716, November 2000.

[14] G. Tzanetakis and P. R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 2002.

[15] G. Xia, T. Huang, Y. Ma, R. Dannenberg, and C. Faloutsos. Midifind: Similarity search and popularity mining in large midi databases. In M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad, editors, *Sound, Music, and Motion*, volume 8905 of *Lecture Notes in Computer Science*, pages 259–276. Springer International Publishing, 2014.

[16] C. Yang. Efficient acoustic index for music retrieval with various degrees of similarity. In *ACM MM*, 2002.

[17] Y. Yu, M. Crucianu, V. Oria, and E. Damiani. Combining multi-probe histogram and order-statistics based lsh for scalable audio content retrieval. In *ACM MM*, 2010.