MATEC Web of Conferences 22, 01024 (2015)
DOI: 10.1051/matecconf/20152201024

Design and Realization of Music Retrieval System Based on Feature Content

Lei Li
College of Construction Engineering, Weifang University of Science and Technology, Shouguang, Shandong, China

Jing Shi
School of Chemical Engineering & Environment, Weifang University of Science and Technology, Shouguang, Shandong, China

ABSTRACT: As computer technology develops rapidly, retrieval systems have also undergone great changes. People are no longer contented with singular retrieval means, but are trying many other ways to retrieve feature content. When it comes to music, however, the complexity of sound is still preventing its retrieval from moving further forward. To solve this problem, systematic analysis and study is carried out on music retrieval system based on feature content. A music retrieval system model based on feature content consisting of technical approaches for processing and retrieving of extraction symbols of music feature content is built and realized. An SML model is proposed and tested on two different types of song sets. The result shows good performance of the system. Besides, the shortfalls of the model are also noted and the future prospects of the music retrieval system based on feature content are outlined.

Keywords: feature content; SML model; music retrieval system

1 INTRODUCTION

The retrieval of feature content is somewhat different from that after manual interference as it follows the objective law. The retrieval content typically includes audios, images, videos and other information that shows digital content. Audios take up more than half of the content of music retrieval. The unique rhythm and audio frequency of music makes it very helpful to develop music retrieval system based on feature content.

Previous authors have made efforts and appreciable achievements in the research of music retrieval systems. Guo H.P. noted the future development direction in his studies on content-based music retrieval system in the light of successful cases of the existing music retrieval algorithms. He also designed a test system that queries on names of songs and proved satisfactory result after verifying the rationality of this system by conducting a listening test [1]. Yang B. refined the end-point retrieval and target music segmentation for hum input after examining the key techniques of polyphonic music retrieval, improved the pitch extraction model of fundamental frequency matrix, and successfully completed the design of music retrieval functions and tested the system [2]. Meng X.W. studied the expression of music content and the processing of signals in his research on music information retrieval, discussed the main technical methods, built a music retrieval model and pointed out its defects [3]. On the basis of previous findings, this paper looks further into the music retrieval systems based on feature content and builds a retrieval model in hopes of providing theoretical reference for future studies on music retrieval systems.

2 DESIGN OF MUSIC RETRIEVAL SYSTEM BASED ON FEATURE CONTENT

2.1 *Music retrieval system model*

In the image retrieval system related to semantic text, Carneiro et al introduced an SML (supervised multiclass labeling) model that manually builds picture sets and extracts features that are spatially subject to polynomial distribution. This kind of retrieval is a multivariate distribution that obtains automatic labeling of retrieval by calculating model parameters. The final result indicates extensive application potential of this model for retrieval purposes.

After Carneiro, Turnbull et al applied the SML retrieval model to music retrieval and also confirmed the high universality of this model.

Figure 1 shows the music framework of the music retrieval system SML based on feature content.

In the manually labeled original music data, after processing the data, we can obtain the semantic feature and the music feature models, and build a model according to the spatial distribution of these two features. The most important of all is the GMM

Article available at http://www.matec-conferences.org or http://dx.doi.org/10.1051/matecconf/20152201024

model that refines the correlation between the two features by adjusting the parameters, thereby realizing a music retrieval system based on feature content.

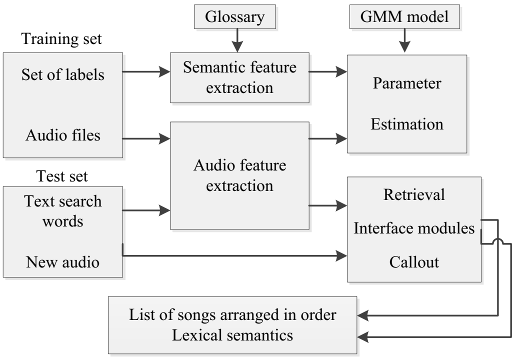

Figure 1. Schematic model

The SML algorithm is in the core of the entire model system. It classifies each individual word as a class and labels to allow automatic retrieval. The exact process is described below.

Suppose each of the words in the glossary V is labeled w_i , then we have $w_i \in V$. For a known song s_q , we look for a group of related words:

$$W = \{w_1, w_2, L \ w_A\}, \tag{1}$$

The feedback song priority in the system is

$$S = \{s_1, s_2 L \ s_R\} \tag{2}$$

According to the labeled data above, we obtain the vector of each song:

$$y = \{y_1, y_2 L \ y_{|V|}\} \tag{3}$$

If a song is labeled $w_i = 0$ and y_i is used to represent the correlation between the song and the word, what we call semantic weight, then $y_i = 0$. If $w_i > 0$, then we have $y_i > 0$. If we map y_i to $\{0,1\}$, then it can be regarded as the class label. The set of feature content of a song's name can be expressed as:

$$X = \{x_1, x_2, L \ x_T\}, \tag{4}$$

Where, T decides the length of the song and represents the total number of frames of the song and x_t is the feature vector extracted from the signals of frame t audio. The set of the label tree of the song is expressed as:

$$D = \{(x_1, y_1), (x_2, y_2) L \ (x_D, y_D)\} \tag{5}$$

2.2 *Realization principle of music retrieval system*

In automatic labeling, each word can be regarded as a class. The word having the largest correlation coefficient is used to describe the labeled song. Under spatial condition w_i , the word set distribution of the music features is $P(x|i)$, $i \in \{1,2,L \ |V|\}$. The feature vector of a song is then $X = \{x_1, x_2, L \ x_T\}$, According to Bayes, we obtain the posterior probability of the audio feature space of each word:

$$P(x|i) = \frac{P(x|i) \ P(i)}{P(x)} \tag{6}$$

Where, the prior probability of the word is expressed as $P(i)$. If the feature vectors are independent from each other under spatial condition W_i , then we have:

$$P(x|i) = \frac{\left[\Pi_{t=1}^{T} P(x|i)\right] P(i)}{P(x)} \tag{7}$$

Under spatial condition w_i , the features vectors of the audios of the word are time independent. However, this assumption does not conform to the fact. To overcome this problem, the $\Pi_{t=1}^{T} P(x|i)$ in the formula above is averaged to $\left[\Pi_{t=1}^{T} P(x|i)\right]^{\frac{1}{T}}$. Assume that the prior probability of the word is uniform, and then the prior probability of the song is expressed as $\Sigma_{v=1}^{|V|} P(x|v) p(v)$. By bringing this into the formula above, we obtain the expression for the automatic labeling:

$$P(i|x) = \frac{\left[\Pi_{t=1}^{T} P(x|i)\right]^{\frac{1}{T}}}{\Sigma_{v=1}^{|V|} \left[\Pi_{t=1}^{T} P(x|i)\right]^{\frac{1}{T}}} \tag{8}$$

According to Bayes and the word distribution, $P(x|i)$, $i \in \{1,2,L \ |V|\}$. Using the formula above, we obtain the distribution function of the spatial feature vector of the audio, and the songs in the music set can be expressed as the posterior probability $P = \{p_1, p_2 \ L \ p_{|V|}\}$, where, $P_i = P(i|x)$ and $\sum_i P_i = 1$.

Systematic retrieval is realized according to the semantic polynomial distribution of the music piece. If we enter text query, we can establish a corresponding polynomial distribution form of query $q = \{q_1, q_2 L \ q_{|V|}\}$. Using uniform distribution method, we define the probability of each word as $1/N$. For the word that no longer exists in the dictionary, we can define its probability as a smaller value ε , such as $\varepsilon = 10^{-6}$.

To calculate the similarity of a high-dimensional space, commonly accepted methods include angle cosine:

$$D_{E^2}(p,q) = \sum_{i=1}^{N} \left[p(i) - q(i)\right]^2 \tag{9}$$

and the Euclidean distance formula:

$$D_{E^2}(\text{p,q})=\frac{\sum_{i=1}^{N}\left[p(\text{i})\text{q(i)}\right]}{\sqrt{\sum_{i=1}^{N}p(\text{i})^2\times\sum_{\text{i}=1}^{N}\text{q(i)}^2}} \quad (10)$$

Entropy of KL distance method examines problems on the basis of information theory, while the distance calculation methods of the all formula above looks at problems from the geometric perspective and rely on semantic feature space and assumptions of audio feature polynomial. KL method is more suitable.

KL refers to the distance between semantic polynomial p and query polynomial q, expressed as:

$$KL(\text{q}\|\text{p})=\sum_{\text{i}=1}^{|V|}q_i\log\frac{q_i}{p_i} \quad (11)$$

The actual distribution is the polynomial distribution discussed above. When the inputted word does not appear in the list of words, then $\text{q}_i=0$. The test result will therefore return to the song corresponding to the largest w_i in the music database. As this simply means a piece of semantic information is added when a new song is inputted and the final retrieval is hardly affected, this prepares the system for good updates.

2.3 *GMM model*

The SLM model in the system is a multivariate Gaussian mixed model (GMM), whic h describes the spatial distribution of probabilities by combining a number of probability densities with Gaussian distribution. The d-dimensional GMM of R mixed numbers is expressed as:

$$g(\text{x})=\sum_{r=1}^{R}\pi_r N(x/\mu_r,\textstyle\sum_r) \quad (12)$$

Here: $\sum_r$ is called the covariance matrix; μ_r is the mean value; $N(x/\mu_r,\sum_r)$ is the Gaussian function; π_r is called mixed weight with $\sum_r\pi_\text{r}=1$; X is the observation vector.

When designing the retrieval system, we calculate the probability of the audio feature in the semantic space to be $P(\text{x}|\text{i})$, and then the corresponding expression is:

$$P(\text{x}|\text{i})=\sum_{r=1}^{R}\pi_r N(x/\mu_r,\textstyle\sum_r) \quad (13)$$

Here: the covariance matrix is the GMM distribution of $\sum$ and the average is μ , expressed as $N(x/\mu,\sum)$. As covariance alone will result in inadequate number of iterations, and the full use of covariance will cause overflow, we used diagonal covariance to avoid this problem.

2.3.1 *Parameter design of GMM model*

In the presence of implicit variables, the most frequently used algorithm is the EM algorithm, which is an expectation maximization method that calculates the most essential parameters. Normally, the EM algorithm causes $\text{E}[\ln\text{p}(\text{Y}|h')]$ to have the largest h' by searching and calculates this expectation under the law followed by Y. $\text{p}(\text{Y}|h')$ is the likelihood of all data under the assumption.

The distribution of the probability functions followed by Y, as determined by the to-be-determined estimation parameter θ is unknown, but the EM algorithm used an assumed h to replace θ and thereby estimates the distribution of Y. When $\theta=h$, and under all the assumptions of x for the observed part of the data Y, we have:

$$Q(\text{h}'|h\ )=\text{E}[\ln\text{p}(\text{Y}|h'),\text{x}] \quad (14)$$

However, this algorithm mainly includes two repeated steps:

Step 1: Estimate the probability distribution of Y.

$$Q(\text{h}'|h\ )\leftarrow\text{E}[\ln\text{p}(\text{Y}|h')\,\text{h}\,,\text{x}] \quad (15)$$

Step 2: Assume that maximize function Q .

$$\text{h}\leftarrow\arg\max Q(\text{h}'|h) \quad (16)$$

If Q is continuous, then $\text{p}(\text{Y}|h')$ is a fixed point. If the likelihood function has the maximum value, then the EM algorithm can converge to an estimated global maximum likelihood. Otherwise it will converge to a local maximum.

2.4 *Application of EM algorithm in music retrieval system*

When estimating by the EM algorithm, the parameters of the GMM model can be estimated simply by direct training to build audio feature vector sets.

2.4.1 *Model weighting*

As the distribution of word set cannot be directly estimated, the music songs are distributed. w_i is extracted from the audio feature vectors of a song related to it and trained using the EM algorithm to calculate the GMM model and weight it to obtain the needed GMM model that reflects the distribution of the word set. The correlation between each word and the song is expressed as:

$$P_{X|Y}(\text{x}|\text{i})=\frac{1}{c}\sum_{\text{d}=1}^{|D|}[\text{y}_d]\text{i}\sum_{k=1}^{K}\pi_\text{r}^{(\text{d})}N(x/\mu_\text{r}^{(\text{d})},\textstyle\sum_r^{(\text{d})}) \quad (17)$$

Where, $|D|$ is the total number of trainings; $c=\sum_{\text{d}=1}[\text{y}_d]_i$ is the total semantic weight; k is the number of trainings for each song, as shown in formula (2) above.

In real operation, when each new song is added, the many Gaussian distributions of each audio feature vector has to be recalculated, which is also the principal defect of this algorithm, to solve this problem, mixed hierarchical EM is introduced.

2.5 *Hybrid hierarchical EM algorithm*

By using mixed hierarchical EM algorithm, we can effectively estimate the distribution of the word set. The weighting portion of this algorithm, like the algorithm above, first calculates the distribution of the song set as shown in formula (3) above. Using the resulting $|D|\,k$ Gaussian distributions, we can estimate the word and the distribution GMM model parameters. Here we need two iteration methods *E-step* which measures the contribution $h^r_{(d),r}$ of the distribution element r of the word set to the song d , and M-*step* that calculates the distribution parameters of the word set once again. The formula of *E-step* is:

$$h^r_{(d),r}=\frac{[y_d]i[N(\mu^d_K|\mu_r,\Sigma_r)e^{-\frac{1}{2}Tr\{(\Sigma\ r)^{-1}\}}]^{\pi^{(d)}_k N}\pi_r}{\Sigma t[N(\mu^d_K|\mu_r,\Sigma_r)e^{-\frac{1}{2}Tr\{(\Sigma\ r)^{-1}\}}]^{\pi^{(d)}_k N}\pi_r} \quad (18)$$

Where, custom parameter is N , Order $N=k$, then the average of $\pi^{(d)}_{k}N$ is 1.

M-*step*

$$\mu^{new}_r=\sum_{(d)} z'_{(d),k}\mu^{(d)}_k \text{ , thereinto } z'_{(d),k}=\frac{h^r_{(d),k}\pi^{(d)}_k}{\Sigma_{(d),k}h^r_{(d),k}\pi^{(d)}_k} \quad (19)$$

$$\Sigma^{new}_r=\sum_{(d),k} z'_{(d),k}\mu^{(d)}_k\left[\Sigma^{(d)}_k+(\mu^{(d)}_k-\mu_t)(\mu^{(d)}_k-\mu_t)^T\right] \quad (20)$$

$$\pi^{new}_r=\frac{\Sigma_{(d),k}h^r_{(d),k}}{W,K} \text{ , thereinto } W=\sum_{d=1}^{|D|}\left[y_d\right]i \quad (21)$$

By sampling the mixed components of the distribution of the word set, we obtain the distribution of a song set. The distribution samples of the song set are the audio features. Throughout the process, all the data and the number of parameters of the algorithm used must be exactly the same as those used in the training. By this way, the mixed hierarchical algorithm takes the place of the standard algorithm.

It is noted that the selection of a new π_r,μ_r,Σ_r in the mixed hierarchical algorithm is determined according to the initial parameters of GMM. The closer the selected initial parameters are to the center of mass, the more satisfactory the result. Here K-means clustering algorithm is used to process the data to derive the initial parameters.

3 REALIZATION OF MUSIC RETRIEVAL SYSTEM BASED ON FEATURE CONTENT

The purpose of must retrieval based on feature content is to design a large music database that allows quick search and store of music with automatic labeling and retrieval functions.

3.1 *Framework structure*

Figure 1 shows the SML model proposed. This includes model training, feature extraction and preprocessing to enable automatic labeling and search.

The collection of songs in this system mainly includes 500 highly representative Chinese songs personal music dataset PMD500 (Personal Music dataset-song) popular over the last few decades and 500 different western songs of University of California over five decades after its foundation (Computer Audition Lab 500-song).

The feature content of music is closely associated with the semantic glossary. PMD500 selects six types of vocabularies including feelings, musical instruments and popular factions and adjusted the features of Chinese music on the basis of CAL500.

3.2 *System realization and testing*

The programming interface was designed under vc.net using the main program of MatLab. MFCC song coefficients were extracted using Audio Processing Toolkit. Two music pieces of PMD500 and CAL500 were trained to realize automatic labeling and search.

The advantages of the MIR system were evaluated using systematic test to improve the system effectiveness. Particularly, precision and recall were tested. The former is tested according to the formula below:

$$\text{Precision} = \frac{\text{Number of songs retrieved}}{\text{Total number of songs returned from retrieval}} \quad (22)$$

The latter is tested according to the formula below:

$$\text{Recall} = \frac{\text{Number of songs retrieved}}{\text{Total number of songs in the song set}} \quad (23)$$

Songs collected were tested using these formulae. The result is presented in the table below:

Table 1. Testing of music retrieval system

Retrieve content	*PMD*500	*CAL*500
Re*call*@10	0.2899	0.2582
Pr*ecision*@10	0.7842	0.6832
Re*call*@5	0.2015	0.1785
Pr*ecision*@5	0.8251	0.8321

From this table, we can see the retrieval performance in terms of precision and recall. The precision is higher than 80%, but the recall is not high.

The average recall and precision of the two types of song sets were calculated to measure the labeling performance of the system by labeling the songs of the test sets with ten label words and five label words. The result is given in Table 2 below.

Table 2. Testing of automatic music labeling system

Automatic annotation	*PMD*500	*CAL*500

Re*call*-10 *word*	0.1854	0.1789
Re*ecision*-10 *word*	0.2989	0.2733
Re*call*-5 *word*	0.1236	0.0964
Re*ecision*-5 *word*	0.4643	0.4768

A comparison of the results from the use of different datasets reveals very close results, which confirms that our SML model is quite advantageous among music retrieval systems based on feature content.

4 CONCLUSIONS

An SML model involving a smaller amount of data was used to build a music database with the help of manual label. Despite some unconformity between the glossary and the datasets, the use of the SML model in the music retrieval system succeeded in achieving semantic polynomial and provides good foundation for the realization of labeling and retrieval functions. However, we have also noticed some of the defects of this model and will try to improve it in subsequent studies.

REFERENCES

[1] Guo, H.P. 2007. *The Research of Algorithms and system Accomplishment Based Content Musical Retrieval*. Changchun: Northeast Normal University. pp: 5-13.

[2] Yang, B. 2012. *Design of a Content-based Polyphonic Music Retrieval System*. Wuhan: Wuhan University of Science & Technology. pp: 4-11.

[3] Meng, X.W. 2009. *Research and System Implementation on Content Based Music Information Retrieval*. Beijing: Beijing University of Posts and Communications. pp: 6-17.

[4] N. Kosugietal. 1999. Music retrieval by humming, *IEEE PACRIM*.

[5] G. Tzanetakis. & P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*. pp: 292-300.

[6] N. Kosugietal. 2000. A practical query-by-humming system for a large music database, *ACM Multimedia*.

[7] Zhang, X. 2013. The influence of MPEG-7 standard on content-based retrieval. *Journal of Information*. (12)

[8] Sun, G.C. 2008. *Study on Content-based Audio Retrieval Technology*. Wuhan: Huazhong University of Science & Technology.

[9] Wu, C.J. 2011. Design and Realization of Melody-based Music Retrieval System. Beijing: Beijing University of Posts and Communications.

[10] Ginanjar, Rikip. 2011. MIDI Conversion to musical notation. *Proceedings-1st International Conference on Informatics and Computational Intelligence*, *ICI 2011*, pp: 95-98.