Perceptual Visualization of A Music Collection

Jiajun Zhu*
Dept. Comp. Sci. & Eng., Fudan University
zhujiajun@fudan.edu.cn

Lie Lu
Microsoft Research Asia
llu@microsoft.com

Abstract

Music visualization provides users with a new interface to browse, search, and navigate their personal digital music collection. Although there are several previous works on visualizing a music collection based on some "surface" musical metadata, such as artist, album and genre, there has been few works on content-based or perception-based visualizations. In this paper, we developed an algorithm to automatically estimate human perceptions on rhythm and timbre of a music clip. Then, based on these two values, each music clip is mapped into a 2D (timbre-rhythm) space. Thus a 2D perception-based visualization is built. Experimental evaluation indicates that this kind visualization is efficiently helpful in many cases of music management manipulations, such as music navigation, similar music search and music play list generation.

1. Introduction

The rapid developments of the Internet, portable devices, multimedia compression and storage technologies have largely increased the amount of personal digital music data. Now it is common that users store hundreds of CDs and thousands of tracks in their computers, which makes better management of personal music collection a more interesting and challenging problem. Most of current music management tools use "conventional" folder-based organization and textual list interface to represent the music collection, such as Windows Media Player, Real Player, Winamp, and Apple iTune. Although users have been familiar with these conventional interfaces, one of the critical disadvantages of the folder-based music organization is that they can just list tracks in the alphabet order of one user-specified music attributes, such as title, album, artist and genre, if these information is available.

Therefore, in order to give a better music organization and facilitate users to manage and navigate their music collection, some novel visualizations of music collection have been proposed. For example, disc visualization, rectangle visualization, and tree-map visualization are presented to allow users to better organize personal music collections and create play lists [1]. Tracks in these visualizations are organized in a 2D space according to their metadata such as genre, artist, year, and album. Compared with conventional folder-based interface, these visualizations are better organized and easier to use.

However, the above textual information (artist, genre) only describes "surface" concepts and sometimes is not enough for a user to manage and navigate the music collection, especially when the user is exploring newly released or unfamiliar music. For example, in many cases, it is more likely that the users want to select tracks according to some perceptual attributes of the music (e.g. joyous or gloomy, fast or slow), instead of music title or artist. Moreover, traditional textual metadata also could not provide enough information on music similarity, which the users highly expect in their music experiences [2].

Some previous works have been done to generate 2D displays of image collections based on perceptual similarities [3][4], however, there are few similar works on perception-based music visualization. In order to provide the users with the possibility of exploring music collection based on perceptual attributes of music, in this paper, we propose a new approach to visualizing a music collection based on two perceptual attributes, rhythm and timbre. We also develop an algorithm to automatically estimate rhythm and timbre for an acoustic music track. The reasons we select rhythm and timbre as the dimensions in our 2D visualization are as following. Firstly, they are able to reflect the music characteristics in temporal and spectral domain respectively. Secondly, they are important factors of musical emotion and thus can facilitate music selection in some contexts. For example, joyous music usually has fast rhythm and bright timbre; thus if a user wants to listen to joyous music, he can find them in the corresponding region in the 2D visualization. Thirdly, humans basically have consistent perception on timbre and rhythm, as our user study (Section 2) shows. However, rhythm and timbre are too complex and contain many components. Therefore, in our implementation, we do some simplifications. The tempo is used to roughly represent rhythm, while brightness is used to represent timbre, since tempo and brightness are much easier for human to perceive. This visualization also implies the similarity between any two tracks. That is, the similarity between tracks can be indicated by the distance between their positions in the 2D space. Experimental evaluation also indicates that this novel visualization can save time for users to find their preferred music and is very helpful in exploring newly released music collection.

The rest of the paper is organized as follows: A user study on human perception consistency is given in Section 2. Section 3 describes our approaches to rhythm-timbre based visualization and the corresponding interface design. Section 4 presents the evaluations of the proposed visualization. Conclusions are given in Section 5.

2. A Pre User Study

At the beginning, it is necessary to learn whether and to what extent different people share similar perceptions in tempo (corresponding to rhythm) and brightness (corresponding to timbre). It is the basis that we can use these two attributes in the visualization, where a user can easily find a preferred track in a region he expects. Therefore, we perform a corresponding user study.

* This wok was performed when the first author was a visiting student in Microsoft Research Asia.

2.1 Design

In our user study, we prepare a collection of 600 tracks which vary in different genres including classical, pop, and rock. 20 subjects who have experiences on music participate in our user study and 5 of them are females. Each subject annotates 60 tracks and each track is annotated by 2 different subjects, and the overlap between any 2 subjects is 3 tracks.

We also prepare five reference music clips on musical tempo or brightness respectively. It facilitates user annotation since they provide some references to compare with. Each subject is asked to drag a slider to a corresponding position based on their perceptions and the reference clips, from which we can specify a perceptions value (from 0 to 1) for each music track, where 0 stands for the slowest tempo or the lowest brightness and 1 stands for the fastest tempo or the highest brightness. The whole annotation process varies from 30 to 45 minutes for each subject. The results are used to indicate human perception consistency and train the mapping model from high-dimension feature vector to 1-dimension tempo or brightness perception value (Section 3).

2.2 Results

The information we learn from the user study is encouraging. It indicates the rationality of designing a visualization based on tempo and brightness.

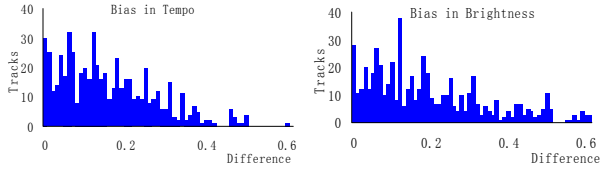

Figure 1. The histogram of the perceptual difference between individuals, where x-axis is the annotation difference ranging from 0.0 to 1.0, and the y-axis is the number of corresponding music tracks

Firstly, we evaluate the perceptual difference on musical tempo and brightness between individuals. Figure 1 illustrates the histogram of the annotation differences of 600 tracks on both tempo and brightness. From the figure, it can be seen that almost all the annotation difference in tempo between individuals are lower than 0.4, and the dominance of that in brightness lies from 0.0 to 0.3. These facts reflect that the perception bias on tempo and brightness between different subjects is pretty small.

However, some annotation differences may be introduced by different annotation preferences but not the perception difference. For example, the annotation given by an individual might be always 0.2 higher than that of another for each track. In order to compensate for this kind annotation preference, in our user study, we also look into whether individuals have more consistency in the annotation difference given two tracks. Figure 2 shows the corresponding comparison results. If two subjects have opposite perceptual direction comparing a track pair (e.g. one thinks track 1 is brighter than track 2, but another annotates oppositely), the bias (the difference between two individuals regarding the annotation difference between two tracks) is showed as a negative value; otherwise, a positive bias is obtained. From Figure 2, it can be seen that most of the values are positive and very close to zero. It indicates that, in most cases, the subjects have the same perception direction and small bias in the perception difference comparing two different tracks.

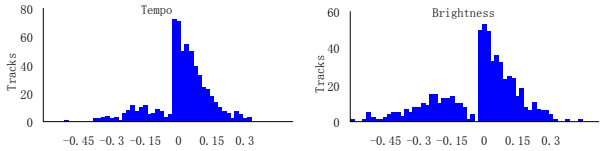

Figure 2. The histogram of perceptual differences in pair-comparison between individuals

Overall, the user study indicates that the perception biases between different individuals are not large. Due to this consistency, it is reasonable to build a 2D visualization based on these two perceptual attributes.

3. Building the Visualization

In this section, we extract a number of acoustic features to describe the timbre and rhythm of a music clip. A support vector regression (SVR) is then employed to predict human perceptions on musical tempo and brightness from these high dimensional features. Finally, the estimated attributes are utilized in generating the visualization interface.

3.1 Feature Extraction

In our approach, each track is first down-sampled into a uniform format: 16kHz, mono channel, and divided into non-overlapping 64ms-long frames. 20-dimension timbre features and 4-dimension rhythm features, most of which are addressed in previous works [6][7][8], are extracted to represent the tempo and brightness.

3.1.1 Timbre (Brightness) Features

Since the features are expected to represent musical brightness which is greatly related to frequencies, they are calculated directly from the spectrums which can be obtained by performing the FFT algorithm. It is noted that the silent frames are removed firstly, since in most cases individual ignores silence when determining the musical brightness. Previous work [5] shows that the timbre is determined primarily by the spectrum of different sub-bands. Therefore, we also divide the frequency domain into several sub-bands and the energy of each sub-band is normalized by the total frame energy. In our implementation, 19 sub-bands are experimentally selected, in order to obtain the optimal brightness estimation. Besides the normalized energies of the 19 sub-bands, we also extract the mean (centroid) of the FFT spectrum, since it is also highly correlated with brightness [9] (previous works called this feature “brightness” as [5]).

Suppose all the non-silent frames have the same importance in determination of the musical brightness, in our approach, the timbre features of the whole music track are simply computed by averaging each frame’s feature vector.

3.1.2 Rhythm (Tempo) Features

Tempo is usually measured by the frequency of beats (BPM, beats per minute). However, it is noted that some other rhythm factors may also affect the human perception on the speed of music (or perceived tempo), such as rhythm strength and rhythm regularity. From the user study, we also note that a track with regular and strong drums usually sounds faster than the one with the same BPM but without drums. To address these factors, in

our approach, the following four rhythm features are used to represent the perceptual music speed, based on the previous work [6], including average tempo (measured by BPM), average onset duration, average rhythm strength, and average rhythm regularity.

Both the first and second features are designed for the music speed information, considering the average frequency of beats and onsets. The third feature represents the average strength of the beats and onsets, while the fourth feature describes the period variance of the onsets. The more regular or stronger the rhythm is, the larger the corresponding feature is. More details on definition and implementation can be found in [6].

It is different from the timbre feature extraction that the rhythm features are calculated on each 10-second-clip of a track instead of each frame. Then the rhythm features of a track are represented by averaging all the corresponding 10-second-clips.

3.2 Mapping Model

As the previous user study shows, people have similar understanding on musical tempo and brightness, thus we can map each song to a 2D space based on its tempo and brightness value. Since tempo and brightness is one-dimension value (from 0.0 to 1.0) while the corresponding features are high dimensional, we should build a model to map each high-dimension feature vector into one-dimension tempo or brightness. In our approach, Support Vector Regression (SVR) [10], which is an application of Support Vector Machine, is utilized to develop such a mapping model.

In our approach, the training data is also obtained from the annotations in the previous user study, where the tracks with large perception difference between two different subjects are discarded to keep the training set representative. We randomly select 80% of the tracks as training set and the rest 20% as testing. SVR builds a satisfying regression model. Figure 3 illustrates the histogram of the difference between the prediction values and the annotations regarding the testing set. It shows that most of the prediction bias are lower than 0.4. Based on the previous user study, 0.2 is assumed as an acceptable deviation range. Thus, the proposed method is able to satisfyingly predict more than 84% perceptual brightness under this acceptable range, while the prediction satisfaction of tempo is up to 90%.

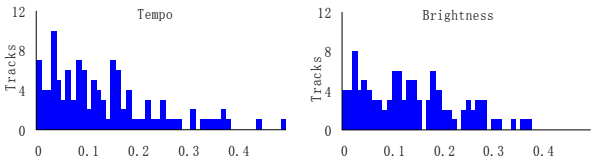

Figure 3. The histogram of the difference between the perception prediction and human annotation

3.3 Interface Design

Figure 4 illustrates a snapshot of our 2D visualization of a music collection, where each square represents a track whose position is determined by its tempo and brightness value. The size and color of each square can be used to represent other attributes of the corresponding track. In this interface, different size and color of the square indicate the popularity and musical genre, which are obtained from user history and metadata. Zoom in and zoom out are also implemented to provide users with a closer view of parts of the whole music collection.

This visualization design is very suitable for contextual aware play list generation. For example, when an individual comes back home from work, and wants to listen to some relaxing music, he can select the corresponding tracks in the region with high brightness and slow tempo in the 2D space. Another application on play list generation is that, given the first and the last track, the tracks most match the path can be selected to compose a play list, as the Figure 4 shows. Each track in such a play list changes gradually in musical rhythm and timbre.

It is noted that the 2D visualization can be easily extended into 3D mode if one more musical attribute is chosen as the third dimension. Thus more information can be included.

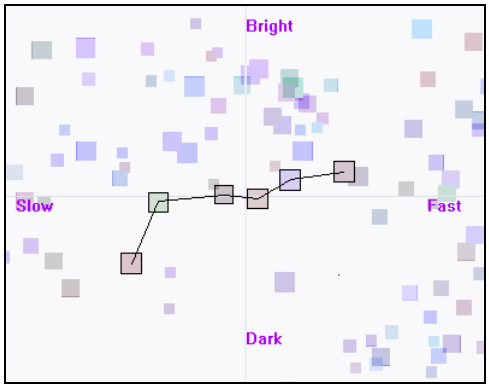

Figure 4. A snapshot of the 2D visualization of an example music collection

4. Evaluations

In this section, we conduct a user study and a post-study questionnaire to evaluate the proposed rhythm-timbre based 2D visualization interface, from which encouraging results are obtained.

4.1 Design and Setup

To evaluate the proposed visualization interface, we conduct another small scale user study to compare our visualization with the conventional interface. 5 females and 5 males with age ranging from 20 to 45 participate in this study. All of them are familiar with most of the existing music playback software, and each of them has a music collection of more than 200 tracks.

Since searching for preferred music and identifying a target track are two very common tasks in music management and experiences, the time cost for these two tasks can indicate the efficiency of the user interfaces. Therefore, we firstly perform a user study to compare the difference of navigating time and matching time cost between our visualization interface and the conventional textual list interface. Here, the navigating time is measured as the time cost for a user to select tracks from a music database according to their personal music taste, and the matching time is the time cost to identify the music which is played before the test. Since the previous visualizations [1] are also based on the same “surface” textual information as used in textual list interface, and do not provide any content information, we only compared our approach with conventional textual list interface in this experiment.

In the user study, we select a collection of 200 newly released tracks as the testing set, most of which are not familiar to the subjects. All the subjects are asked to navigate the database

using our visualization interface and the textual list interface respectively. In each navigation process, they are asked to select three preferred tracks; while in the matching process, a random track is firstly played, then the user need to identify it in both interfaces.

Besides the interface study, we also design a post-study questionnaire, which aims at getting more feedbacks from the subjects. In the experiment, the following questions are asked,

1) Can the brightness-tempo 2D space provide a supplement way to better describe your music collection?
2) Does the automatic perception estimation result conform to your own perception?
3) Does the visualization interface facilitate navigating newly released music collection?
4) Any suggestions to improve the visualization interface?

4.2 Results and Feedbacks

Figure 6(a) illustrates two box plots of the navigating time comparing our visualization with conventional textual-list interface. The bottom and the top line of the boxes represent the shortest and longest time cost, respectively; and the average time cost is represented by the symbol "+". From the Figure, it can be seen that using the proposed visualization can save time significantly. The average time cost is only 40 seconds with our proposed visualization interface, while it cost 74 seconds with the conventional interface. Without the visualization interface, the users can only linearly search through the file list based on very limited attributes (such as genre) to find their most preferred music when they have not listened to most of songs in the collection. Figure 6(b) shows the matching time comparison between the different interfaces. The average time cost is 42 seconds with our interface and 145 seconds with the conventional way.

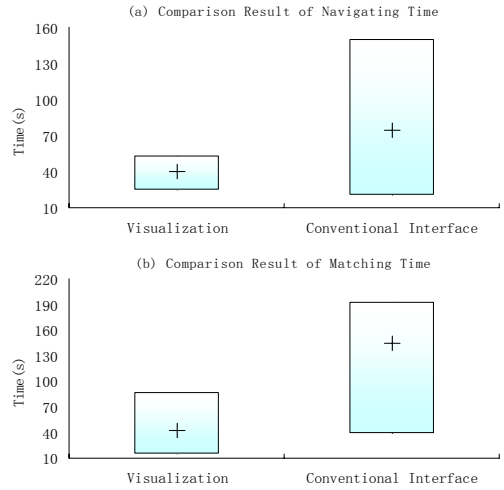

Figure 6. (a) Comparison result of navigating time. (b) Comparison result of matching time

The experiment results are encouraging. They indicate that our visualization interface could save both the navigating time and matching time compared with the conventional ways, which is one of the important factors in the success of a new music management system.

In the post-study questionnaire, the subjects are asked to rate each question based on three levels, including good, acceptable and bad. During the process, they are encouraged to switch between our visualization interface and their current music management software to make further comparisons.

Table 1 shows a brief summary of the results in the post-study questionnaire (on the first 3 questions). 7 subjects think that a timbre and rhythm analysis is definitely helpful for music management; all the 10 subjects say that the automatic perception estimation results are satisfactory or acceptable; and 9 subjects agree that the new visualization interface facilitates the process of navigating and managing a music collection.

Questions	Good	Accept-able	Bad
1. On the brightness-tempo 2D space	7	1	2
2. On the perception estimation	9	1	0
3. On the visualization interface	9	0	1

Table 1. A summary of the results in the post-study questionnaire

As for the extra feedbacks, almost all the feedbacks are positive. The subjects claim that they enjoy and like the creative interface. An interesting suggestion from the subjects is that it would be much better if we could analyze and present the gender of the artist in visualization since it is one of the most important factors in the music taste.

5. Conclusion

Based on the user study which indicates the perception consistency on music tempo and brightness among different individuals, in this paper, we build a novel 2D space to visualize a music collection, with dimensions of these two attributes. We also build a SVR-based model to map the high dimensional feature vector to one-dimensional tempo and brightness. Final user study compares the proposed visualization with the conventional textual list interfaces, and shows the new interface can save much navigating time and matching time in music selection and searching process. Feedbacks from the users also show that the proposed visualization could be a promising alternative user interface for the traditional music management tools.

6. References

[1] M. Torrens, P. Hertzog, J.L Arcos "Visualizing and Exploring Personal Music Libraries", I*nt. Symp. Music Information Retrieval (ISMIR04)*, 2004

[2] F. Vignoli. "Digital Music Interaction Concepts: A User Study", I*nt. Symp. Music Information Retrieval (ISMIR04)*, 2004

[3] Yossi Rubner. "Perceptual Metrics for Image Database Navigation", *Phd Thesis*, Stanford University, 1999

[4] B. Moghaddam, Q. Tian, Thomas S. Huang, "Spatial Visualization for Content-Based Image Retrieval", *ICME01*,pp.229-232, 2001

[5] L. Lu, H.J. Zhang, S. Li, "Content-based Audio Classification and Segmentation by Using Vector Machines", *ACM Multimedia Systems Journal*, 8 (6), pp. 482-492, 2003

[6] D. Liu, L. Lu, H.J. Zhang. "Automatic Music Mood Detection from Acoustic Music Data". I*nt. Symp. Music Information Retrieval (ISMIR03)*, pp. 81-87, 2003

[7] D. N. Jiang, L. Lu, H.J. Zhang, et. al. "Music type classification by spectral contrast features", *ICME02*, Vol. 1, pp. 113-116, 2002

[8] G. Tzanetakis, and P.Cook, "Music genre classification of audio signals", *IEEE Trans. Speech Audio Processing*, 10 (5), 293-302, 2002

[9] J. M. Grey, and J. W. Gordon. "Perceptual effects of spectral modifications on musical timbres." *J. Acoust. Soc. Am.*, Vol. 63(5), 1493-1500

[10] A. J. Smola, and B. Scholkopf, "A Tutorial on Support Vector Regression", *NeuroCOLT2 Technical Report Series*, NC2-TR-1998-30