

Music Database Retrieval and Media Conversion System Based on Impression

Toshihiko NODA*, Dong ZHAO * Hideyuki TAKAGI**

Kyushu Institute of Design, *Graduate School, **Dept. of Art and Information Design
4-9-1, Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
Phone & FAX: +81-92-553-4555, E-mail: takagi@kyushu-id.ac.jp

Keywords: MIDI database retrieval, media converter, image database, impression-based media retrieval, psychological space, neural network, genetic algorithms

Abstract— We propose a music database retrieval system based on *KANSEI* and a media converter that translates the same impression onto a different media. We first construct a universal psychological space that expresses human impressions of physical features from several kinds of media described by sets of adjective pairs and by the principle component analysis. Retrieving MIDI phrases is conducted by a genetic algorithm (GA) that searches a music physical feature space and a neural network that maps the searched physical feature vectors onto the psychological space. The distance between the mapped positions and target position in the psychological space is used for the GA search, and when the distance becomes less than a certain value, the retrieval search ends. By combining database retrieval systems for music and images with a common psychological space, a media converter is formed. We evaluate the MIDI file retrieval system with a subjective test and show that the system is efficient.

1 Introduction

In this current so-called multimedia era, media databases have become so large that efficient the development of retrieval methods for huge databases are in high demand. When we retrieve music or images, which is handled by our audio-visual sensors, it is important to consider how human users *feel*.

Some media retrieval methods using relationships among various kinds of media and their impressions have been proposed, for example, impression-based image retrieval using interactive evolutionary computation [2, 3] and retrieval with adjective pairs [1, 6].

In this paper, we present a MIDI database retrieval system based on a user's impression handled in a psychological space. This psychological space is a key component of our system and is constructed by applying principle component analysis to adjective-pair

sets.

One of our goals is to construct a media conversion system by combining multiple media database systems. The approach of our MIDI database retrieval system is similar to our image database retrieval based on impressions [4]. To combine the database retrieval systems for MIDI data and images and construct a media converter, we design a common psychological space for the two systems. We will describe this point in section 5.

We describe our music database retrieval system in section 2 and evaluate the system with subjective tests as described in section 3. After evaluating the experimental results, we discuss a media conversion system that combines the music retrieval system and the image retrieval system in section 5.

2 Music Retrieval System Based on Impression

Figure 1 shows the procedure of our music retrieval system. First, a user inputs a musical impression that he or she wants to retrieve into the system as the coordinate of a psychological space. Then, a genetic algorithm (GA) searches the physical features whose musical impression is similar to the impression that the user desires. The similarity of the searched musical phrase and the user's input is described as the distance between two coordinates in the psychological space. To measure the distance, a neural network (NN) maps the searched physical feature to the psychological space. Our system retrieves music based on the mapping relationship between the impression of the music in the psychological space and its physical feature in the physical feature space using GA and NN.

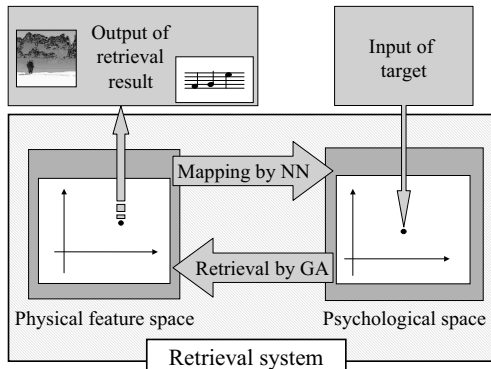


Figure 1: Our media retrieval system is based on impression. A user points a target impression in the psychological space, and GA searches media data whose impression is similar to the target impression. GA searches the media data in a physical feature space, and NN maps the coordinates onto the psychological space. A fitness value for the GA is the distance between the searched coordinate and target coordinate in the psychological space.

2.1 Construction of Psychological Scale Space

We construct a psychological space to express the impression of a retrieval target and the impressions of the searched musical physical features using the semantic differential (SD) method. The SD method reduces the redundancy among the adjectives pairs and obtains fewer factors describing the original adjective pairs by using principle component analysis. The principle component analysis is applied to the subjective evaluation data for each scale described by adjective pairs and obtains the factors in the adjectives pairs. The obtained factors forms the orthogonal factor axes of a space whose dimensional number is less than that of the original adjective pairs. We call this space a psychological space in this paper and use it to express our impression of a media as a coordinate of the space. Due to the smaller number of dimensions in the psychological space, it becomes easier for retrievers to express their impression of the search target.

The psychological space used in our system has four dimensions whose axes are shown in Table 1. These factors are obtained from the 14 adjective pairs listed in Table 2, and the 14 pairs are selected from 151 adjectives by eliminating the adjectives whose semantics may depend on a person’s interpretation, the similarity, or the type of media [4, 5]. Note that the psychological space is constructed by using images, but we

carefully select the 14 adjective pairs commonly applied to music and other media retrieval and construct a media converter in our next research step [5].

This psychological space is the core of several of our media retrieval systems. By installing the psychological space obtained from subjective evaluation enables computers to *feel* the impressions of music.

Table 1: Four factors used as the axes of our psychological space for music retrieval.

Activities factor	based on <i>passionate</i> , <i>jaunty</i> , etc.
Humanity factor	based on <i>warm</i> , etc.
Complexities factor	based on <i>complex</i> etc.
Spaces factor	based on <i>perspectively wide</i> , etc.

Table 2: Fourteen adjective pairs used to construct a psychological space. Original words used in the experiment [4],[5] were in Japanese.

<i>bright</i>	—	<i>dim</i>
<i>vivid</i>	—	<i>subdued</i>
<i>clear</i>	—	<i>fainted</i>
<i>gaudy</i>	—	<i>plain</i>
<i>passionate</i>	—	<i>dispassionate</i>
<i>hard</i>	—	<i>soft</i>
<i>jaunty</i>	—	<i>placid</i>
<i>pure</i>	—	<i>impure</i>
<i>warm</i>	—	<i>cool</i>
<i>simple</i>	—	<i>complex</i>
<i>comical</i>	—	<i>serious</i>
<i>emotionally attractive</i>	—	<i>emotionally unattractive</i>
<i>perspectively wide</i>	—	<i>perspectively narrow</i>
<i>dry</i>	—	<i>wet</i>

2.2 Physical Features of MIDI Data

The music database consists of eight-measure phrases of standard MIDI files collected from the Internet. Musical physical features are obtained from each phrase; they are the number of the notes, the average and variance of the length, and the pitch of notes for every two measures. This means that one phrase of eight measures has a vector of $5 \times 8/2 = 20$ -dimension physical features.

2.3 Mapping Relationship from a Physical Feature Space to an Impression Space

To obtain the mapping relationship from a physical feature space to an psychological space for impression, we train a feed-forward NN that inputs a 20-D of physical features of MIDI data and outputs the vector of the 4-D psychological features [5]. The NN has one

hidden layer consisting of 14 neurons. This number of neurons is determined by changing the number of neurons from 10 to 30 and comparing their training errors.

Training data for the NN are the coordinates of the psychological space that express the impression of human subjects for music. MIDI files are divided into eight measure phrases for a total of 250 phrases, and the obtained 250 phrases are displayed to 13 subjects. The subjects are requested to express their musical impressions for given musical phrases at five levels on each of four axes whose factor names are listed in Table 1. The four axes form a 4-D psychological space, and the musical impressions of subjects are expressed as the coordinates of the psychological space.

2.4 Database Retrieval by GA and NN

Our MIDI database retrieval system searches music phrases whose impressions are close to the target impression. The search in the the physical feature space is conducted by GA, and the mapping searched vectors from the physical feature space to the psychological space is conducted by an NN. Once the searched vectors are mapped on a psychological space, the distance between the mapped coordinate and the target coordinate in a same space. This distance is used for GA search in the physical feature space as a fitness value [5].

The following total retrieval procedure is based on Figure 1. First, a user inputs a musical impression that he or she desires to retrieve as a coordinate in the psychological space whose axes are the scale of factors shown in Table 1; this is start of the database retrieval. The retrieval system randomly selects some coordinates in the physical feature space as GA individuals in the first generation. The individuals are mapped in a psychological space by an NN. The distances between the target coordinate given by the user and the coordinates where individuals are also mapped. These distances are fed-back into the GA as fitness values, and the GA generates offspring physical feature vectors. This search is repeated until the distance becomes less than a certain threshold, and the retrieved musical phrases are displayed to the user.

This is the first retrieved result shown to the user while GA and NN iterates their internal searches using two spaces. As the psychological space is designed for an average of 13 subjects, there may be differences in the impressions of the four factors among the users. If the first retrieved result is different from what the user expects, the user, modifies the input coordinate for music retrieval according to the difference between

the retrieved result and his or her impression in the metric of the psychological space. This user-computer interaction is repeated until desired musical phrase is retrieved.

3 Evaluation of Proposed Music Database Retrieval System

In this section, we evaluate how the retrieved outputs of our system match the user's query intention with subjective tests. The database used in this experimental evaluation consists of 4064 8-measure phrases extracted from 281 MIDI files, i.e. 281 pieces of music. Experimental conditions of GA are: 95% crossover rate, 5% mutation rate, 100 population size, and the 5,000 maximum searching generations.

3.1 Evaluation Procedure

The performance of our retrieval system is measured as the distance between the impression of the received musical phrase and the target impression in the psychological space.

The target impression, (x_1, x_2, x_3, x_4) in the 4-D psychological space, is the position in the (*activity factor*, *humanity factor*, *complexity factor*, *space factor*) space, where $1 \leq x_i \leq 5$. This scale originates from the rating of five levels. To avoid biased evaluation, uniformed impression targets in the psychological space are chosen as the 17 coordinates of $(3, 3, 3, 3)$, the center of the space, and $(3 \pm 1, 3 \pm 1, 3 \pm 1, 3 \pm 1)$, 16 neighbor points of the center.

Subjects listen to the best 5 searching results for the 17 target impressions and point out the coordinates of their impression in the same psychological space. As the subjects express their impression at five levels on each axis, we may suppose the differences of 0.5, 1, 1.5, and 2 on a 1-D axis correspond to *very close*, *close*, *a little farther*, and *far*, respectively. Expanding this assumption to the distance in a 4-D space, we may suppose the differences of 1, 2, 3, and 4 correspond to *very close*, *close*, *a little farther*, and *far*, respectively, because Euclidian distance in a 4-D space is given by

$$\sqrt{\sum_{i=1}^4 d_i^2}.$$

3.2 Retrieved Results

Searching for the 5 five musical phrases for 17 target impressions results in $5 \times 17 = 85$ phrases. We finally obtained 46 phrases after avoiding duplicate searching results.

We compare the retrieved results for two opposite targets which are symmetrical with respect to the center of the psychological space, $(3, 3, 3, 3)$, i.e. $(3 - a, 3 - b, 3 - c, 3 - d)$ and $(3 + a, 3 + b, 3 + c, 3 + d)$ are compared, where a, b, c , and d are either 1 or -1 .

The outputs of these symmetrical targets have few duplications among the 4,064 phrases. For example, the all outputs for the target, $(2, 2, 4, 2)$, are the Concerto of J. S. Bach, and those for the opposite one, $(4, 4, 2, 4)$, are popular music.

We examine the continuity of the mapping from a music physical feature space to a psychological space. Figure 2 shows the best five retrieval outputs for the three targets: the center of the psychological space, $(3, 3, 3, 3)$, and two opposite points, $(2, 2, 4, 4)$ and $(4, 4, 2, 2)$. This figure shows that retrieved results gradually changed according to the position in a psychological space.

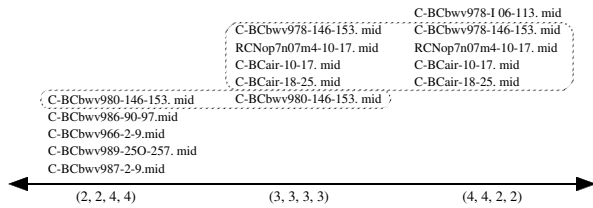


Figure 2: Psychological space coordinates of three target impressions and the best five searching results for each target.

3.3 Subjective Test for Output of the System

Subjects listen to the 46 phrases obtained in section 3.2 and evaluate their impression position in the 4-D psychological space mentioned in section 2.1. Six subjects who joined experiment in section 2.3 listened to the phrases and evaluated their impression at five levels for each axis.

We examine the distances between the impression positions of the target and the average for six subjects. Table 3 shows the number of the phrases and the percentage for each distance. Phrases whose distances are less than or equal to 3, i.e. those evaluated within *very close*, *close*, and *a little farther*, occupy 84.7% of 85 phrases.

Table 4 lists phrases whose evaluations were the best five and the worst five with their target impression coordinates and their average evaluation in the 4-D psychological space. It shows the concrete samples of upper and lower distribution of 6% per each;

Table 3: The distances between the target impression and the averages of the six subjects' evaluations on 4-D psychological space and the number and ratio of 85 phrases.

distance on 4-D space	0-1	1-2	2-3	3-4	4-
# of phrase	2	23	47	13	0
percentage (%)	2.4	27.0	55.3	15.3	0

the best five are included in the evaluation of *very close* or *close*, and the worst five are included in *a little farther* and *far* in our supposed evaluation categories in a 4-D space. For example, in the best one's case, distance = $\|(2, 4, 2, 4) - (1.5, 4, 2.5, 3.5)\| = \|(0.50, 0.00, 0.50, -0.50)\| = 0.86$ (*very close*).

4 Discussion

The result in section 3.2 that 84.6% of retrieved results are *very close* or *close* to the given target impression implies high efficiency of our music database retrieval system for different retrievers. Note that this performance was obtained in just only first retrieval and users who could not be satisfied with the retrieved musical phrases continued on to the next retrieval and narrowed the searching phrases down by considering the difference between their impression in the psychological space.

In this experiment, 37 pieces of music from 281 pieces of music are retrieved in this experiment. Although we cannot say that all phrases of the same music make the same impression, the fact that the retrieved results are biased to different measures in the same piece of music implies that there is a tendency that the music in our MIDI database makes similar impressions and that the same music has similar physical features. This may come from the fact that our database includes much baroque music and modern music; if there were many symphonies of romantic music composers, the change of phrase impressions within same music would had become larger than that of baroque music.

To obtain the better performance in Table 3, two conditions must be satisfied: (1) NN mapping from a physical feature space to a psychological space must be precise, and (2) the variation of definitions of impression among humans must be few. The high performance of 84.7% retrieved phrases were less than or equal to little farther shows that these two conditions are roughly satisfied.

Table 4: Retrieved results of given target impression and music phrases. Columns mean retrieved phrase, given target coordinates in psychological space, average of subjective evaluations for the retrieved phrases, and the distance between the coordinates of the target impressions and those of the average impression of the evaluational coordinates.

(a)The best five phrases.

composer, title of the music, measures	target	evaluation	distance
L. Grandell "Fake Sex"(58-65)	(2, 4, 2, 4)	(1.50, 4.00, 2.50, 3.50)	0.86
J. S. Bach "Air on the G string" (18-25)	(4, 2, 2, 2)	(3.50, 1.83, 2.66, 1.50)	0.98
J. S. Bach "BWW966" (2-9)	(2, 2, 4, 4)	(2.66, 2.33, 3.00, 3.66)	1.29
L. Grandell "Fake Sex" (66-73)	(2, 4, 2, 4)	(1.16, 3.50, 1.83, 3.16)	1.29
J. S. Bach "BWV1003" (26-33)	(4, 2, 4, 2)	(4.16, 3.00, 3.16, 2.50)	1.40

(b)The worst five phrases.

composer, title of the music, measures	target	evaluation	distance
A. L. Albinoni "Adagio in G minor" (138-145)	(2, 4, 2, 4)	(4.16, 2.50, 3.66, 1.50)	3.99
J. S. Bach "Air on the G string" (18-25)	(2, 4, 2, 4)	(3.50, 1.83, 2.66, 1.50)	3.69
J. R. Bosset "Time For Me" (66-73)	(4, 2, 4, 4)	(1.16, 3.16, 2.16, 3.50)	3.60
J. S. Bach "Air on the G string" (10-17)	(2, 4, 2, 4)	(3.83, 1.83, 2.33, 1.83)	3.58
G. Pacchioni "The first movement of Concerto in C minor, Op.43" (26-33)	(4, 4, 4, 2)	(1.83, 2.33, 2.16, 2.66)	3.35

5 Media Converter: Preliminary Research

One of our final goals is to convert media such as music, images, and movies with keeping same impression. The point is to keep same impression as humans feel. Our approach to realize this purpose is to commonly use a universal psychological space, to connect each media retrieval system, and realize mutual media retrieval through the common psychological space [5] (see Figure 3.) The input to the media converter is media themselves, while the input to our media retrieval systems is the impression expressed as a coordinate in a psychological space. Since each media retrieval system has three functions of calculating a physical feature vector of inputted medium, mapping from a physical feature space to a psychological space by an NN, and searching physical feature vectors by GA, it is easy to use the mapped point in a physical feature space as an input impression to other media retrieval systems. This mapping from a space to another space realizes a media converter, and mapping through common psychological space realizes kept same impression.

There are several possible applications of the media converter. When we want to attach music to our web page, photo or computer graphics, or home video movie, the media converter finds out suitable music whose impression is similar. Another possible application is visualization of sound or music. It is impos-

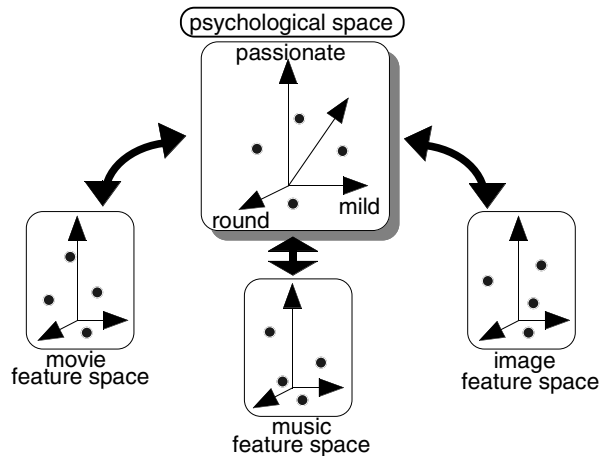


Figure 3: Outline of the media conversion system. When user inputs a image to this system, system estimates impression of the image, retrieves music that give user similar impression to the input. User input his expected impression directly to retrieve media.

sible to glance music except their keywords. If auto-indexing for music with images of same impressions is realized, it becomes easier for us to view the contents of music database at once.

We combine the the MIDI database retrieval system proposed in this paper and our previous image retrieval system manually and preliminarily check the converted output of the media converter. The im-

age database retrieval system is same structure as the MIDI database retrieval system except using a 20-D image physical feature space.

We inputted the images in Figure 4 to the image database retrieval system and obtained the mapped coordinate in the psychological space. Then, we read the the mapped coordinate and inputted it to the MIDI database retrieval system manually. The output of the MIDI database retrieval system was 17th – 24th measures of *Aria On the G String* by J. S. Bach. Since our image database retrieval system uses color information as physical features, Figure 4 may not show all information. As we have not evaluated how these images and the retrieved music are similar, we do not discuss in detail in this paper; we would like to leave it to readers' discretion.



Figure 4: When these two images are inputted to future our media converter, it is expected that *Air On the G String* by J.S. Bach is retrieved.

6 Conclusion

We proposed a multimedia retrieval system based on impression and a media conversion system as our next research. To realize the media retrieval system and the media converter; first, we constructed a common psychological space used by these systems; second, we constructed a MIDI database retrieval system; and then, we evaluated the system with subjective tests. Higher retrieval performance was shown. Finally, we manually constructed a media converter and show an example of the media conversion from images to music. Our next step in this research is to construct the automated media converter and evaluate the similarity of the impression of the converted media have to the original media.

Acknowledgements

We are grateful to Dr. Akito Teraoka of Institute of Systems & Information Technologies/Kyushu for his programming to handle MIDI files.

References

- [1] Hayashi, T. and Hagiwara, M., “An image retrieval system to estimate impression words from images using neural network,” IEEE Int'l Conf. on Systems, Man and Cybernetics (SMC'97), Orlando, Florida, USA, pp.150–155 (Oct., 1997).
- [2] Lee, J.-Y. and Cho, S.-B., “Interactive genetic algorithm for content-based image retrieval,” Asian Fuzzy Systems Symposium (AFSS'98), Masan, Korea, pp.470–484 (June, 1998).
- [3] Lee, J.-Y. and Cho, S.-B., “Interactive genetic algorithm with wavelet coefficients for emotional image retrieval,” Int'l Conf. on Soft Computing and Information/Intelligent Systems (IIZUKA'98), Iizuka, Fukuoka, Japan, pp. 829–832 (Oct., 1998).
- [4] Takagi, H., Cho, S.-B., and Noda, T., “Evaluation of an IGA-based image retrieval system using wavelet coefficients,” IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE'99), Seoul, Korea, Vol.3, pp. 1775–1780 (Aug., 1999).
- [5] Takagi, H., Noda, T., Cho, and S.-B., “The Psychological Space of Common Media Impressions Held in a Media Database Retrieval System,” IEEE Int'l Conf. on Systems, Man, and Cybernetics, Tokyo, Japan, Vol.4, pp. 263–268 (Oct., 1999).
- [6] Yoshida, K., Kato, T., and Yanaru, T., “Image retrieval system based on subjective interpretation,” Int'l Conf. on Soft Computing and Information/Intelligent Systems (IIZUKA'98), Iizuka, Fukuoka, Japan, pp. 247–250 (Oct., 1998).