Media Converter with Impression Preservation Using a Neuro-Genetic Approach

Takagi, Hideyuki
Fcaulty of Design, Kyushu University : Professor

Noda, Toshihiko
Graduate School, Kyushu Institute of Design

Media Converter with Impression Preservation
Using a Neuro-Genetic Approach

Hideyuki TAKAGI[(1)]**and Toshihiko NODA**[(2)] [*]

(1) Faculty of Design, Kyushu University, (2) Graduate School, Kyushu Institute of Design
4–9–1, Shiobaru, Minami-ku, Fukuoka, 815-8540 Japan
takagi@design.kyushu-u.ac.jp

Abstract. We propose a media converter framework which takes images, music, and other media as input and outputs different media while preserving the same impression for humans. The media converter is realized by combining multiple media database (DB) retrieval systems that use a common psychological impression space. Each of the media DB retrieval systems consists of a physical media feature extraction part, a physical feature space for the extracted features, an impression space where the psychological impressions of media are expressed as their coordinates, a neural network (NN), and a genetic algorithm (GA). The NN maps the coordinates of the media in the physical feature space to those in the impression space, and the GA search the coordinates of media in the physical feature space using the target coordinate in the impression space and the NN mapping. A user specifies a coordinate in the impression space that corresponds to his or her target impression. The media DB retrieval system extracts features of the images, music, or other media on the Internet or a commercial packaged media DB and stores them in inside as physical feature space coordinates. Media whose impression is similar to the user-specified target impression are searched by an NN and GA. The media converter is realized by combining multiple media DB retrieval systems. A medium, medium A, is converted as follows: (1) it is mapped from a physical feature space A to an impression space by NN_A and (2) medium B is searched for in a physical feature space B from the impression space using GA_B. Prototypes of an image DB retrieval system and a music DB retrieval system are made and evaluated for their mapping & searching performance. Finally, we make a prototype of a media converter by combining the media DB retrieval systems and show the potential of its realization.

Keywords: media converter, media database retrieval system, neural networks, genetic algorithms, *KANSE* information processing.

1   Introduction

The rapid spread of broadband throughout homes has made it easier to obtain multimedia such as images or music through the Internet. To Internet users, it looks like the Internet links to a huge media database (DB). In fact, it is already very common for Internet users to access these media through web browsers and retrieve them.

There are three main methods for retrieval from a media DB: searching using keywords, searching for similar images or music, and searching based on human impression.

The first method is to find media files whose associated keywords are the same or similar to the keywords inputted by the user. Keywords for image file CD-ROMs or other practical commercial DBs are entered manually (Chang 1986). Robot search engines, such as the one used by google, search by keywords not only in media file names but also in text files linked to the media files. There is ongoing research into methods of automatically associating keywords to media files. These automatic methods compare physical features of the target media, such as histograms or wavelet transforms, with those of the training media, and attach the training media's keywords to the target media if their physical features are similar (Amagasa et al. 2000). Another approach is to match keywords on the semantic level. This enables searches to be performed not only on symbolically identical keywords but also on semantically similar ones by introducing a thesaurus and measure of distance between keywords (Motro 1988).

The second method is to draw images or input media similar to the desired target media and retrieve images whose physical features are similar to those of the inputted images (Kato et al. 1991). In music retrieval, a corresponding method is used whereby a melody is inputted using a keyboard or by humming, singing,

---

[*]He is working for Sony now.

or tapping and the system searches for music containing similar phrases (Kageyama et al. 1993, Ghias et al. 1995, Chen&Jang 1998, Kataoka et al. 1998, Sonoda et al. 1999, Tsien 2000, Jang et al. 2001).

The third method is impression-based media retrieval. An impression space is designed from pairs of polar adjectives or by a factor analysis such as the semantic differential (SD) method. Here the mapping the relationship between a physical feature space and the impression space is defined by a certain mapping method, and a target media in the physical space is retrieved from the given coordinate in the impression space. A user either specifies a coordinate in the impression space (Hayashi&Hagiwara 1997, Yoshida et al. 1998) or gives an evaluation value on similarity to the retrieved media by interactive evolutionary computation (IEC) (Takagi 2001) and lets the IEC search the target media (Cho&Lee 2002). When the IEC approach is adopted, it is also possible to access the physical feature space directly without using the impression space.

The approach of this paper is an extension of the third method. The first feature of this paper is to propose and evaluate media DB retrieval systems that use hybrid neural network (NN) and genetic algorithms (GA) to search for target media in a physical feature space corresponding to the given impression coordinate in a psychological impression space. The second feature is to propose a framework of a media converter with impression preservation that combines multiple media DB retrieval systems and evaluate its realized potential.

We describe the structure of the proposed media DB retrieval systems and the media converter and design their parts in section 2. We evaluate the media DB retrieval systems, the most important component of the media converter, in section 3 and evaluate the media converter in section 4.

2 Proposed of Media Converter

2.1 Proposed systems

The proposed media converter is a system that converts media to other media whose impression is similar, e.g. it outputs music whose impression is similar to that of an inputted image. If one were to anthropomorphize, it is as if the computer were to listen to music, understand how humans feel the music, search for images whose impression is similar to that of the music, and then display those images.

The media converter consists of an impression-based image DB retrieval system (Takagi et al. 1999), an impression-based music DB retrieval system (Noda et al. 2000), and other media DB retrieval systems (see Figure 1). A media DB retrieval system consists of a part for extracting the physical features of the media, a physical feature space, an impression space that expresses the psychological impression of media, an NN that maps coordinates from the physical feature space to the impression space, and GA that search for coordinates in the physical feature space, based on a coordinate in the impression space, by the cooperation with the NN.

A media converter is realized by designing a common impression space for multiple media and combining multiple media DB retrieval systems through the common impression space. Main idea is to involve the unique common impression space as a *bridge* connecting the multiple media DB retrieval systems.

The searching accuracy of the systems depends on the design of the impression space. If the space is designed by only one user, the impression of the input and output media should be close for that user; if it is
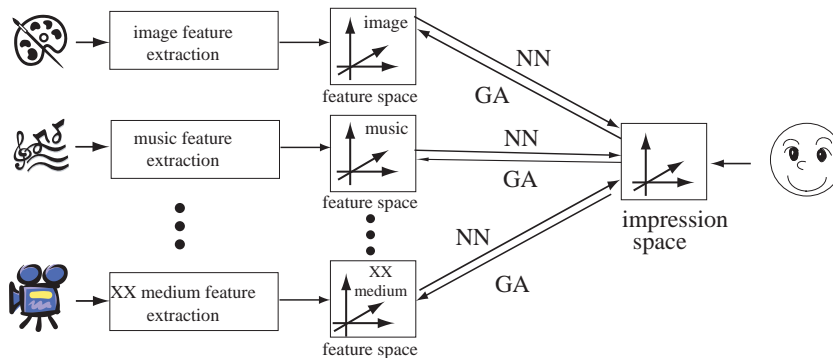

Figure 1. System framework of a proposed media converter. Users indicate the coordinates in an impression space to express impressions in their minds when they use media DB retrieval systems. When the media DB is part of a media converter they do not directly interact with the impression space.

designed by multiple users, the impressions of input and output media should be close for the average of the group. Naturally the searching accuracy of the former is higher.

Although the proposed media converter can theoretically convert any medium to any other medium, our prototype of the media converter designed and evaluated in this paper converts between images and music.

2.2 Impression space

The psychological space for expressing human impressions must be common to all the media used in the media converter. Conversion performance expressed by the distance in the impression space between the two media before and after media conversion depends on the design precision of the impression space. As it takes several weeks to design a psychological space with subjective tests, it is too costly for us to design several impression spaces for each of multiple users. At the same time, it is necessary to increase the number of subjects to increase the reliability of evaluation. Thus, we decide to design only one common impression space with multiple subjects, implement the impression space for the image and music DB retrieval systems and media converter, and evaluate the proposed system.

We use the semantic differential (SD) method to extract factors from pairs of polar adjectives by a factor analysis for designing the impression space.

(a) Extraction of pairs of polar adjectives

We displayed 50 images to 3 subjects, asked them to list any adjectives that they felt, and collected a total of 151 adjectives. We then deleted the adjectives whose impressions were person-dependent or image-dependent and merged them into semantically similar adjectives. We especially paid attention to delete adjectives that represented different impressions if used for images and music when implementing the media converter prototype between images and music. Finally, the 14 pairs of polar adjectives shown in Table 1 were obtained. Note that the adjectives are translated here from the original Japanese, and the impressions given by the translations may not be the same as the the original Japanese and may depend on culture.

Table 1. English translation of 14 pairs of polar adjectives used to construct an impression space.

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

(b) Selection of images

To design an impression space with the SD method, 610 images were culled from a total of 3,000 images by deleting similar images.

(c) Designing an impression space by factor analysis

An image is displayed to a subject and he or she is requested to rate the impression with 14 pairs of polar adjectives in 7-step scales. This rating is conducted for 10 subjects $\times$ 610 images. These data were analyzed by factor analysis. When we approximated the 14 dimensional (14-D) adjective space with a 3-D to 6-D factor space, the accumulated cover rates were 59%, 67%, 73%, and 77%, respectively. We chose the four factors shown in Table 2 as a balance between their lower dimensionality and the higher coverage rate of the original 14 adjectives. Using these four factors for the axes, a 4-D impression space was thus constructed.

Table 2. Four axes constructing an impression space using in the image and music DB retrieval system and a media converter in this paper.

(1) *dynamic* factor, such as 'passionate', 'jaunty', etc.
(2) *emotional* factor, such as 'warm', "emotionally attractive", etc.
(3) *complex* factor, such as 'complex', etc.
(4) *spatial* factor, such as "perspectively wide", etc.

2.3   Physical features for DB retrieval

We adopt 6 types of physical features for an image DB retrieval, for a total of 21 dimensions. The features are 9-D brightness distribution, Fourier spectrum, horizontal cross-correlation among pixels, 4-D hue distribution mapped to both the vertical and horizontal axes, brightness distribution mapped to the vertical axis, and a clustering value of pixel saturation.

Our experimental music DB consists of eight-measure phrases of standard MIDI files collected from the Internet. The musical physical features were obtained from each phrase and aggregated for every two measures. The features are: the number of notes, the average and variance of the note lengths, and the pitch of the notes. This means that 1 phrase of 8 measures has a vector of $5 \times 8/2 = 20$-D physical features.

These physical features are selected by taking account of Table 2.

2.4   NN-GA hybrid search for physical features and impression spaces

In impression-based media DB retrieval, users specify their target impression as a coordinate in an impression space, and a search system finds a different media whose coordinate in a physical feature space is closest to the coordinate corresponding to the target impression. However, it is not easy to map from the lower dimensional (4-D) impression space to the higher dimensional (20-D or 21-D) physical feature space even if we seek to only find one or a few target media among many corresponding ones in the physical feature space.

Our solution is to use an NN-GA hybrid search. We make an NN train the mapping relation from the higher dimensional physical feature space to the lower dimensional impression space, which is a natural application for NNs. When a media coordinate in the physical feature space is inputted to the trained NN, the NN outputs the corresponding coordinate in the impression space. As the distance between the mapped coordinate and the user specified coordinate can be used as a search performance index, the Euclidian distance between the two coordinates is used as a fitness value. Based on this fitness value, the GA search for the best media in the physical feature space.

The NN configuration was determined by comparing the computational cost and convergence performance in a preliminary experiment. The obtained number of neurons of the three layer NNs are $(20 \times 32 \times 4)$ and $(21 \times 27 \times 4)$ for image and music DB retrieval systems, respectively, from input layer to output layer. Although a media DB retrieval system and a media converter use NNs, their users do not train the NNs but use previously trained NNs. GA coding parameters are a coordinate in the physical feature space, and the GA parameters are shown in Table 3 in the next section.

Unlike media DB retrieval systems, human users do not specify the target coordinate in an impression space; instead a media converter automates the process without human intervention. However, because the impression space is designed based on human impression characteristics and a NN-GA hybrid search system is used, so the system is inside the same as media DB retrieval systems.

3   Evaluation of Media DB Retrieval Systems

3.1   Experimental method

The performance of our proposed media converter is decided by the performance of media DB retrieval systems that form the converter. These retrieval systems are evaluated by measuring how close the impressions of the retrieved media are to the target impression specified in the impression space by the users.

In this sub session, we explain how experimental evaluation is conducted in detail with case of an image DB retrieval system. The same experimental evaluation is conducted for a music DB retrieval system as well. See the experimental conditions for both retrieval systems in Table 3.

Target Media: A target images is displayed to an experimental subject, and he or she is requested to identify images whose impression is similar. This retrieval experiment is iterated for 17 target images $\times$ 7 subjects. To obtain the 17 target images, 17 coordinates are input into the impression space. These coordinates are equally distributed in the space in order to evaluate the image DB retrieval system without bias. Given that the axes are scaled from 1 to 5, the 17 coordinates selected are: the center coordinate (dynamic factor, emotional factor, complexity factor, spatial factor) = (3, 3, 3, 3) and the surrounding 16 coordinates (3 ± 1, 3 ± 1, 3 ± 1, 3 ± 1). The user's task is not to search for images identical to the given targets but rather for images of similar impressions, which is close to the real use of the media DB retrieval system.

Procedure: Each subject views a displayed target image and rates the impression of the image on the 5-

Table 3. Parameters used in image DB retrieval. The GA stop condition is met when either the termination generation is exceeded or the distance between the searched image and the target image becomes less than a threshold.

parameters	value
# of images in the image DB	12,000
# of music phrases in the MIDI DB	4,064
# of dimension of the image feature space	21
# of dimension of the music feature space	20
crossover rate	95%
mutation rate	2%
population size	50
termination generation	5,000
threshold distance for termination (image)	0.05
threshold distance for termination (music)	0.20

step scale for each of the four factor axes. This 4-D vector in the impression space is the user's impression's target coordinate, and an image DB retrieval experiment starts. If, after the maximum 5,000 generations, the user feels that the impressions of the 5 best images that the NN and GA found are different from the target image, he or she adjusts the input target coordinate and tries again. The experimental evaluation system is made using a CGI script to collect the subjects' evaluation data through a web browser.

Numerical Evaluation: The best image, the image whose impression is the closest from amongst the five presented to that of the target image, is mapped to the impression space along with the target image. The performance of the image DB retrieval system is evaluated by the distance between the two images in the impression space. We use Manhattan distance in a 4-D impression space whose axes are all in 5-step scales. This means that the maximum distance in this space is 16 and that 2 images whose mutual distance is 1 are, on average, 0.25 far from each other on each axis.

Convergence of User Trials: In the case where the user was forced to adjust the target coordinate, convergence was evaluated by measuring whether the results in the impression space began to approach the target image. The distance between the target coordinate and the mapped coordinate of the best image that GA found in a physical feature space as an elite is measured for this convergence evaluation.

Subjective Test: The target image and the best 34 images retrieved in this experiment are randomly displayed to the user one after the other. The user evaluates the difference between their impressions of the retrieved images and the target image on a 5-step scale for the four factor axes of the impression space. This subjective evaluation is conducted approximately two weeks after the retrieval experiment without informing the subjects about the purpose of this subjective evaluation in order to avoid a bias from the users' memories of their impressions. Through this experiment, we can evaluate how well each of the users could retrieve images using our image DB retrieval system.

3.2 Experimental results

The distances between the target images and retrieved images in an impression space for each user and for each factor axis are shown in Table 4 (a) and (b), respectively. The same measurement results for music retrieval from the MIDI phrase DB are also shown in the table. There is no significant difference between users; the retrieval performance both systems was stable.

The average distances, 1.30 and 2.37, of image and music DB retrieval systems in the 4-D space corresponds to mean distances of 0.33 and 0.59 along each of the 5-step factor axes, as seen in Table 4. From these performance results, we can say that these two impression-based DB retrieval systems can find images and music phrases of similar impression to the target media.

When a media DB retrieval system is used alone, a user specifies the impression of target media on an impression system and runs the system, observes the retrieved media, and adjusts the target coordinate and tries again if the obtained media satisfy the user. This iteration is similar to IEC process.

Figure 2 shows the convergence performance for this case; this is an example from subject A among seven subjects. The vertical axis represents the distance in the impression space, and the horizontal axis represents the normalized number of trials. When a user adjustment resulted in a big distance, the distance decreased in the following trial. This change means that the user reacted to an adjustment that was too large and it also suggests that the user may have been trying to find the real target through trial-and-error. It may not be possible

Table 4. Manhattan distances between the target medium and the best retrieved media in an impression space: (a) per each user, and (b) per each factor axis .

(a)

subjects	images		music phrases	
	average distances	standard deviations	average distances	standard deviations
A	1.28	0.77	2.49	2.67
B	1.30	0.62	2.06	2.36
C	1.23	0.72	2.21	2.49
D	1.57	0.72	2.75	2.98
E	1.17	0.42	2.62	2.17
F	1.29	0.59	1.81	2.30
G	1.25	0.43	2.65	3.09
average	1.30	0.64	2.37	2.62

(b)

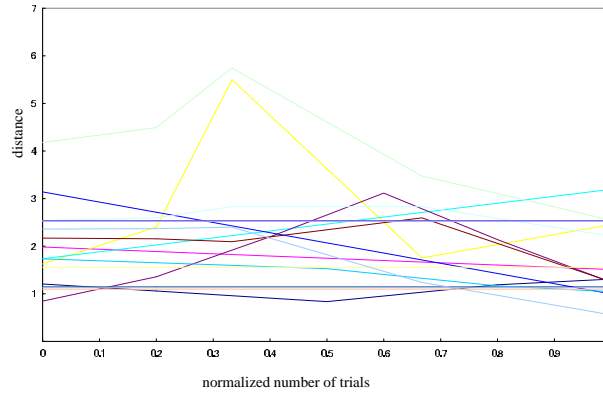| subjects	images		music phrases	
	average distances	standard deviations	average distances	standard deviations
dynamic factor	0.36	0.27	0.90	1.20
emotional factor	0.35	0.27	0.39	0.49
complex factor	0.29	0.22	0.69	0.89
spatial factor	0.30	0.24	0.39	0.52

Figure 2. Convergence characteristics of several retrieval experiments by the subject A among seven subjects. The horizontal axis is the normalized number of trials from 0 to 1, and the vertical axis shows distances between the target image and retrieved images in the impression space.

to prevent this trial-and-error method of adjustment unless the impression space is designed specifically by and for one user. Similar characteristics were found in music DB retrieval.

Subjective test results are shown in Table 5 for the image and music DB retrieval systems. Each of the best 34 images retrieved in this experiment was displayed to users in random order approximately two weeks later and the differences in impression were evaluated for each of the 4 factors on a 5-step scale.

4   Evaluation of Media Converter

The performance of our media converter is evaluated for four kinds of conversion: image to image, image to music, music to image, and music to music. Its performance is the product of the performances of both the image and music DB retrieval systems evaluated in section 3. The difference of impressions between media before and after conversion is evaluated by the distance between the two media in the impression space. The media converter includes 12,000 images and 4,064 music phrases in the image and MIDI DBs.

Table 5. Users' subjective evaluation for the distances between the target medium and the retrieved media on a 5-step scale.

subjects	images		music phrases	
	average distances	standard deviations	average distances	standard deviations
A	2.94	1.95	2.06	1.76
B	4.12	1.28	3.53	2.38
C	3.88	2.40	3.65	2.70
D	4.47	2.81	3.65	2.70
E	4.18	1.58	4.76	2.96
F	3.41	3.09	3.29	3.30
G	4.65	2.25	4.35	1.45
average	3.85	2.34	3.61	2.66

Table 6. Distances between media before and after conversion in the impression space.

media conversion	average distances	standard deviations of distances
image → image	1.50	0.75
image → nusic	3.70	1.01
music → image	3.44	1.02
music → music	1.73	1.24

The GA termination condition for the media DB retrieval systems inside the media converter is that the distance in the impression space between the corresponding coordinate mapped from the input medium by an NN and those mapped from the output media found by the GA becomes less than 0.35.

Average distances and standard deviations of media before and after conversion are shown in Table 6. The distances are measured between the input medium and the best converted medium in the impression space, though the media converter outputs the best five.

All distances in Table 6 are less than 4, so given that the impression space is 4-D and with 5-step scaled axes, it can be said that the converted medium preserves a certain amount of the impression of the input medium from the distance point mentioned in section 3.1.

5   Discussion

It seems strange that the distances in Tables 4 (a) and 5 are different despite being measured in the same impression space. The distances in the former are calculated with real numbers, while those in the latter are calculated with integers from 1 to 5 on each factor axis, i.e. the latter includes round errors of ± 0.5 per each of 4 axes. This round error results in a maximum error of 2 in coordinate positions, so that the maximum possible rounding errors between two media becomes 4. If this view is correct, the difference in Tables 4 (a) and 5 seems reasonable.

The performance of the music DB retrieval system is slightly poorer than that of the image DB retrieval system. Table 4 (b) implies a method that might be used to improve its performance; we need to find better physical music feature parameters, especially to express the *dynamic* factor well. Collecting diverse MIDI phrase data with a wide variety of impressions is another important factor for improving the performance. A good mapping relation cannot be trained without wide coverage of the impression space.

In fact, our analysis shows that the distribution of the 4,064 MIDI data is not uniform, but concentrates towards the center of impression space, (3, 3, 3, 3). Music training data of extreme impression may be needed to cover the peripheral areas of the impression space.

We designed the impression space with the data of 10 subjects. If, instead, the impression space were to be designed for a specific user, the performance of proposed systems must be improved for that user. When critical performance is required for certain tasks, this is one immediate solution.

6 Conclusion

We proposed an NN-GA approach for an impression-based media DB retrieval system and a framework for a media converter consisting of the media DB retrieval systems. We showed that the retrieval systems could retrieve images and music phrases whose impressions are similar to a target impression through experimental evaluation.

We also evaluated a prototype of a media converter and demonstrated that the converted media preserved, to some degree, a similar impression to input media.

Still, there are many places left where we can further improve our systems, such as the bias of the impression space which can be decreased in the future by increasing the number of media samples. Ultimately, we imagine that, by overcoming these issues, we can realize a computer that *feels* the taste of images and music.

References

Amagasa, T., Nakai, M., Hatano, K., Yoshikawa, M., and Uemura, S. (Sept., 2000), "Keyword assignment to images using sliding windows," *Proc. of 2000 ADBIS-DASFAA Symposium on Advances in Databases and Information Systems*, Prague, Czech Republic, pp. 1–10.

Chang, S. K. (1986), "Image database systems," *Handbook of pattern recognition and image processing*, Young, T. Y. and Fu, K. S. (eds.), pp. 371–393, Academic Press.

Chen, B. and Jang, J.-S. R. (Aug., 1998), "Query by Singing," *11th IPPR Conf. on Computer Vision, Graphics, and Image Processing*, Taiwan, pp. 529–536.

Cho, S.-B. and Lee, J.-Y. (2002), "A human-oriented image retrieval system using interactive genetic algorithm," *IEEE Trans. on Systems, Man and Cybernetics, Part A*, vol. 32, no. 3, pp. 452–458.

Ghias, A., Logan, H., Chainberlin, D., and Smith, B.C. (Nov., 1995), "Query by humming: musical information retrieval in an audio database," *ACM Conf. on Multimedia*, pp. 231–236.

Hayashi, T. and Hagiwara, M. (Oct., 1997), "An image retrieval system to estimate impression words from images using neural network," *IEEE Int. Conf. on Systems, Man and Cybernetics*, Orlando, Florida, USA, pp. 150–155.

Jang, J.-S. R., Lee, H.-R., and Yeh, C.-H. (Oct., 2001), "Query by tapping: a new paradigm for content-based music retrieval from acoustic input," *2nd IEEE Pacific-Rim Conf. on Multimedia*, Beijing, China, pp. 590–597.

Kataoka, M., Kinouchi, M. and Hagiwara, M. (Oct., 1998), "Music information retrieval system using complex-valued recurrent neural networks," *IEEE Int. Conf. on Systems, Man and Cybernetics*, San Diego, CA, USA, pp. 4290–4295.

Kato, T., Kurita, T., and Shimogaki, H. (1991), "Intelligent visual interaction with image database systems –Toward the Multimedia Personal Interface–," *J. of Information Processing*, vol.14, no.2, pp. 134–143.

Kageyama, T., Mochizuki, K., and Takashima, Y. (1993), "Melody retrieval with humming," *Int. Computer Music Conf.*, pp. 349–351.

Motro, A. (1988), "VAGUE: A user interface to relational databases that permits vague queries," *ACM Trans. on Office Information Systems*, vol.6, no.3, pp. 187–214.

Noda, T., Zhao, D. and Takagi, H. (Oct. 2000), "Music database retrieval and media conversion system based on impression," *6th Int. Conf. on Soft Computing*, Iizuka, Fukuoka, Japan, pp. 144–150.

Sonoda, T., Goto, M., and Muraoka, Y. (1998), "A WWW-based melody retrieval system," *Int. Computer Music Conf.*, pp. 349–352.

Takagi, H., Cho, S.-B., and Noda, T. (Aug., 1999), "Evaluation of an IGA-based image retrieval system using wavelet coefficients," *IEEE Int. Conf. on Fuzzy Systems*, Seoul, Korea, vol.3, pp. 1775–1780.

Takagi, H. (2001), "Interactive evolutionary computation: fusion of the capacities of EC optimization and human evaluation," *Proceedings of the IEEE*, vol.89, no.9, pp. 1275–1296.

Tsien, Y.-H. (Feb.–Mar. 2000), "Music indexing and retrieval for digital music libraries," *5th Joint Conf. on Information Sciences*, Atlantic City, NJ, USA, vol.2, pp. 533–536.

Yoshida, K., Kato, T., and Yanaru, T. (Oct., 1998), "Image retrieval system based on subjective interpretation," *Int. Conf. on Soft Computing and Information/Intelligent Systems*, Iizuka, Fukuoka, Japan, pp. 247–250.