

Evaluation of an IGA-based Image Retrieval System Using Wavelet Coefficients

Hideyuki Takagi* Sung-Bae Cho** Toshihiko Noda*

* Kyushu Institute of Design, Dept. of Acoustic Design

Shiobaru, Minami-ku, Fukuoka 815-8540, Japan, E-mail: takagi@kyushu-id.ac.jp

** Yonsei University, Dept. of Computer Science

Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea, E-mail: sbcho@csai.yonsei.ac.kr

Abstract— In this paper, we evaluate the performance of an interactive genetic algorithm (IGA)-based image retrieval system with a subjective test. First, the IGA-based system that retrieves images based on wavelet analysis is introduced. Second, a psychological scale space is constructed to quantitatively express mental images to handle subjective retrieval. Third, the IGA-based system is compared with a random search-based system using the psychological space and a subjective test. The two sign tests and a statistical test have shown that the IGA-based system is significantly quicker in image retrieval than a random search-based image retrieval system.

Keywords: *subjective test, image retrieval, psychological space, interactive genetic algorithm, wavelet transform*

1 INTRODUCTION

Computer users have requested the ability to retrieve information from multimedia databases as the computer power increases. In particular, users are interested in methods that allow them to retrieve information based on stored contents rather than keywords. Such systems include the QBIC system of IBM [1, 7], the QVE system of NEC [2], chabot of UC Berkeley [8], photobook of MIT [11, 10, 9], and Image Surfer of Interpix Software [3]. Since these systems allow users to develop image search criteria from queries, they perform better than conventional methods based on keyword matching.

However, most of these systems are based on engineering approaches which have little relevance to human preference and emotion.

To solve this problem, we have proposed an image database retrieval system based on interactive genetic algorithms (GA) that performs content-based

image retrieval using human preferences [5, 6].

Interactive GA, interactive evolutionary computation (EC) in general, measures fitness values based on the user's subjective evaluation. For the past 10 years, the number of papers on the interactive EC have increased, and its application fields have expanded into the artistic, engineering, and educational & edutainment fields [13].

In the IGA-based image retrieval system, a human evaluates each individual image within a population, and the GA determines the next retrieved images as individuals in next generation based on the evaluation, which the system might incorporate human preference into the process of image retrieval. Our IGA-based system uses wavelet transform to extract features from images.

The objective of this paper is to evaluate our IGA-based image database retrieval system using a subjective test. We overview our system in section 2, make a psychological space for the subjective test in section 3, and conduct the test in section 4.

2 IMAGE DATABASE RETRIEVAL BY INTERACTIVE EC

GA is a model of machine learning based on evolution in nature and is frequently used for optimization problems. Usually, a fitness function is used as a cue to optimize the given tasks.

Unlike normal GA, interactive GA adopts a user's evaluation as fitness, which is useful when a fitness function cannot be exactly determined. This property allows a system to be developed based on human preference or emotion. Therefore, our IGA-based image retrieval system obtains fitness values for images from a user, which are used to select individuals for the next generation.

2.1 Chromosome Representation

Wavelet coefficients are obtained by decomposing an image using wavelet transform. Through the above procedure an $r \times r$ matrix, T , contains the average

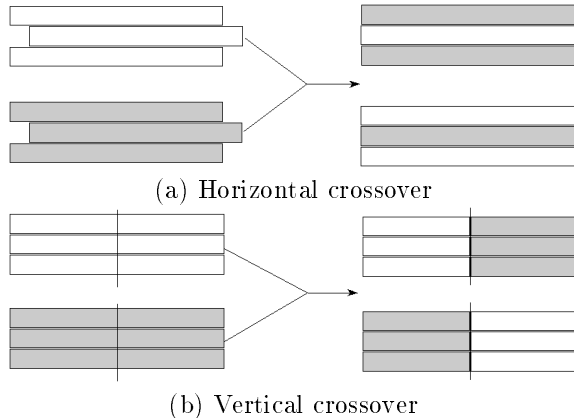


Figure 1: Horizontal and vertical crossover operations.

color of the image in entry $T[0, 0]$ and wavelet coefficients in the remaining entries of T . We can reconstruct original image without loss using this information, but because there is no need to maintain all the information for searching, we extract the largest 50 coefficients in RGB channels and construct a chromosome in a 3×50 array. Previous work showed that storing the 40~50 largest-magnitude coefficients in each color works best and truncating the coefficients appears to improve the discrimination power of the metric [4]. Therefore, we only store the sign information of coefficient values of the chromosome.

2.2 Genetic Operators

The size of the population is 12, and the fitness values for shape and color are obtained from a user. The selection strategy is governed by the expected frequency of each individual, and one point horizontal and vertical crossover operators are used [5]. Mutation is not considered because its effects would be disproportionately large for the small population. Figure 1 shows the schematic diagrams of horizontal and vertical crossover operations.

2.3 Implementation

The entire system is constructed as shown in Figure 2. In the preprocessing step, we perform wavelet transform to every image in the database and store the overall average color and the indices and signs of the m magnitude wavelet coefficients in a table.

To search images, we use the interactive GA. Our IGA-based image retrieval system displays 12 images, obtains fitness values for them from a user, and selects candidates based on the fitness. Either horizontal or vertical crossover is applied to the selected candidates. To find the 12 best images in the current generation, the stored image information is evaluated using each criterion. Twelve images of the

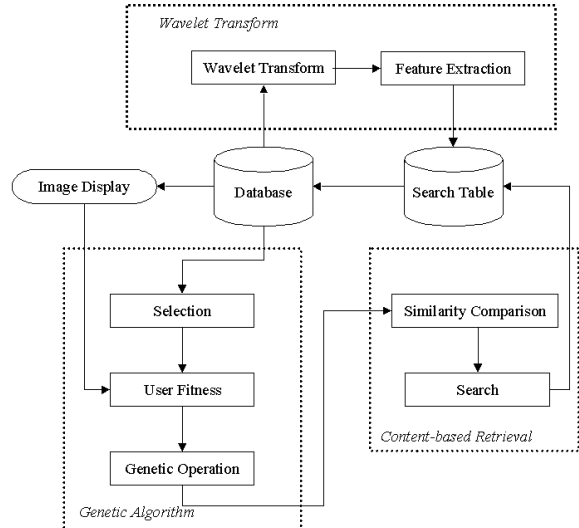


Figure 2: System structure.

higher magnitude value are provided as a result of the search. The similarity between the potential target image and a candidate image is calculated by the following equation:

$$\|Q, T\| = w_{0,0} |Q[0, 0] - T[0, 0]| + \sum_{i,j} w_{i,j} |Q[i, j] - T[i, j]|$$

where $Q[i, j]$ and $T[i, j]$ represent single color channels of wavelet decomposition of the candidate and target images, and $Q[0, 0]$ and $T[0, 0]$ are the overall average intensities of those color channels. The system repeats this process to search for new candidates until the user finds the image that he or she wants.

The system was developed in Microsoft Visual C++, and runs on Pentium PC. A searching table was constructed by a batch job over 256×256 JPEG images.

3 CONSTRUCTION OF PSYCHOLOGICAL SPACE FOR IMAGE DATABASE RETRIEVAL

To evaluate image retrieval systems, it is necessary to prepare several different motifs. We let human operators retrieve images matching the given motifs using the systems, and evaluate how close the retrieved images are to the given motifs and how quickly they are obtained. It is important to prepare unbiased image motifs for universal evaluation. Since the *bias/unbias* in image retrieval is a psychological issue, we constructed a psychological space of images and determined the motifs that were widely distributed in the space for image retrieval evaluation. We cannot judge if the retrieval experiment is unbiased without the psychological space.

Table 1: 14 pairs of adjectives used to construct a psychological scale space for image. Original words used in the experiment in section 3 are 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>
emotionally attractive	—	emotionally unattractive
<i>perspectively wide</i>	—	<i>perspectively narrow</i>
<i>dry</i>	—	<i>wet</i>

In this section, we examine the kinds of psychological scales used as humans watch images and construct a psychological scale space. The scale construction procedure includes the careful selection of adjective pairs, SD (semantic differential) method, and principal component analysis.

We first considered 151 adjectives. Then, we eliminated the adjectives whose semantics may depend on a person’s interpretation or those that are similar. Finally, we extracted the 14 pairs of adjectives listed in Table 1.

In our experiment using the SD method, subjects were required to evaluate the images using 14 pairs of adjectives varied in 7 scale values. For example for the *dry—wet* scale, a subject would have the scale of (*very wet, wet, a little wet, normal, a little dry, dry, very dry*) and assign his or her impression of given images on the scale. We used 610 images selected from the 3,000 images of our preliminary experiment in the SD method. The subjects were 10 undergraduate and graduate students in their 20s.

We applied the principle component analysis to the obtained $14 \times 610 \times 10$ data. As a result, it was determined that the original 14 dimensional adjective space could be approximated by 3, 4, 5, 6, and 7 factors with the coverage of 58.9%, 66.5%, 72.3%, 77.3%, and 82.0%, respectively. When we use these approximated factor scales to retrieve images, the dimension of the final factor space should be constructed by considering the coverage of original space and the user’s capacity to evaluate images and input the evaluation values into a retrieval system.

The purpose of this experiment was not to develop the final image retrieval system but to select some retrieval motifs for the evaluation of the IGA-based

Table 2: 14 pairs of adjectives in 3 factors. Original words used in the experiment in section 3 are in Japanese.

<i>1st factor</i>	<i>2nd factor</i>	<i>3rd factor</i>
<i>pure</i>	<i>passionate</i>	emotionally attractive
<i>simple</i>	<i>gaudy</i>	<i>warm</i>
<i>jaunty</i>	<i>vivid</i>	<i>soft</i>
<i>dry</i>	<i>clear</i>	<i>comical</i>
<i>bright</i>		
perspectively wide		

image retrieval system. Experimentally, we approximated the original 14 dimensional adjective space with 3 factors. Table 2 shows the 14 original pairs of adjectives in the 3 factors. Although only one adjective from each pair is described in the table, each adjective implies an adjective pair. For example, *pure* in the first factor means the *pure—impure* axis.

The first, second, and third factors may be able to be summarized as the factors of *pure or simplicity, vibrancy with life, and emotion*, respectively.

We determined the image retrieval motifs used in section 4 by the following three adjective pairs as representatives of three factors:

1st axis: *simple — dense*
 2nd axis: *passionate — mild*
 3rd axis: *round — rugged*

Then, we chose four noncontiguous subspaces in the eight subspaces separated by the three axes and used them as the retrieval motifs in the experiment in section 4. They are:

motif #1: *simple, passionate, and round* image
 motif #2: *simple, mild, and rugged* image
 motif #3: *dense, passionate, and rugged* image
 motif #4: *dense, mild, and round* image

The “four noncontiguous subspaces” in a three-factor space is illustrated in Figure 3.

4 SUBJECTIVE TEST FOR IMAGE DATABASE RETRIEVAL SYSTEM

In this section, we evaluate the retrieval performance of both IGA-based and random search-based image retrieval systems. The evaluation experiment includes two phases. The first retrieves images that match the given retrieval motifs, and the second compares the retrieved images to those in the first phase using the paired-comparison subjective test.

In the image retrieval experiment of the first phase, each subject is given two of the four motifs

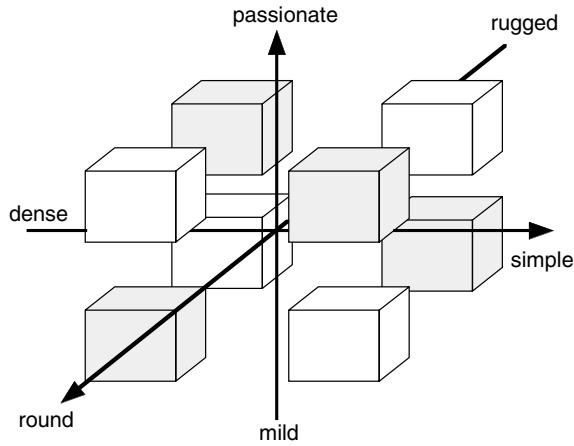


Figure 3: Retrieval motifs in a 3-D psychological space used in the experiment in section 4.

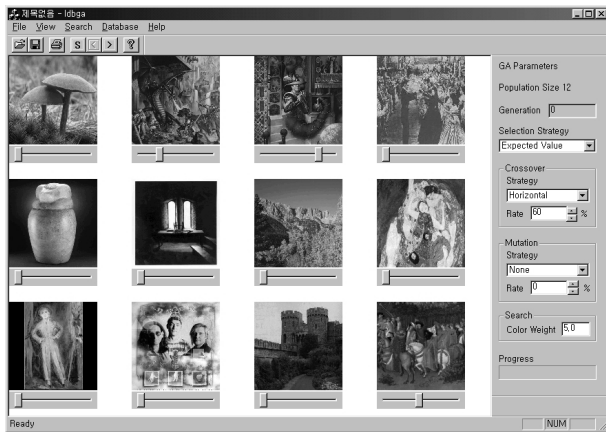


Figure 4: User interface of image retrieval experiment in the first phase.

determined in section 3 and retrieves images that he or she feels their mental image will best match the given motifs. Figure 4 shows the user interface implemented in this experiment. The subjects are requested to retrieve images for 13 generations using the IGA-based and random search-based image retrieval systems. The reason we did not give all four motifs to each subject is that we are afraid of decreasing the reliability of our subjective test due to subject fatigue. Pairs of a retrieved image and fitness value given by each subject in each generation are saved for the subjective test in the second phase.

In the subjective test of the second phase, image-pairs obtained by the two systems in the first phase are compared, and the performance of our IGA-based image retrieval system is compared to the random search-based system using statistical tests. The

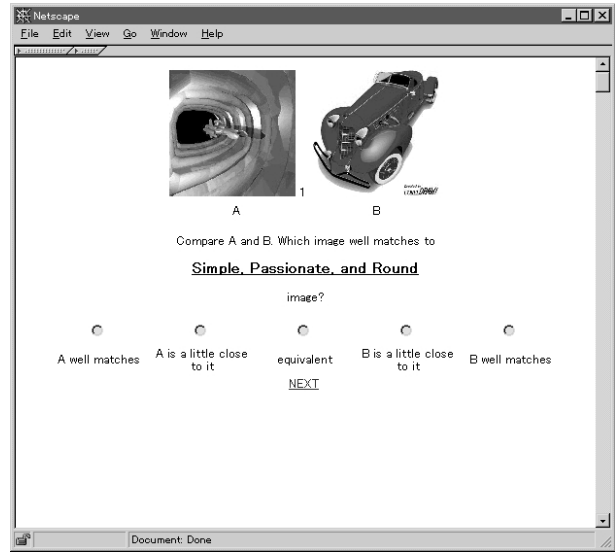


Figure 5: User interface of the comparison of two images and rating in five scale used in the second phase. Letters in this window are an English translation of original Japanese version actually used in our experiment.

first step is to choose the image with the highest fitness in each generation of the saved data. Then, the two best images of the two retrieval systems in same generation are displayed at once and compared in five scales by each subject (see Figure 5.) The subjects do not know which system retrieved which image. The best image pairs in 2–13 generations are compared. Since the images in the first generation are based on random initialization of GA, they are not compared. The 12 comparison data are collected and statistically tested using two sign tests.

The images that each subject evaluates in the second phase are those that the same subject retrieved in the first phrase. Since we constructed a psychological scale space of images using the total evaluation of 10 subjects in section 3, it can be considered that this evaluation experiment based on the three factors of the space is commonly applicable to multiple subjects, so that it would not be necessary to limit the evaluation to the image made by subject himself or herself. However, we decided to adopt the strict condition of using the same psychological scale for both image retrieval and evaluation experiments to avoid the possibility that the psychological scale may depend on the person.

The 12 subjects were undergraduate and graduate students ranging from 22 to 34 years in age and included 11 Japanese and 1 Chinese. All subjects participated in the experiment of constructing a factor space in section 3. An examiner explained the

meaning of the retrieval motifs to subjects using the adjective list used in the experiment in section 3.

The interfaces of both IGA-based and random search-based image retrieval systems are same and shown in Figure 4. The IGA-based system took slightly longer for the retrieved images to be displayed on a window due to the calculation time of the GA operation, although the time difference is just on the order of *msec*. This time difference might be to the advantage of the random search-based system psychologically. The experiment in the first phase requests the retrieval of images with two motifs. To avoid the psychological interference of the two motifs, the second retrieval experiment followed the first after half to two days. The database used in this experiment includes 2,000 images.

5 EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Convergence performance by generation

Table 3 shows summed up data in five ranks by generation. (-2, -1, 0, 1, 2) in the table means (*R well matches to give motif, R is a little close to it, R and I are equivalent, I is a little close to it, I well matches to it*), where R and I are a retrieved image by the random search-based and IGA-based systems, respectively. As the 12 subjects evaluated the two retrieval motifs, the total number of evaluations was 24. The positive (or negative) ranks mean that the IGA-based image retrieval system (or the random search-based system) is superior to the other. Note that the *positive* and *negative* are used here as numerical signs only and do not have any value such as good or bad.

The result of the sign test and the Wilcoxon sign-ranks test [12]⁰ are also shown in Table 3. The sign test checks the significance of the difference between the sum of (“+1” and “+2”) and that of (“-1” and“-2”) in Table 3. The sign-ranks test checks the significance of the difference when taking account of the evaluation rank. Basically the sign-ranks test is applicable when the evaluation psychological scales of all subjects were assumed to be similar.

Table 3 clearly shows that the IGA-based image retrieval system searches out desire images more quickly than the random search-based system. This faster search performance has been statistically shown in the total data of 12 generations, former half generations, i.e. the sum of 2–7th generations, and latter half generations, i.e. the sum of 8–13th generations with ($p < 0.01$).

⁰The table of Wilcoxon sign-ranks test in this book includes a few incorrect data, so we used a table in another book.

Table 3: Results of a subjective test and statistical tests by generation. Five ranks means that IGA based image retrieval system or random search based image retrieval system retrieves specified image faster than another when positive or negative ranks is chosen. (See five ranks in the first paragraph in section 5.1.) **, *, and Δ in two sign tests mean that the difference between retrieval performances of the two systems is significant with risk rate of 1% ($p < 0.01$), 5% ($p < 0.05$), and 10%($p < 0.1$).

ranks	generation #						
	2	3	4	5	6	7	
2	3	2	8	7	7	7	
1	9	15	4	6	3	10	
0	3	4	2	3	5	4	
-1	6	1	10	7	5	2	
-2	3	2	0	1	4	1	
sign test		**				**	
sign-ranks test		*				**	
	generation #						total
	8	9	10	11	12	13	
2	8	7	7	3	8	6	36
1	8	7	10	11	7	6	48
0	3	2	3	6	3	4	21
-1	3	5	2	1	5	5	30
-2	2	3	2	3	1	3	9
sign test	*		**	*			**
sign-ranks test	*		*		*		**

5.2 Convergence performance by retrieval motif

Table 4 shows the sign tests’ results by retrieval motif. They have shown that the IGA-based system is superior to the random search-based system with ($p < 0.1$) for the third retrieval motif, i.e. *dense, passionate, and rugged* image, and with ($p < 0.01$) for other retrieval motifs. Although the performance depends on the retrieval motifs, this experimental result, as well as the result in section 5.1, shows the superiority of the IGA-based system is significant. Some subjects reported that it had been difficult to draw mental images of the third motif. This difficulty of imagining mental impressions might result in the difference.

In conclusion, it was shown that the IGA-based system can retrieve images belonging to the widely distributed and separated four spaces in a psychological space shown in Figure 3 quicker than the random search-based system.

6 CONCLUSION

In this paper, we evaluated the performance of the IGA-based image retrieval system that uses wavelet

Table 4: Results of a subjective test and statistical tests by retrieval motif. See the caption of Table 3 for five ranks and statistical test marks. Four retrieval motifs are listed in section 3

ranks	motif #				total #
	1	2	3	4	
2	23	19	17	14	73
1	21	23	21	31	96
0	11	10	11	10	42
-1	12	9	19	12	52
-2	5	11	4	5	25
sign test	**	**	Δ	*	**
sign-ranks test	**	**	**	**	**

coefficients to represent physical features of images. The subjective tests and a statistical test have shown that the convergence speed of the IGA-based system that retrieves an actual image being similar to a mental image from an image database is significantly faster than that of the random search-based system.

We constructed the psychological scale space to quantitatively handle the image impressions. By using this psychological space, we evaluated the usefulness of physical image features for different types of images. The *type* in this case means the psychological one, which means that we can quantitatively handle mental images; it does not mean the physical images themselves. Since the evaluation without the psychological space cannot express mental image impressions quantitatively, it is fundamentally difficult to evaluate retrieval systems based on human preference, i.e. IGA-based database retrieval system in detail. This point is another fruit of our research.

From the point of the factor axes in the psychological space, we are going to analyze the usefulness of wavelet coefficients for image retrieval. As a result, if it is found that the wavelet coefficients are not enough to retrieve images located in certain parts of the psychological space, we will also investigate supplementary image features using this method. After all, we aim to determine the fewest and most effective image features, which are not restricted to wavelet coefficients, for image retrieval using this method, besides practically using the image retrieval system.

REFERENCES

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, et al. "Query by image and video content: the QBIC system," *Computer*, vol.28, no.9, pp.23-32 (Sept., 1995).

[2] K. Hirata and T. Kato, "Rough sketch-based image information retrieval," *AIEC Research and Development*, vol.34, no.2, pp.263-73 (April, 1993).

[3] Interpix homepage: <http://www.interpix.com/>

[4] C. E. Jacobs, A. Findkelstein, and D. H. Salesin, "Fast multiresolution image querying," *Proc. of SIGGRAGH95*, pp.277-286, Los Angeles, CA, USA, Edited by: R. Cook, New York, NY, USA: ACM (Aug., 1995).

[5] J.-Y. Lee and S.-B. Cho, "Interactive genetic algorithm for content-based image retrieval," *Asia Fuzzy System Symposium*, Masan, Korea (June, 1998).

[6] J.-Y. Lee and S.-B. Cho, "Interactive genetic algorithm with wavelet coefficients for emotional image retrieval," *5th Int'l Conf. on Soft Computing and Information/Intelligent systems (IIZUKA'98)*, pp.829-832, Iizuka, Fukuoka, Japan, World Scientific (Oct., 1998).

[7] W. Niblack, R. Barber, W. Equitz, M. Flickner, et al., "The QBIC project: querying images by content using color, texture, and shape," *Proc. of the SPIE - The Int'l Society for Optical Engineering, Storage and Retrieval for Image and Video Databases*, San Jose, CA, USA, pp.173-187 (Feb., 1993).

[8] V. E. Ogle and M. Stonebraker, "Chabot: retrieval from a relational database of images," *Computer*, vol.28, no.9, pp.40-48 (1995).

[9] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: content-based manipulation of image databases," *Int'l J. of Computer Vision*, vol.18, no.3, pp.233-254 (1996).

[10] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: tools for content-based manipulation of image databases," *23rd AIRP Workshop. Image and Information System: Applications and Opportunities*, Washington, DC, USA, (Oct. 1994). *Proc. of the SPIE - The Int'l Society for Optical Engineering*, vol.2368, pp.37-50 (1995).

[11] A. Pentland, R. W. Picard, and S. Sclaroff, *Photobook: tools for content-based manipulation of image databases. Storage and Retrieval for Image and Video Databases II*, San Jose, CA, USA, (Feb., 1994). *Proc. of the SPIE - The Int'l Society for Optical Engineering*, vol.2185, pp.34-47 (1994).

[12] S. Siegel and N. J. Jr. Castellan, "Nonparametric statistics for the behavioral science," 2nd ed., McGraw-Hill Book Company (1988).

[13] H. Takagi, "Interactive Evolutionary Computation, - Cooperation of computational intelligence and human *KANSEI* -," *5th Int'l Conf. on Soft Computing and Information/Intelligent systems (IIZUKA'98)*, pp.41-50, Iizuka, Fukuoka, Japan, World Scientific (Oct., 1998).