Multimed Tools Appl
DOI 10.1007/s11042-007-0109-8

Video scene retrieval with interactive genetic algorithm

Hun-Woo Yoo · Sung-Bae Cho

Abstract This paper proposes a video scene retrieval algorithm based on emotion. First, abrupt/gradual shot boundaries are detected in the video clip of representing a specific story. Then, five video features such as "average color histogram," "average brightness," "average edge histogram," "average shot duration," and "gradual change rate" are extracted from each of the videos, and mapping through an interactive genetic algorithm is conducted between these features and the emotional space that a user has in mind. After the proposed algorithm selects the videos that contain the corresponding emotion from the initial population of videos, the feature vectors from them are regarded as chromosomes, and a genetic crossover is applied to those feature vectors. Next, new chromosomes after crossover and feature vectors in the database videos are compared based on a similarity function to obtain the most similar videos as solutions of the next generation. By iterating this process, a new population of videos that a user has in mind are retrieved. In order to show the validity of the proposed method, six example categories of "action," "excitement," "suspense," "quietness," "relaxation," and "happiness" are used as emotions for experiments. This method of retrieval shows 70% of effectiveness on the average over 300 commercial videos.

Keywords Emotion-based Retrieval · Video scene retrieval · Commercial video retrieval · Interactive Genetic Algorithm (IGA)

1 Introduction

A variety of data available in the image and video format are being generated, stored, transmitted, analyzed, and accessed with advances in computer technology and communication network. To make use of these data, an efficient and effective technique needs to be developed for the retrieval of multimedia information. This paper proposes an emotion-based video scene retrieval algorithm dealing with the retrieval of video based on their contents.

H.-W. Yoo (✉)
Center for Cognitive Science, Yonsei University, 134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-749, Korea
e-mail: paulyhw@yonsei.ac.kr

S.-B. Cho
Department of Computer Science, Yonsei University, 134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-749, Korea
e-mail: sbcho@cs.yonsei.ac.kr

In the earlier days, most related systems used a text-based approach [16, 22]. However, because of the burden of manual indexing and the subjectivity of data contents, the content-based retrieval method that indexes data automatically by a computer became popular. This approach retrieves data based on color, texture, shape, or spatial relationship of objects within the image/video [1, 5, 11, 18, 20, 24, 35, 36]. For images, techniques of image processing, pattern recognition, and computer vision are frequently employed for that purpose. For videos, techniques of shot boundary detection, key frame extraction and clustering, and feature extraction were generally used [12, 15, 29, 33, 37–39].

Contrary to the expectation of good results, however, the low-level information such as color, texture, shape, etc did not capture sufficiently semantic information that people had in mind. For example, when a user want images/videos with blue sky, the use of color features would retrieve blue sea as well as blue sky. Furthermore, it is hard to differentiate a plane from a bird in the sky through current computer vision technology. Hence, it is important how to extract semantic information and employ it for retrieval [10, 19, 23, 31, 32]. Object segmentation and learning [19], classification based on Bayes' theorem [31, 32], and relevance feedback [10, 23] have been frequently exploited for that purpose.

Representation of emotions with visual medium is very important. Many posters and movie previews are designed to appeal to the potential customer by containing specific moods or emotions. Hence, emotion-based retrieval would be one of the essential applications in the near future. Yet the related studies associated with the computer vision are still in infancy.

Several researchers have studied retrieval methods based on emotion [7–9, 27, 30]. In [7], based on wavelet coefficients, gloomy images were retrieved through feedbacks called interactive genetic algorithm (IGA). This method, however, is limited to differentiate only two categories, gloomy images or not. In a similar approach, Takagi et al. [27] designed psychology space (or factor space) that captures human emotions and mapped those onto physical features extracted from images. Also, IGA was applied for emotion-based image retrieval. Based on Soen's psychological evaluation of color patterns [25], Um et al. [30] proposed an emotional model to define a relationship between physical values of color image patterns and emotions. It extracted color, gray, and texture information from an image and input into the model. Then, the model returned the degree of strength with respect to each of 13 emotions. It, however, has a generalization problem because it experimented on only five images and could not be applied to the image retrieval directly since, for image retrieval, an emotion keyword should be presented first as a query, and then images with the presence of the associated emotion are retrieved (Um's model has a reverse procedure). Moreover, the above methods [7, 27, 30] deal with only images, not with videos that would be handled here in this paper.

Based on the color theory of Itten [14], Colombo et al. [8] retrieved art paintings from Renaissance and contemporary eras by mapping expressive and perceptual features onto emotions. It segmented the image into homogeneous regions, extracted features such as color, warmth, hue, luminance, saturation, position, and size from each region, and used its contrast and harmony relationship to other region for capturing emotions. However, it was designed for art painting retrieval only. Authors also proposed emotion-based retrieval method for video case. Based on the semiotic category, commercials are classified into four emotions of utopic, critical, practical, and playful [8, 9].

Audio-based retrieval with relation to emotions is also found in several studies [6, 21, 28]. In [28], audio information in the form of prosody features is extracted to describe the emotional content of audio files. In [21], music retrieval is performed by so-called query by humming approach. Most recently, in [6], audio information as multi-modal approach is surveyed for video indexing purposes.

So far, the researches on multimedia retrieval in terms of emotion are very few (in particular, for video case). In this paper, we propose a new emotion-based video retrieval method that uses emotion information extracted from a video scene. The proposed method retrieves video scenes that a user has in mind through the evolutionary computation called *interactive genetic algorithm*. In the preprocessing step, for each database video, video shots are detected and a key frame is selected from each shot, and from the key frames among all the shots, the features that describe emotional content are extracted such as the average color histogram, average brightness, average edge histogram, average shot duration, and the gradual change rate. These features are encoded as a chromosome and indexed into the database. For all the database videos, the above procedure is repeated. For retrieval, an interactive genetic algorithm is exploited that performs optimization with the human evaluation. The system displays 15 videos, obtains a relevance feedback of the videos from a human, and selects the candidates based on the relevance. A genetic crossover operator is applied to the selected candidates. To find the next 15 videos, the stored video information is evaluated by each criterion. Fifteen videos that have higher similarity to the candidates are provided as a result of the retrieval.

The IGA is different from the well-known genetic algorithm (GA) in that a human being is involved as a fitness function. Since retrieving videos that a user has in mind depend on personal preference, IGA is a technique suitable to emotion-based video retrieval. Moreover, the proposed method is not limited to specific emotions (utopic, critical, practical, and playful) of Colombo et al. [8, 9], providing a more general retrieval mechanism through IGA. The overall retrieval procedure is described in Fig. 1.

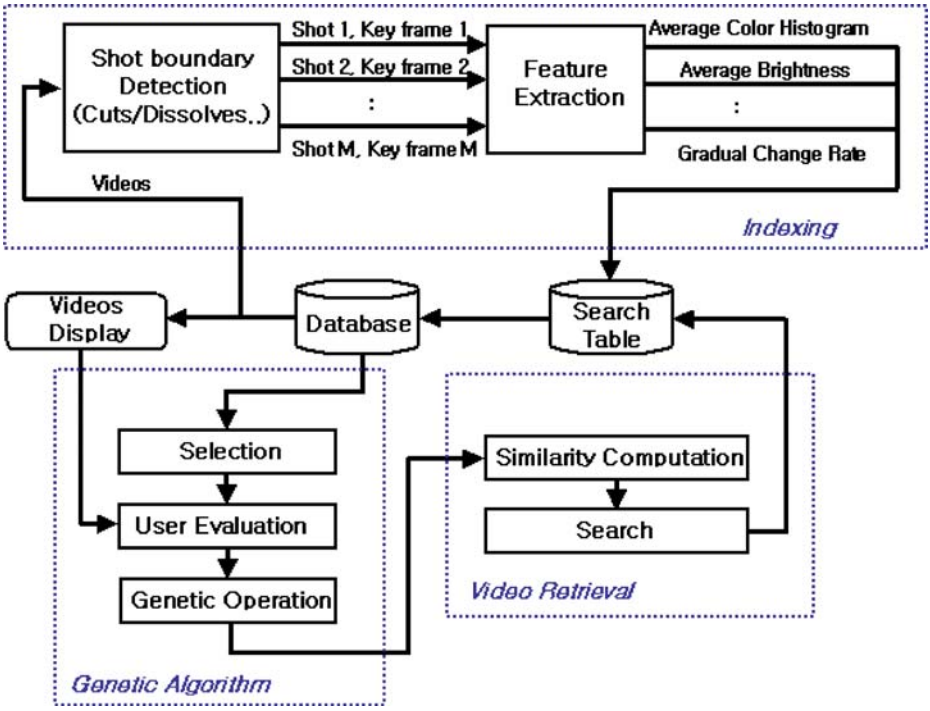

Fig. 1 The proposed method

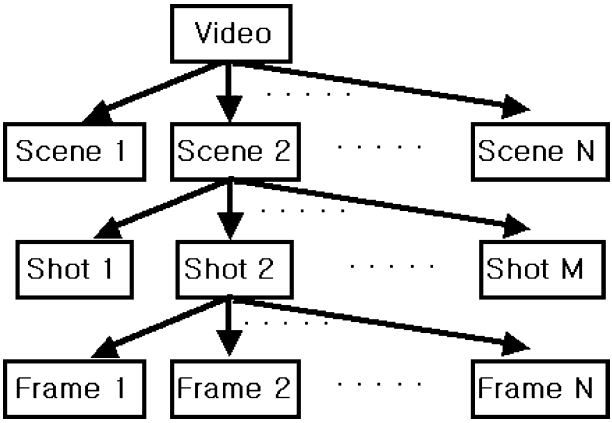

Fig. 2 Hierarchical structure of video

2 Shot boundary detection

A video consists of many still frames, which are compressed by spatial–temporal information. For the emotion-based scene retrieval, we have to segment a video into shots, and choose key frames and extract meaningful features from the shots, which capture a specific emotion (mood) of the video. As in Fig. 2 and Table 1, a video is constructed hierarchically in the form of video/scene/shot/(key) frame from top to bottom-levels.

In general, shot boundaries are detected for two cases: the abrupt transition called cut and the gradual transitions called fade-ins, fade-outs, or dissolves. For shot boundary detection, we use the correlation coefficient of gray-level histogram and edge histogram of our earlier investigation [34].

2.1 Correlation sequence using gray-level histogram

Let m_{lk} and σ_{lk} be the average and variance of the gray-level information in the kth frame. The gray-level inter-frame correlation between two consecutive frames k and $(k+1)$ is obtained as follows:

$$S_{\mathrm{LIFC}}(k,k+1)=\frac{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left(X_{lk}[i][j]-m_{lk}\right)\left(X_{l(k+1)}[i][j]-m_{l(k+1)}\right)}{\sigma_{lk}\sigma_{l(k+1)}}\quad -1\le S_{\mathrm{LIFC}}(k,k+1)\le 1 \tag{1}$$

Table 1 Terminology of video

Terminology	Meaning
Shot	A sequence of frames, which is generated during a continuous camera operation (that is, between record and stop)
Key frame	Frame that contains the most important information in the shot. According to the complexity, one or two frames are chosen from the shot. In this paper, the first frame of the shot is selected as key frame
Scene	Assembly of neighboring shots that have semantically common events or story.

where W and H are the width and height of a frame, and $X_{lk}[i][j]$ is the gray-level of (i, j) coordinate in the kth frame.

2.2 Correlation sequence using edge histogram

The use of gray-level information only for detection purpose would lead to false detections when abundant variations of gray-levels occur in the similar frames. Thus, the edge information from a frame is also used. First, Canny edge detector [35] is applied to a frame. Then, the number of edges is acquired in each 3×3 uniform rectangular region of the frame. Take the number of edges in each region as histogram bin values in that frame. Let m_{ek} and σ_{ek} be the average and variance of the number of edges in the kth frame. The edge inter-frame correlation between two consecutive frames, k and $(k+1)$, is obtained as follows:

$$S_{\mathrm{EIFC}}(k,k+1)=\frac{\sum_{i=0}^{8}\left(X_{ek}[i]-m_{ek}\right)\left(X_{e(k+1)}[i]-m_{e(k+1)}\right)}{\sigma_{ek}\sigma_{e(k+1)}}\quad -1\le S_{\mathrm{EIFC}}(k,k+1)\le 1 \tag{2}$$

where $X_{ek}[i]$ is the number of edges of the ith region in the kth frame.

2.3 Integrated correlation

Finally, two correlations are integrated for detection purpose.

$$S_{\mathrm{IFC}}(k,k+1)=w_l S_{\mathrm{LIFC}}(k,k+1)+w_e S_{\mathrm{EIFC}}(k,k+1) \tag{3}$$

where w_1 and w_e are the weights for relative importance between the use of gray-levels and edges, and are subject to $w_1+w_e=1$.

2.4 Abrupt and gradual detection

A basic detection mechanism is similar to the research [15]. We consider a current frame as an abrupt shot boundary frame when it is sufficiently different from a previous frame, and a current frame is considered as a gradual shot boundary frame when it is sufficiently different from a previous shot boundary frame. The criterion of deciding the sufficient difference is based on the inter-frame correlation. This scheme is described in Fig. 3. The algorithm uses two thresholds for detection purposes (threshold t_1 for abrupt boundary and t_2 for gradual boundary).

Detection processes are as follows: The correlation between the previous frame (p) and the current frame (c) is computed using (3), and compared with the predetermined threshold t_1 for abrupt detection. If this correlation is lower than t_1, an abrupt shot change is declared. If not, the current frame (c) is compared with the previous shot boundary frame (b) (if previous shot boundary frame does not exist, the first frame of video is considered as a default shot boundary frame). If the correlation between (c) and (b) is lower than a predetermined threshold t_2, a gradual shot change is declared. Otherwise, the shot boundary is not declared, and the current frame (c) is set to the previous frame (p). Then the above process with the next frame is repeated until the end of the video frames.

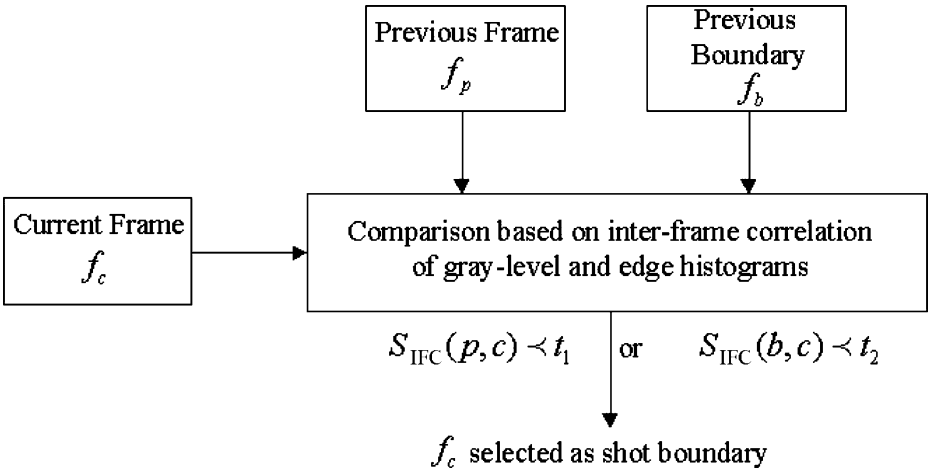

Fig. 3 Diagram of shot boundary detection

It requires so much computational costs that the above procedure is applied to every consecutive frame. To reduce the cost, we have reduced temporal resolution by sampling every 15 frames instead of every frame. Since videos are used to have temporal redundancies, it is reasonable to reduce the temporal resolution for detection purpose.

3 Feature extraction

To retrieve video scenes with the presence of specific emotion, the features for representing scene content effectively should be extracted. For that purpose, the shot boundaries in the scene are detected, the key frames in each of shots are chosen, and the features from each key frame are extracted to represent the scene content. In this research, visual content such as color, gray intensity, and edge information and visual rhythm such as shot duration in time and the ratio of the number of gradual boundaries among total number of shot boundaries are used for describing features.

To represent emotional content of a shot, feature extraction is performed from a single key frame rather than all the frames in a shot, because in our experiments two methods did not yield a prominent difference for the system performance. Furthermore, feature extraction using the information of all the frames could cause high computational overhead.

3.1 Average color histogram

Traditionally, color is one of the basic attributes to represent video content. It has been successfully used in content-based image/video retrieval for a long time. This attribute could be used effectively to describe a specific emotion. For example, the images with "warm" sensation tend to have much of red colors. The images with "cold" sensation tend to have much of blue or cyan colors. Also, the "action" videos contain red or purple colors and the videos with "quietness" sensation contain blue, green, and white colors [8].

In our method, for describing the color in a shot, we first extract an RGB joint histogram from a key frame and multiply the value that each RGB joint histogram bin has by the

number of frames within the shot. Then, by applying the above procedure to every shot within a scene and dividing it by the total number of frames, we obtain the average color histogram in the scene as follows.

$$\mathrm{Avg}H_{\mathrm{RGB}}[i] = \frac{\sum_k H^k_{\mathrm{RGB}}[i] \times \mathrm{ShotLength}[k]}{N_T}, i = 0, 1, 2, ...26 \tag{4}$$

where $H_{\mathrm{RGB}}[i]$ is the i^{th} bin value of 27-dimensional joint RGB histogram (each of R, G, and B channels is equally quantized, i.e. $3 \times 3 \times 3$) in the key frame of the k^{th} shot, ShotLength[k] is the length of the k^{th} shot, which is denoted in the number of frames, and N_{T} is the total number of frames within the scene.

3.2 Average brightness

In general, people feel light and happy from bright images. Contrarily, people tend to feel hard, heavy, and gloomy from dark images. In videos, the scenes with "quietness" tend to have bright tones. Therefore, we include an average brightness within a scene for describing a feature.

$$\mathrm{AvgBright} = \frac{\sum_k \mathrm{LocalAvgBright}[k] \times \mathrm{ShotLength}[k]}{N_{\mathrm{T}}} \tag{5}$$

where LocalAvgBright[k] is the average brightness of a key frame in the k^{th} shot, ShotLength[k] is the length of the k^{th} shot, which is denoted in the number of frames, and N_{T} is the total number of frames within the scene.

3.3 Average edge histogram

It is easily noticeable that in the images with "gloomy" sensation the number of edges are relatively few, and vice-versa in the images with "joyful" sensation [7]. We extract dominant edges through Canny edge detector [35] and use the number of edges with respect to 72 edge directions as a feature.

$$\mathrm{Avg}H_{\mathrm{EDGE}}[i] = \frac{\sum_k H^k_{\mathrm{EDGE}}[i] \times \mathrm{ShotLength}[k]}{N_{\mathrm{T}}}, i = 0, 1, 2, ..., 71 \tag{6}$$

where $H_{\mathrm{EDGE}}[i]$ is the ith bin value of 72-dimensional direction edge histogram in the key frame of the k^{th} shot, Shotlength[k] is the length of the k^{th} shot, which is denoted in the number of frames, and N_{T} is the total number of frames within the scene.

3.4 Average shot duration

In general, the scenes such as "action" and "excitement" tend to have short shots to increase tension. On the other hand, "quietness," "relaxation," and "happiness" scenes tend to have

long shots and do not have much change in the shot content. In this paper, we compute shot durations within the scene and sum them up. Then, by dividing the summation by the total number of shots, we obtain an average shot duration as a feature.

$$\text{AvgShotTime} = \frac{\sum_{k} \text{ShotDuration}[k]}{N_S} \tag{7}$$

where ShotDuration*[k]* is the duration time of the k^{th} shot, which is denoted in seconds and N_S is the total number of shots within the scene.

3.5 Gradual change rate

Often, gradual shot boundaries in videos induce specific emotion. Many "quietness" scenes contain gradual shot boundaries such as dissolve. Hence, the ratio of the number of gradual boundaries over the total number of shot boundaries is used.

$$\text{GradRate} = \frac{N_G}{N_S} \tag{8}$$

where N_G and N_S are the total number of gradual shot boundaries and the total number of shots within the scene, respectively.

4 Emotion-based video retrieval

4.1 Interactive genetic algorithm

Genetic algorithm (GA) provides a powerful technique for searching large problem spaces. A GA solves problems by evolving a population of potential solutions to a problem, using standard genetic operations like crossover and mutation, until an acceptable solution emerges [13]. In other words, solutions appropriate only to a certain fitness function survive for every evolving step. However, most of the conventional applications of GA lack of the capability of utilizing human intuition or emotion appropriately. Consider retrieving the most favorite videos from human-machine interaction systems. Such systems must be subjectively evaluated, and it is hard or even impossible to design a fitness function.

Interactive genetic algorithm is a technique that searches possible solution based on human evaluation. A human can obtain what he has in mind through repeated interactions with the method, when the fitness function cannot be explicitly defined. It has been applied to a variety of areas such as graphic art, industrial design, musical composing, voice processing, virtual reality, information retrieval, education, games, etc [2–4, 7, 17, 27]. An extensive survey on IGA can be found in [26].When the videos are presented to a user for emotion-based video scene retrieval, the user has to decide fitness whether each video agrees with what he or she has in mind or not. Therefore, in this paper, we try to achieve video scene retrieval by formulating it to the problem of the interactive genetic algorithm. For that purpose, five features are extracted from a video as in the previous section and encoded as a chromosome. Then, the retrieval is achieved by evolving the chromosomes of initially selected videos until satisfactory videos are obtained.

The method does not focus on the extraction of features that capture certain emotion. Rather, it focuses on performing emotion-based retrieval through human–computer interaction with IGA technique after extracting possible features from videos. If videos

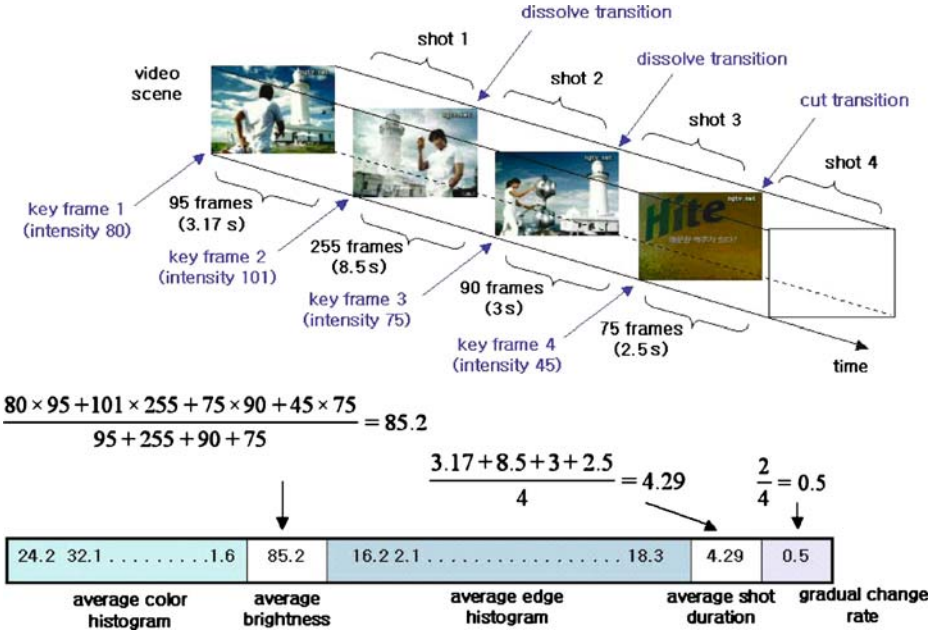

Fig. 4 Chromosome representation from a video

should be retrieved for the fixed query emotions, more sophisticated features can be extracted. Figure 4 shows how a chromosome is encoded from a video.

For scene-based retrieval, videos must be divided into many meaningful scenes, and those scenes must be indexed into database, or short videos composed of one scene (for example, commercial videos such as CF) must be indexed into database. In this paper, we have used the short videos of single scene for experimental convenience.

4.2 Retrieval method

The overall retrieval procedure is based on the following steps.

Step 1. Extract features from videos, represent them as chromosomes, and index them into the database.

Select a video from a database and input it to the shot boundary detection module. Detect abrupt shot boundaries (cuts) and gradual shot boundaries (dissolves etc) and take the first frame of each shot boundary as a key frame of the corresponding shot. Next, input each key frame to the feature extraction module, extract five features: average color histogram, average brightness, average edge histogram, average shot duration, and gradual change rate. Then, index them into the database and search table. Remember that it is represented as chromosome in Fig. 4. Iterate the above procedure for all database videos (See the indexing box in the top of Fig. 1).

Step 2. Present initial videos to the user and select videos similar to what a user has in mind.

The system presents randomly selected 15 videos to the user. Only the 1st frames from each video are displayed (see the system GUI in Fig. 5). Then a user examines whether his

Fig. 5 The system GUI (Graphic User Interface)

or her preference videos exist among them through watching each video sequence by pressing the play button (Fig. 6a). If a video is similar to what he or she has in mind (we called it target emotion here), select that video by checking. If not, do not check. Sometimes, since a user can feel it burdensome to check each video sequence because of fatigue or lack of time, we also include the function of examining only the key frames of every video (Fig. 6b).

Step 3. Obtain target chromosomes by applying crossover genetic operator to checked videos.

Suppose that M videos are checked in Step 2. Extract M associated chromosomes of checked videos from the database. Produce 15 chromosomes by listing M associated chromosomes plus 15-M chromosomes by selecting randomly some chromosomes more than once (if M is lower than 15). Finally, obtain 15 target chromosomes by applying a crossover operator among some of 15 chromosomes as follows: (1) select random pairs among some of 15 chromosomes; (2) select crossover points randomly in each pair; (3) swap a part of chromosome on the basis of those points. In our research, one point crossover of four possible points is used as shown in Fig. 7.

Step 4. Obtain result videos by computing the similarity between each of target chromosomes and each of chromosomes of database videos.

The 15 new videos where each of their chromosomes is most similar to each of target chromosomes are obtained based on the similarity function S(Q,D) between the presented target chromosome Q and the chromosome D of database video. The degree of similarity is computed by integrating Euclidean distances between Q and D for the five features.

Fig. 6 Interface for examining a video: **a** the way of seeing the whole frame sequence (by pressing the ▶ button under key frame image) **b** the way of seeing only key frames (by pressing ⓘ under key frame image)

$$
\begin{aligned}
S(Q,D) = {} & w_{\mathrm{RGB}} \times S_{\mathrm{RGB}}(Q,D) + w_{\mathrm{BRIGHT}} \times S_{\mathrm{BRIGHT}}(Q,D) + w_{EDGE} \times S_{EDGE}(Q,D) \\
& + w_{\mathrm{SHOTTIME}} \times S_{\mathrm{SHOTTIME}}(Q,D) + w_{\mathrm{GRADRATE}} \times S_{\mathrm{GRADRATE}}(Q,D) \\
S_{\mathrm{RGB}}(Q,D) = {} & \sum_{i=0}^{31} \left(\mathrm{Avg}H^{Q}_{\mathrm{RGB}}[i] - \mathrm{Avg}H^{D}_{\mathrm{RGB}}[i] \right)^{1/2} \\
S_{\mathrm{BRIGHT}}(Q,D) = {} & \left(\mathrm{AvgBright}_{Q} - \mathrm{AvgBright}_{D} \right) \\
S_{\mathrm{EDGE}}(Q,D) = {} & \sum_{i=0}^{71} \left(\mathrm{Avg}H^{Q}_{\mathrm{EDGE}}[i] - \mathrm{Avg}H^{D}_{\mathrm{EDGE}}[i] \right)^{1/2} \\
S_{\mathrm{SHOTTIME}}(Q,D) = {} & \left(\mathrm{AvgShotTime}_{Q} - \mathrm{AvgShotTime}_{D} \right)^{1/2} \\
S_{\mathrm{GRADRATE}}(Q,D) = {} & \left(\mathrm{GradRate}_{Q} - \mathrm{Gradrate}_{D} \right)^{1/2}
\end{aligned}
\tag{9}
$$

where $S_{\mathrm{RGB}}(Q,D)$, $S_{\mathrm{BRIGHT}}(Q,D)$, $S_{\mathrm{EDGE}}(Q,D)$, $S_{\mathrm{SHOTTIME}}(Q,D)$, and $S_{\mathrm{GRADRATE}}(Q,D)$ are the similarities of average color histogram, average brightness, average edge histogram, average shot duration, and gradual shot ratios between the target chromosome Q and the chromosome D of database video, respectively. w_{RGB}, w_{BRIGHT}, w_{EDGE}, w_{SHOTTIME}, and w_{GRADRATE} are weights of relative importance among $S_{\mathrm{RGB}}(Q,D)$, $S_{\mathrm{BRIGHT}}(Q,D)$, $S_{\mathrm{EDGE}}(Q,D)$, $S_{\mathrm{SHOTTIME}}(Q,D)$, and $S_{\mathrm{GRADRATE}}(Q,D)$, respectively, and are subject to $w_{\mathrm{RGB}} + w_{\mathrm{BRIGHT}} + w_{\mathrm{EDGE}} + w_{\mathrm{SHOTTIME}} + w_{\mathrm{GRADRATE}} = 1$.

The lower value $S(Q,D)$ has, the higher similarity between Q and D is. In our research, $w_{\mathrm{RGB}}=0.2$, $w_{\mathrm{BRIGHT}}=0.2$, $w_{\mathrm{EDGE}}=0.1$, $w_{\mathrm{SHOTTIME}}=0.25$, and $w_{\mathrm{GRADRATE}}=0.25$ are empirically chosen as default weights.

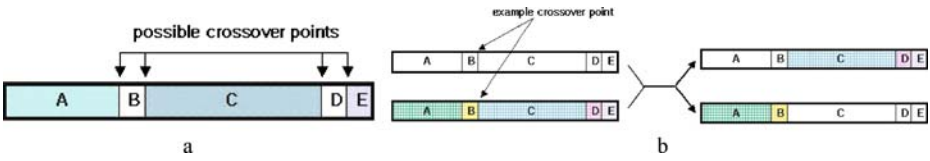

Fig. 7 Crossover application: **a** possible crossover points **b** example of one crossover point

a

b

c

Fig. 8 Example key frames with the presence of six emotions: **a** action **b** excitement **c** suspense **d** quietness **e** relaxation **f** happiness

Fig. 8 (continued)

Step 5. Iterate step 2- step 4 until ending conditions are met.

By iterating step 2 through step 4 until at least one of the ending conditions is satisfied, we can obtain the user preference videos. Ending conditions are that (1) the user is satisfied with the current results, or (2) there is no significant change even after many iterations, or (3) the user does not want to search the videos because of time or fatigue.

5 Experiments

To evaluate the proposed video retrieval method, several experiments were performed on a Pentium PC. The computer programs have been implemented in Visual C++. The experimental videos consist of 300 TV commercials (total 2.5 GB). They have a single scene for a very short running time (less than 1 min). We chose crossover rate as 0.54 to take eight chromosomes from 15 chromosomes for crossover candidates.

The system presents an initial population that consists of 15 videos selected randomly as shown in Fig. 5. A user watches each video by playing it or examining key frames, and then selects all videos similar to what he or she has in mind. Based on the selected videos,

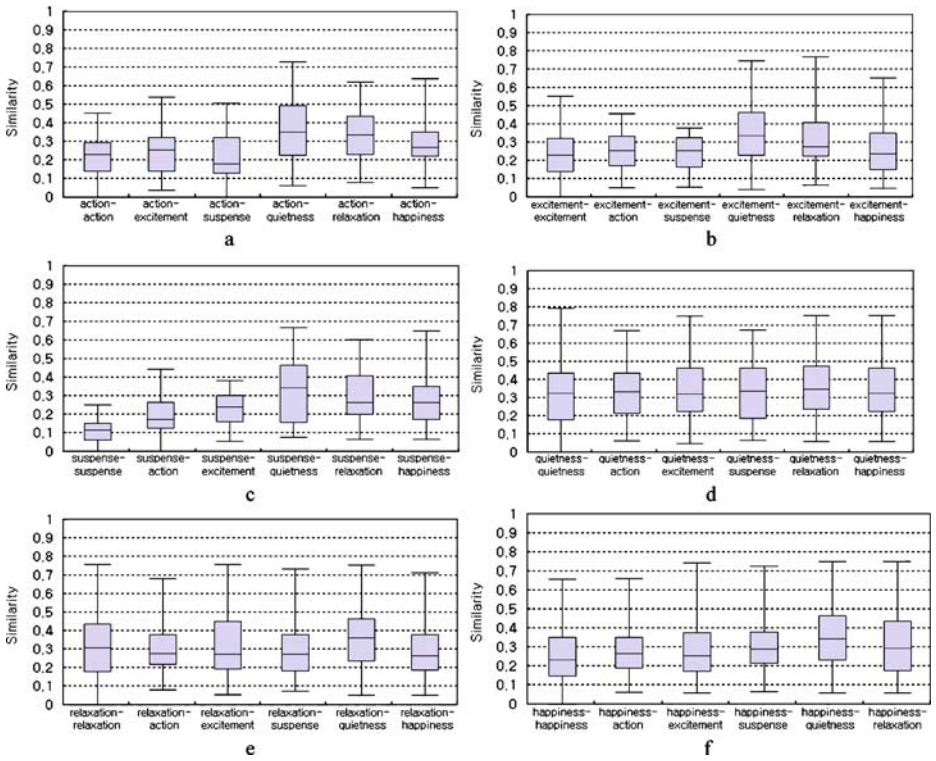

Fig. 9 The intra-class similarity and the inter-class similarity for videos in terms of six emotion categories: **a** action **b** excitement **c** suspense **d** quietness **e** relaxation and **f** happiness

15 chromosomes are produced. Then, among 15 chromosomes, take eight chromosomes randomly for crossover candidates and also, take a crossover point randomly. Then, list 15 target chromosomes by taking eight crossovered chromosomes plus seven non-crossovered chromosomes. Next, the system presents 15 database videos that are similar to the 15 target chromosomes through the similarity function (9). This procedure is repeated until at least one of the ending conditions is met. In case that the results of the next iteration are not satisfactory, the system allows the user to increase or decrease the weights of each feature.

For what he/she has in mind, six emotions of "action," "excitement," "suspense," "quietness," "relaxation," and "happiness" were used for experiments. These are also used for commercial video retrieval in [8]. However, the proposed method is not limited to the six emotions above, and it can be easily extended to other emotions. According to [8], the "action" scene has the presence of red and purple, short sequences joined by cuts, and a high degree of motion. "Excitement" scene has short sequences joined through cuts. "Suspense" scene that is associated with "action" has both long and short sequences joined through frequent cuts. "Quietness" scene has the presence of blue, orange, green, and white colors, and long sequences joined through dissolves. "Relaxation" scene does not have relevant motion components. "Happiness" scene shares "happiness" features and has relevant motion components. These characteristics can be represented as chromosomes with five features proposed in this paper. We have asked ten subjects to categorize 300 videos into the six emotions above. Resultant categories are 36 videos for "action," 56 videos for

"excitement," 22 videos for "suspense," 72 videos for "quietness," 71 videos for "relaxation," and 82 videos for "happiness." Since emotion is subject to individuals greatly, all the 300 videos used in the experiments are what all participants agree with for each of emotions. The example key frames of each category are listed in Fig. 8.

To evaluate the proposed method, we have examined two things in the experiments. Fist, we examine whether or not the five features (average color histogram, average brightness, average edge histogram, average shot duration, and gradual change rate) are appropriate for the genetic representation. Second, we perform the retrieval process up to maximum ten iterations and check the user's satisfaction (effectiveness) on the resulting videos.

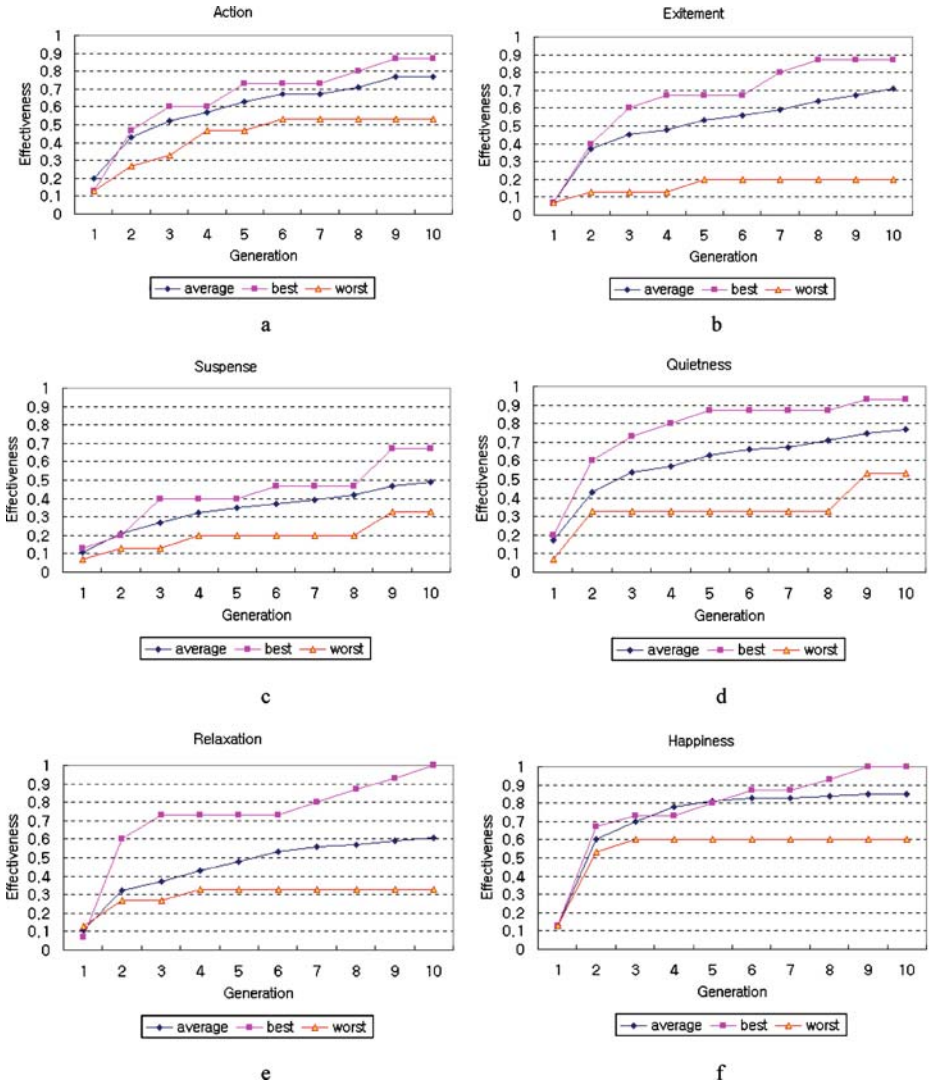

Fig. 10 User's satisfaction (system effectiveness) according to six emotions: **a** action **b** excitement **c** suspense **d** quietness **e** relaxation and **f** happiness

5.1 Usefulness of features

Among the six categories, we compute the intra-class similarity and the inter-class similarity using (9). For experimental convenience, we have randomly selected 30% of the videos from each category (12 for "action," 17 for "excitement," 7 for "suspense," 22 for "quietness," 21 for "relaxation," and 25 for "happiness"). The two similarities give the ground truth of whether videos in the same category are closer than those in the other category over feature space. The intra-class similarity and the inter-class similarity according to the emotions are depicted with box plots in Fig. 9. The box covers 50% of similarity data. Three horizontal lines within the box describe the lower (25%), median (50%), and upper (75%) quartiles in the order from bottom to top. Two horizontal lines outside of the box describe the maximum value and the minimum value. The lower the values are, the closer the videos are over the feature space (That is, videos are more similar to one another). For example, in Fig. 9a, when 12 videos within the cluster "action" are presented as queries, the values of intra-class (i.e. "action–action") are lower than the values of inter-class such as "action–quietness." It means that videos of intra-class are more similar to one another than those of inter-class "action–quietness." Also, for other cases such as "action–excitement," "action–relaxation," and "action–happiness," the values of "action–action" are lower as the whole. However, in the case of "action–suspense," it is hard to judge because "action–action" is lower for a maximum value and a 75% quartile but higher for a median and a 25% quartile. In Fig. 9b, the values of intra-class ("excitement–excitement) are lower than the values of inter-class such as "excitement–quietness," "excitement–relaxation," and "excitement–happiness." However, it is hard to judge for "excitement–action" and "excitement–suspense." Also, in Fig. 9c, the values of intra-class "suspense–suspense" are lower than the values of inter-class. It is, however, hard to judge for emotions such as "quietness" and "relaxation" categories as in Fig. 9d–e. In Fig. 9f, the values of intra-class "happiness–happiness" are lower than the values of inter-class. Consequently, except for "quietness" and "relaxation," most of the intra-class videos are more similar than inter-class videos as a whole. It indicates the five features are appropriate for emotion-based retrieval.

5.2 System effectiveness

After conducting up to ten feedback iterations, the user's satisfaction on retrieved videos has been examined. Ten graduate students, who have been trained about how the system works and what the six features mean, participate in the experiment. For each emotion, the user's satisfaction (we call it the system effectiveness in this paper) is computed using (10). Final results were obtained by averaging ten results from the ten graduates.

$$\text{Effectiveness} = \frac{N_{\text{CORRECT}}}{N_{\text{TOTAL}}} \tag{10}$$

where N_{TOTAL} is the number of displayed videos (i.e. 15) and N_{CORRECT} is the number of videos that the user satisfies among 15 videos.

Table 2 Average effectiveness after ten feedbacks

	Action	Excitement	Suspense	Quietness	Relaxation	Happiness
Effectiveness	0.77	0.71	0.49	0.77	0.61	0.85
Average	0.7					

For the six emotions, the system effectiveness up to the tenth feedback (it is called as the tenth generation in genetic algorithm) is depicted as the best, worst, and average cases as shown in Fig. 10. As we expect, it increases according to the feedback. In the tenth generation, it shows about 0.7 on the average. This means that, among the 15 retrieved videos after the tenth feedback, the user has found it among ten videos that he or she has in mind. Except for "suspense" and "relaxation," the effectiveness is more than 0.7. "Happiness" shows the best result 0.85. "Suspense" shows the worst result 0.49 (see Table 2).

For "action," the effectiveness after ten feedbacks shows the maximum of 0.87, the minimum of 0.53, and average of 0.77. Since "action" videos consist of many cuts for a short time, time information such as average shot duration is very important. For "excitement," the effectiveness after ten feedbacks shows the maximum of 0.87, the minimum of 0.2, and average of 0.71. The low satisfaction of 0.2 is attributed to the fact that some videos have long shot duration and contain dissolve transitions, which are not common in "excitement" videos. For "suspense," the effectiveness after ten feedbacks shows the maximum of 0.67, the minimum of 0.33, and average of 0.49. This is the lowest among the results of the six emotions. For "quietness," the effectiveness after ten feedbacks shows the maximum of 0.93, the minimum of 0.53, and average of 0.77. Since "quietness" videos consist of many cuts and dissolve for a long time, time information such as the average shot duration and the gradual change rate are important. For "relaxation," the effectiveness after ten feedbacks shows the maximum of 1.0, the minimum of 0.33, and average of 0.61. Since "relaxation" videos consist of many cuts and dissolve for a long time, time information such as average shot duration is important. For "happiness," the effectiveness after ten feedbacks shows the maximum of 1.0, the minimum of 0.6, and average of 0.85. This is the highest among the results of the six emotions. Among 300 videos, 82 videos are related to "happiness" and most of them have white, green, and blue colors and long average-shot duration.

As you can see in Fig. 10, it is interesting to note that some cases have room for improvement after more than ten feedbacks. Thus, the user can perform more feedbacks to obtain better results (if time is allowed and if the user does not feel fatigued).

Commercials used in these experiments have numerous content variations because they have to deliver many pieces of information in a short time. On the other hand, videos such as news, drama, and movie do not. This observation has led us to believe that our method will yield better effectiveness for the scene retrieval of general videos (for example, scenes of an anchor person in the news video, conversation scenes, and so on). Also, in the GUI, we give two levels for user's satisfaction on each video by just checking a box. The checked video is considered to be the one that the user has in mind whereas non-checked video is not. In [7], a slider bar is provided to give a variety of levels for the user's degree of satisfaction. This method will have a chance that in the next generation, more videos that the user is satisfied with are retrieved since more satisfactory videos are likely to be chosen for crossover. However, in this paper, we did not use that approach because the use of slider bar can cause greater fatigue for the user than the use of a check box.

6 Conclusions

In this paper, a new emotion-based video scene retrieval method has been presented. The videos that a user has in mind can be retrieved by mapping his/her emotion space onto physical feature space through interactive genetic algorithm. Experiments on 300 commercial videos have yielded an average of 70% satisfaction for six selected emotions ("action," "excitement,"

"suspense," "quietness," "relaxation," and "happiness") after ten generations. In addition, it has provided a room for better results if the user tries to perform more than ten generations.

However, although the performance of the proposed method could be a solution for emotional video retrieval application, it has some limitations to overcome. One of them is that the five features (average color histogram, average brightness, average edge histogram, average shot duration, and gradual change rate) do not describe emotions induced from moving objects in the video. Hence, it would be necessary to add related features to the chromosomes by using a motion vector or optical flow technique. Also, one of the major problems of IGA is that it cannot evolve over many generations because of the user's fatigue. To deal with the problem we presented a quick overview of key frames as well as playing whole videos. However, evolving over many generations cannot eliminate the user's fatigue completely. Interface that reduces the burden of the user and fast convergence methodology of IGA remains to be the main subject for further research. To convince more the usefulness of the method, subjects more than ten and subjects naïve to the five features should be participated in the experiments. Currently, asian commercials were used for experiments. The use of well-known popular movies in the world would also show the performance of the method in a more reliable manner.

References

1. Bach JR, Fuller C, Gupta A, Hampapur A, Horowitz B, Humphrey R, Jain RC, Shu C (1996) The virage image search engine: an open framework for image management. In: Proc. SPIE, vol. 2670: Storage and Retrieval for Images and Video Databases IV, pp 76–86
2. Banzhaf W (1997) Interactive evolution. Handbook of evolutionary computation. IOP, Oxford
3. Biles JA (1994) GenJam: a genetic algorithm for generating jazz solos. In: Proc. Int. Computer Music Conf, pp 131–137
4. Caldwell C, Johnston VS (1991) Tracking a criminal suspect through face–space with a genetic algorithm. In: Proc. Int. Conf. Genetic Algorithm, pp 416–421
5. Carson C, Belongie S, Greenspan H, Malick J (2002) Blobworld: image segmentation using expectation–maximization and its application to image querying. IEEE Trans Pattern Anal Mach Intell 24(8):1026–1038
6. Snoek Cees GM, Worring M (2005) Multimodal video indexing: a review of the state-of-the art. Multimedia Tools and Applications 25(1):5–35
7. Cho S-B (2002) Towards creative evolutionary systems with interactive genetic algorithm. Appl Intell 16 (2):129–138
8. Colombo C, Del Bimbo A, Pala P (1999) Semantics in visual information retrieval. IEEE Multimed 6 (3):38–53
9. Colombo C, Del Bimbo A, Pala P (2001) Retrieval of commercials by semantic content: the semiotic perspective. Multimedia Tools and Applications 13(1):93–118
10. Cox IJ, Miller ML, Minka TP, Papathomas TV, Yianilos PN (2000) The Bayesian image retrieval system, PicHunter: theory, implementation and psycophysical experiments. IEEE Trans Image Process 9 (1):20–37
11. Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, Yanker P (1995) Query by image content: the QBIC system. IEEE Computer 28(9):23–31
12. Gargi U, Kasturi R, Strayer SH (2000) Performance Characterization of Video-shot-change detection methods. IEEE Trans Circuits Syst Video Technol 10(1):1–13
13. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA
14. Itten J (1961) Art of color (Kunst der Farbe). Otto Maier Verlag, Ravensburg, Germany (in German)
15. Jain AK, Vailaya A, Xiong W (1999) Query by video clip. Multimedia Syst: Special Issue on Video Libraries 7(5):369–384
16. Joseph T, Cardenas A (1988) PicQuery: a high-level query language for pictorial database management. IEEE Trans Softw Eng 14(5):630–638

17. Lee J.-Y, Cho S.-B (1998) Interactive genetic algorithm for content-based image retrieval. In: Proc. Asia Fuzzy Systems Symposium, pp 479–484
18. Ma WY, Manjunath BS (1999) Netra: a toolbox for navigating large image databases. Multimedia Syst 7 (3):184–198
19. Minka TP, Picard RW (1997) Interactive learning using a society of models. Pattern Recogn 30(3):565–581
20. Pentland A, Picard RW, Sclaroff S (1996) Photobook: content-based manipulation of image databases. Int J Comput Vis 18(3):233–254
21. Pickens J, Bello JP, Monti G, Crawford T, Dovey M, Sandler M, Byrd D (2002) Polyphonic score retrieval using polyphonic audio queries: a harmonic modeling approach. In: Proc. ISMIR, pp 13–17
22. Roussopolous N, Faloutsos C, Sellis T (1988) An efficient pictorial database system for pictorial structured query language (PSQL). IEEE Trans Softw Eng 14(5) 639–650
23. Rui Y, Huang TS, Ortega M, Mehrota S (1998) Relevance feedback: a power tool in interactive content-based image retrieval. IEEE Trans Circuits Syst Video Technol 8(5):644–655
24. Smith JR, Chang S-E (1996) VisualSEEK: a fully automated content-based image query system. In: Proc. ACM Multimedia, pp 87–98
25. Soen T, Shimada T, Akita M (1987) Objective evaluation of color design. Color Res Appl 12(4):184–194
26. Takagi H (2001) Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. Proc IEEE 89(9):1275–1296
27. Takagi H, Noda T, Cho S-B (1999) Psychological space to hold impression among media in common for media database retrieval system. In: Proc. IEEE Int. Conf. on System, Man, and Cybernetics, 263–268
28. Toivanen J, Seppänen T (2002) Prosody-based search features in information retrieval. TMH-QPSR 44 Fonetik
29. Truong BT, Dorai C, Venkatesh S (2000) New enhancements to cut, fade, and dissolve detection processes in video segmentation. In Proc. ACM Int. Conf. on Multimedia, pp 219–227
30. Um J-S, Eum K-B, Lee J-W (2002) A study of the emotional evaluation models of color patterns based on the adaptive fuzzy system and the neural network. Color Res Appl 27(3):208–216
31. Vailaya A, Figueiredo MAT, Jain AK, Zhang HJ (2001) Image classification for content-based indexing. IEEE Trans Image Process 10(1):117–130
32. Vailaya A, Jain AK, Zhang HJ (1998) On image classification: city images vs. landscapes. Pattern Recogn 31(12):1921–1936
33. Yeo BL, Liu B (1995) Rapid scene analysis on compressed video. IEEE Trans Circuits Syst Video Technol 5(6):533–544
34. Yoo H.-W, Jang D-S (2004) Automated video segmentation using computer vision technique. International Journal of Information Technology and Decision Making 3(1):129–143
35. Yoo H-W, Jang D-S, Jung S.-H, Park J-H, Song K-S (2002) Visual information retrieval system via content-based approach. Pattern Recogn 35(3):749–769
36. Yoo H-W, Jung S-H, Jang D-S, Na Y-K (2002) Extraction of major object features using VQ clustering for content-based image retrieval. Pattern Recogn 35(5):1115–1126
37. Zabih R, Miller J, Mai K (1999) A feature-based algorithm for detecting and classifying production effects. Multimedia Syst 7(2):119–128
38. Zhang HJ, Kankanhalli A, Smoliar SW, Tan SY (1993) Automatic partitioning of full motion video. Multimedia Syst 1(1):10–28
39. Zhang HJ, Wu J, Zhang D, Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Pattern Recogn 30(4):643–658

Hun-Woo Yoo is a research professor at the Center for Cognitive Science at Yonsei University. He received his B.S. and M.S. degrees in Electrical Engineering from Inha University, Korea and a Ph.D. degree in

Industrial Systems and Information Engineering at Korea University, Korea. From 1994 to 1997, he has worked as a research engineer at the Manufacturing Technology Center of LG Electronics. His current research interests include multimedia information retrieval, computer vision, and image processing.

Sung-Bae Cho received the B.S. degree in computer science from Yonsei University, Seoul, Korea, in 1988 and the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Taejeon, Korea, in 1990 and 1993, respectively. From 1991 to 1993, he worked as a Member of the Research Staff at the Center for Artificial Intelligence Research at KAIST. From 1993 to 1995, he was an Invited Researcher of Human Information Processing Research Laboratories at ATR (Advanced Telecommunications Research) Institute, Kyoto, Japan. In 1998, he was a Visiting Scholar at University of New South Wales, Canberra, Australia. Since 1995, he has been a Professor in the Department of Computer Science, Yonsei University. His research interests include neural networks, pattern recognition, intelligent man-machine interfaces, evolutionary computation, and artificial life.

Dr. Cho is a Member of the Korea Information Science Society, INNS, the IEEE Computer Society, and the IEEE Systems, Man and Cybernetics Society. He was awarded outstanding paper prizes from the IEEE Korea Section in 1989 and 1992, and another one from the Korea Information Science Society in 1990. In 1993, he also received the Richard E. Merwin prize from the IEEE Computer Society. In 1994, he was listed in Who's Who in Pattern Recognition from the International Association for Pattern Recognition and received the best paper awards at International Conference on Soft Computing in 1996 and 1998. In 1998, he received the best paper award at World Automation Congress. He was listed in Marquis Who's Who in Science and Engineering in 2000 and in Marquis Who's Who in the World in 2001.