Intl. Journal of Human–Computer Interaction, 29: 96–109, 2013

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2012.692316

Taylor & Francis
Taylor & Francis Group

Trial Realization of Human-Centered Multimedia Navigation for Video Retrieval

Miki Haseyama and Takahiro Ogawa
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

A trial realization of human-centered navigation for video retrieval is presented in this article. This system consists of the following functions: (a) multimodal analysis for collaborative use of multimedia data, (b) preference extraction for the system to adapt to users' individual demands, and (c) adaptive visualization for users to be guided to their desired contents. By using these functions, users can find their desired video contents more quickly and accurately than with the conventional retrieval schemes since our system can provide new pathways to the desired contents. Experimental results verify the effectiveness of the proposed system.

1. INTRODUCTION

In IDC white papers published in March 2008 (Gantz et al., 2008) and May 2010 (Gantz & Reinsel, 2010), it was reported that the digital universe in 2007 was 281 exabytes and that it will be 44 times larger in 2020 than in 2009. Also, most of the data contained in the digital universe are unstructured data such as images, music, and videos (i.e., multimedia data). Nowadays, users always access the data to find useful information, generally based on the query–response model, in which users input a query into the retrieval interface and then get the desired information.

The query–response model can narrow huge information down to a suitable size for users to check according to the query. Actually, metadata are attached to multimedia contents to be retrieved in advance; retrieval engines present multimedia content, the metadata of which are matched to the queries, to users; after users input queries once or more than once, retrieval engines effectively guide users to the useful information. Traditionally, metadata of the contents were generated from information such as dates, times, and places, which were obtained in their acquisition, and manually annotated keywords were also utilized as metadata. Recently, image analysis and video analysis have been improved, and technologies for automatically extracting metadata have therefore been proposed (Bay, Ess, Tuytelaars, & Gool, 2008; Csurka, Dance, Fan, & Bray, 2004; Lowe, 1999; Mikolajczyk & Schmid, 2005; TRECVID: National Institute of Standards and Technology, 2000). Machine learning has contributed tremendously to the development of these technologies, especially for deriving semantic-level indices, and it is accelerating the progress in content-based image and video retrieval research (Flickner et al., 1995; Greetha & Narayanan, 2008; Smeulders, Worring, Santini, Gupta, & Jain, 2000). However, performance has not been satisfactory, and it is still difficult to perfectly solve the problem of "Semantic Gap" (Picard, 2003). Specifically, machine learning-based video retrieval approaches enable extraction of semantic level indices (i.e., annotation). Then, by using keywords that represent contents desired by users, semantic retrieval becomes feasible. However, to realize accurate estimation of semantic-level indices, a sufficient amount of training data must be provided, and it becomes difficult to perform accurate estimation when a large amount of training data cannot be prepared. Generally, it is difficult to prepare enough training contents for each semantic-level index, and the estimation of semantic-level indices therefore has some errors. In such cases, because the semantic features obtained from the estimated semantic-level indices also have errors, the distances between two video scenes based on those features cannot be accurately calculated. The retrieval results may therefore not be the same as the ones desired by users.

Smeulders et al. (2000) reported that the pivotal point in content-based retrieval is that the user seeks semantic similarity, but the database cannot provide similarity by data processing, and this is the so-called semantic gap. Furthermore, we have to solve the following new problems as well as the aforementioned problems. When users cannot provide specific queries representing their desired contents, it becomes difficult to discover those contents by retrieval based on the query response model (Campbell, 1996). This was also pointed out in IDC report 2010 (Gantz & Reinsel, 2010). It was stated in that report that we will have to consider how we find the

This work was partly supported under project SCOPE (Strategic Information and Communications R&D Promotion Programme) of the Japanese Ministry of Internal Affairs and Communications. This research was partly supported by a Grant-in-Aid for Scientific Research (B) 21300030, from the Japan Society for the Promotion of Science (JSPS).

Address correspondence to Takahiro Ogawa, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0814 Japan. E-mail: ogawa@lmd.ist.hokudai.ac.jp

information we need when we need it, and thus new search and discovery tools must be developed.

In this article, we present a human-centered navigation system for video retrieval as a solution to improving the conventional retrieval systems. Lew, Sebe, Djeraba, and Jain (2006) reported that human-centered computing is one of the research topics that has potential for improving multimedia retrieval by bridging the semantic gap. In human-centered computing, the main idea is also to satisfy the users and allow users to make queries in their own terminology; user studies give us direct insights into interactions between humans and computers. Therefore, human-centered computing is effective for not only bridging the semantic gap but also solving the problem of users not being able to provide specific queries. The proposed human-centered navigation system is equipped with the following three functions.

1. Collaborative use of multimedia data: Sometimes, users cannot represent their desired contents in words. Collaborative use of multimedia data such as visual data, audio data, text data, link relationship, and sensor data plays a significant role. This is known as a multimodal approach (Babaguchi, Ishida, & Morisawa, 2004; Bruno, Moenne-Loccoz, & Marchand-Maillet, 2008; Calic, Campbell, Dasiopoulou, & Kompatsiaris, 2005; Sudha, Shalabh, Basavaraja, & Sridhar, 2008) and is necessary for overcoming the limitation when we use only one type of content. This means that to improve the performance of video retrieval, collaborative use of multimedia data is essential. Note that the basic idea of this approach is based on previous works, and its novelty in our system is less than that of the following two functions.
2. User's preference extraction: No technology can completely satisfy users' demands because their preferences are continuously changing according to their situations. Thus, some schemes for estimating users' preferences are necessary to realize the aforementioned human-centered computing. Then the system can extract useful information from users, and thus users can become the terminologies for providing queries.
3. Adaptive visualization: Retrieval results have to be exhibited for users to be aware of their desired contents. Furthermore, to realize the second point's user's preference extraction, the interface must connect the users and our system. Therefore, adaptive visualization to their interest (Chorianopoulos, 2008) is highly effective for leading them to the desired contents.

The aforementioned functions are necessary to provide a solution and to overcome the limitations of the traditional query–response model. Thus, we have developed a system that is equipped with these functions; it is called the human-centered navigation system hereafter and is shown in the following section.

This article is organized as follows. First, the basic concepts of human-centered systems and key functions for realizing human-centered systems in the proposed method are explained in the second section. In the third section, we present a new human-centered video navigation system that is equipped with the key functions shown in the second section. To verify the effectiveness of the proposed system, results of some experiments are shown in the fourth section. Finally, conclusions are given in the fifth section.

2. HUMAN-CENTERED SYSTEMS AND THEIR KEY FUNCTIONS

This section presents the basic concepts of human-centered systems and key functions for their realization in the proposed method. As pointed out by Lew et al. (2006), current systems have significant limitations, such as inability to understand a wide user vocabulary and the user's satisfaction level in searching for a particular media item. Current research topics that have potential for improving multimedia retrieval by bridging the semantic gap are as follows: human-centered computing, new features, new media, browsing and summarization, and evaluation/benchmarking. In human-centered computing, the main idea is to satisfy users and allow users to make queries in their own terminology. User studies provide direct insights into interactions between humans and computers. By human-centered, we mean systems that consider the behavior and needs of the human user. As noted earlier, the foundational areas of multimedia information retrieval were often in computing-centric fields. However, because the primary goal is to provide effective browsing and search tools for the user, it is clear that the design of the systems should be human-centric. There have been several major recent initiatives in this direction, such as user understanding, experiential computing, and affective computing (Bertino, Hacid, & Toumani, 2005; Jaimesa & Sebeb, 2007; Jain, 2008; Kooper & MacIntyre, 2003; Shneiderman, 1990).

From the prior discussion, we present three key concepts for our human-centered video navigation system. Recent systems are mostly rank list-based browsing interfaces showing retrieval results that are best matched to queries provided by users in turn. In these systems, users should provide specific queries that correctly represent what they want in order to quickly find desired contents. This means that for showing these desired contents in higher ranks, users should provide specific queries. Otherwise, their desired contents cannot be shown in high ranks, and the users must search lower ranks. Furthermore, we have to note another point—semantic gap. Generally, recent retrieval systems automatically perform indexing or annotation for contents in their databases. However, it is well known that their performance is not perfect, and they cannot correctly grasp semantic concepts. Therefore, even if users can provide specific queries, the best-matched contents provided by the systems may not correspond to their truly desired contents. In such a case, users must also search lower ranks. To tackle this problem, collaborative use of multimedia data such as visual data,

audio data, text data, link relationship, and sensor data plays a significant role. Different kinds of data should mutually complement their limitations. Furthermore, users' preferences are continuously changing according to their situations, that is, their specific queries are not always the same even if they watch the same contents. Thus, some schemes for knowing their situations to find their changeable preferences are necessary for solving the aforementioned problem, and this is called user's preference extraction. Finally, even if the aforementioned two key functions are realized, it is difficult to perform perfect retrieval of desired video contents, and thus users must "search for" their truly desired contents through browsing interfaces. Therefore, because the interfaces should enable users to reach such desired video contents, adaptive visualization to their interest is highly effective for leading them to the desired contents.

3. HUMAN-CENTERED VIDEO NAVIGATION SYSTEM

This section presents a new system that is a feasible solution to realize human-centered navigation for video retrieval. This system has been developed for navigating users to desired video contents on the web or broadcasted on TV or in camcorders. First, we show the algorithms in our navigation system. Specifically, it consists of the following five parts.

Part 1: Preparation: Scene Segmentation of Video Sequences

Before processing video contents in a target database, we have to divide them into some basic units. Therefore, in the first part, scene segmentation is performed for each video content. The definition of scenes is given later. In this part, attribution probabilities of audio classes are estimated for audio signals based on principal component analysis (PCA), Mahalanobis's generalized distance (MGD), and a fuzzy algorithm, and scene cuts are detected for target video contents.

Part 2: Distance Calculation Based on Audio-Visual Features

In the second part, distances based on audio-visual features between two scenes are calculated. Our algorithm extracts audio-visual features for these two scenes and calculates the two kinds of distances respectively concerning their audio and visual sequences by using dynamic time warping (DTW).

Part 3: Distance Calculation Based on Music Features

In the third part, distances based on music features between two scenes are calculated. Our algorithm extracts music features based on bass and non-bass, which represent melodies of music, for music pieces in these two scenes and calculates their distances based on DTW.

Part 4: Distance Calculation Based on Text Features

In the fourth part, distances based on text features between two scenes are calculated. First, text features based on terms are extracted by using "Julius"—text extraction and morphological analysis of audio features. Then their distances based on the tf-idf method are calculated.

Part 5: Visualization of Retrieval Results

The final part is visualization of retrieval results. From the distances obtained by Parts 2 through 4, weighted distances between two scenes are calculated. The weights can be obtained from a "preference board" through the interface from users, and its details are shown next. Then our system visualizes retrieval results in the visualization space.

The proposed method utilizes the multiple distances obtained in Parts 2 to 4 to realize "collaborative use of multimedia data," and this multimodal approach enables users to find desired contents through several aspects, that is, several kinds of features. It should be noted that the final distances between video contents are determined by the function of "user's preference extraction" equipped in the interface of "adaptive visualization." Therefore, the three functions are all necessary in the proposed navigation system for video retrieval. The rest of this section presents the details of Parts 1 through 5.

3.1. Scene Segmentation of Video Sequences (Part 1)

This section presents scene segmentation of video sequences, which is the preprocessing of the proposed system. Generally, shots and scenes are regarded as basic units for audio-visual segmentation and classification. In this article, a shot denotes a set of image frames in a video sequence obtained by one camera without interruption (Huang, Liu, & Wang, 2002), and a scene consists of one or more shots that are semantically correlated. The boundaries between two adjacent shots and scenes are called shot-cuts and scene-cuts, respectively.

First, the proposed system divides each video sequence into several shots, that is, shot-cut detection is performed by the method described in Patel and Sethi (1996). This method utilizes the chi-square test for shot-cut detection. The shot-cut detection method that utilizes the chi-square test is a well-known method, and it can detect shot-cuts accurately. However, the level of accuracy of the shot-cut detection results becomes lower when several effects, such as fade and dissolve, are added to the shot-cuts. Therefore, we utilize not only the method proposed in Patel and Sethi (1996) but also the fade and dissolve detection method proposed in Truong, Darai, and Venkatesh (1996) for shot-cut detection.

From the obtained shot-cut detection results, we perform shot classification for realizing scene segmentation. The proposed system adopts our previously reported classification method using PCA, MGD, and a fuzzy algorithm (Nitanda & Haseyama, 2007). This method consists of two parts, an audio analysis part and a shot classification part. In the audio analysis

part, both PCA and MGD are utilized, and the effective features for the analysis can be automatically obtained. In the shot classification part, the method in Nitanda and Haseyama (2007) utilizes a fuzzy algorithm and enables calculation of the attribution probabilities of each shot belonging to shot classes Si, Sp, Mu, No, SpMu, and SpNo, these six classes representing silence, speech, music, noise, speech with music background, and speech with noise background, respectively. The proposed system regards the six attribution probabilities as features of each shot, and a six-dimensional feature vector can be obtained for each shot. Then the distance of feature vectors between two neighboring shots in the target video sequence can be calculated, and scene-cuts for which distances are larger than a predefined threshold value can be detected.

3.2. Distance Calculation Based on Audio-Visual Features (Part 2)

This subsection presents distance calculation based on audio-visual features in the proposed system. The proposed system calculates the following features for each clip, which is a short term in each video scene:

Visual features (48 dimensions). The color histogram of the first frame in each clip is calculated with HSV color space being utilized, and the numbers of the bins are 12, 2, and 2 for Hue, Saturation, and Value, respectively. Some high-order visual features such as SIFT (Lowe, 2004), HOG (Dalal & Triggs, 2005), and Bag of keypoints (Csurka et al., 2004) can also be utilized for the visual features in our system. In this article, we utilize only the color features for simplicity.

Audio features (22 dimensions). The averages and standard deviations of the following 11 features are computed: volume, zero-crossing rate, pitch, frequency centroid, frequency bandwidth, sub-band energy ratio (0–630 Hz) (630–1720Hz) (1720–4400 Hz) (4400–11025 Hz), nonsilence ratio, and zero-ratio. These features are also utilized in several conventional methods (Liu, Wang, & Chen, 1998; Nitanda & Haseyama, 2007; Zhang & Kudo, 2001).

Next, the proposed system calculates two kinds of distances for the aforementioned audio and visual features between two scenes by using DTW (Pikrakis & Kamarotos, 2003). DTW aligns two time series and computes the distance between them. Because DTW takes into account the expansion and contraction of the series, it can appropriately represent the distance between two series even if the series include these effects. Therefore, the proposed method regards video sequences as time series of audio or visual features and calculates the distance between two scenes for which lengths are different. Then two kinds of distances are obtained for audio and visual features.

3.3. Distance Calculation Based on Music Features (Part 3)

In this subsection, distance calculation based on music features between two scenes is explained. In the study field of music, it is well known that three elements—melody, rhythm, and code—are the most important for analyzing music pieces. The proposed system focuses on melodies and derives distance calculation based on their features. It should be noted that we cannot calculate music features for scenes not including any music. In the proposed system, we estimate the attribution probabilities utilized for the scene segmentation and calculate the music features from audio signals with attribution probabilities of the music class that are higher than those of the other classes. If any parts of the target audio signals do not have the highest attribution probabilities of the music class, we cannot compare the music features and do not perform the distance calculation for a target scene. Details of the distance calculation based on music features are shown in the rest of this subsection.

The proposed system adopts our previously reported method for music analysis (Kobayashi & Haseyama, 2007). Specifically, we divide music into two parts, a bass track and a nonbass track, for the calculation of distances based on melody lines. First, the time series of the fundamental frequency of the bass sound is calculated as melody lines of the bass track. The fundamental frequency of the bass track can be accurately estimated because of the following two characteristics: (a) the energy of a bass sound is concentrated in a lower frequency band than those for other sounds, and this band is generally limited to 40 to 250 Hz because there are only a few harmonic sound components, and (b) the duration of a bass sound is longer than the durations of other sounds. Therefore, to estimate the fundamental frequency of the bass track, we calculate time series of the weighted power spectrum for each scene. Then the time series of the fundamental frequency of the bass track can be obtained as those of the maximum point of the weighted spectrum.

Furthermore, the time series of energy corresponding to the pitch notation is calculated as melody lines of the nonbass track. Generally, the sound of nonbass instruments contains harmonic sound components. In addition, some instruments compose the sound in the same frequency bands. Thus, the melody lines of each instrument cannot be easily obtained. Therefore, our system utilizes energy of the frequency that corresponds to the pitch notation as melody lines of the nonbass track.

In this way, we can obtain two kinds of features based on the bass track and the nonbass track for each scene, where the obtained features are the time series concerning melody lines. The proposed system computes the distance of the bass tracks and the nonbass tracks between two scenes by utilizing DTW because the lengths of the two scenes (i.e., two music components) are generally different. Then, by multiplying two distances obtained for the bass track and nonbass track by DTW, the distance based on the melody lines can be calculated between two scenes. Details of the feature extraction and the distance calculation are shown in Kobayashi and Haseyama (2007).

3.4. Distance Calculation Based on Text Features (Part 4)

In this subsection, the distance of topics by utilizing text features based on speech recognition results is defined. First, the proposed system extracts audio signals including speech with attribution probabilities of Sp, SpMu, and SpNo used in the scene segmentation that are higher than the other attribution probabilities for each scene. Then speech recognition is performed for the speech sequences, and terms are extracted from the speech recognition results. Specifically, Julius (http://julius.sourceforge.jp/), a large vocabulary continuous speech recognition system, is utilized for speech recognition in the proposed system. The language model, acoustic model, and dictionary adopted in the proposed method are those of the Julius dictation kit. Furthermore, the extracted terms are weighted by the tf-idf function (Sebastiani, 2002), which is widely used in information retrieval. Note that scenes are regarded as documents to apply the tf-idf function to video contents. Then, for each scene, a vector with elements that are the weights of the terms is obtained. The proposed system defines the distance of text features between two scenes by calculating the distance of their feature vectors. It should be noted that distances cannot be calculated between two scenes that do not contain speech signals. In such cases, we utilize other distances described in the previous subsections.

The speech recognition system used in the proposed system has limitations in recognition accuracy in order not to lose the advantage of a large amount of training data not being needed. Therefore, we define the features based on the speech recognition results as not being sensitive to recognition errors. Because the distance of topics is computed by statistical weights, it tends not to be sensitive to speech recognition errors.

3.5. Visualization of Retrieval Results (Part 5)

This subsection presents the visualization of retrieval results. From the previous subsections, we can calculate the distances of audio-visual features, music features, and text features. By combining these distances through the visualization interface in the proposed system, the collaborative use of multimedia data (i.e., multimodal video retrieval) is realized. Furthermore, in this scheme, users' preferences should be extracted through the adaptive visualization interface. Therefore, the proposed system must be equipped with the aforementioned three key functions for visualization of retrieval results. Details of this interface are given in the rest of this subsection.

Before explaining the specific procedures, we present the interface of the proposed system. The details of the adaptive visualization interface are shown in Figure 1. The library contains video contents, and a new query content can be initially selected. From the query content, we retrieve four similar video contents, where the upper left two contents tend to be retrieved from their visual features and the lower right two contents tend to be retrieved from their audio features, the details of which are shown later. Furthermore, the analysis board shows how similar the query content and the retrieved content are; that is, it shows several calculated distances. Finally, the preference board determines weights of visual, audio-music, and text distances from users' operations for retrieving similar video contents (Takahashi & Haseyama, 2007). Then, through this interface, new video contents are associatively retrieved, and the proposed system leads users to the desired contents.

Figure 2 shows the procedures for visualization of retrieval results. First, the proposed system obtains four distances of visual features, audio features, music features, and text features. The proposed system merges the two distances of audio and music features. Specifically, if the sum of the attribution probabilities of music in two target scenes are higher than those of the other audio classes, we utilize only the distance of music features. Otherwise, the distance of audio features is only utilized.

From the obtained three distances (visual, audio-music, text), the proposed system calculates the final distance. Specifically, the proposed system computes the weighted sum of the three distances, the weights being determined from the preference board. The weight of each element (visual, audio-music, text) is the inverse number of the distance from the point provided by users through the preference board, and it is normalized by using all of the weights. From the obtained final distances, we select four video contents similar to the query content and show them as presented in Figure 2. The positions of the retrieval results shown in the interface are also determined as shown in this figure.

3.6. Implementation of the Proposed System

By performing the aforementioned procedures, we implement the human-centered multimedia navigation system for video retrieval. In this system, if video contents are added to the target database, scene segmentation of these contents is performed (Part 1). Furthermore, the calculation of distances between scenes obtained by the scene segmentation and those in the database is performed (Parts 2–4). It should be noted that the procedures in Parts 1 to 4 are performed in an off-line environment, and the visualization (Part 5) shown in the previous subsection is performed after those procedures. Naturally, the visualization of retrieval results is performed in an online environment.

As previously described, the implemented system in our approach is equipped with three functions, collaborative use of multimedia data, user's preference extraction, and adaptive visualization. Video contents simultaneously contain multiple data such as visual data, audio data, music data, and text data, and they should be collaboratively used. Therefore, the collaborative use of multimedia data is necessary for satisfying user's demands, and this is closely related to recent multimodal approaches. Furthermore, when the aforementioned approach is implemented into the proposed system, the system should not determine which elements are important or not for retrieving

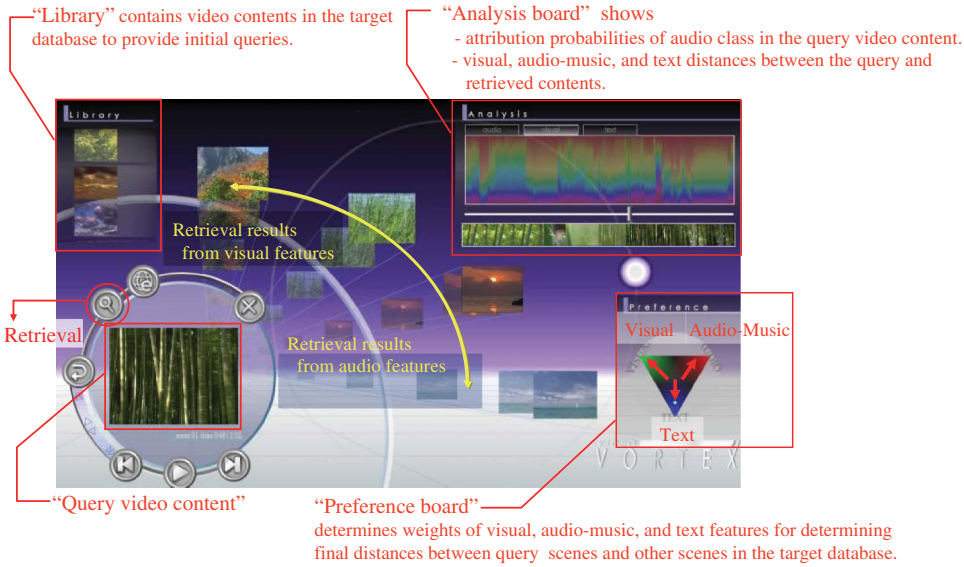

FIG. 1. Interface of the proposed navigation system. A multi-modal approach, which collaboratively utilizes visual, audio, music, and text features, is used in the distance calculation of videos in the proposed system, and an interface that enables extraction of user's preferences leads users to their desired videos (color figure available online).

video contents. This is because the important elements are different for each user. Thus, implementation of user's preference extraction is desirable for improving the performance of video retrieval by our system. In the proposed system, users determine weights of visual data, audio data, music data, and text data as their preferences, but their direct determination is not usually easy. Therefore, we implement the visualization interface equipped with the preference board and enable users to intuitively change these weights. Because the retrieval results adaptively change on the basis of weights extracted from the preference board, users can intuitively understand their own preferences. In this way, the proposed system effectively uses the three functions and realizes a human-centered multimedia navigation system with high flexibility.

3.7. Contributions of the Proposed System

As shown in the previous explanations, the procedures from "Scene Segmentation of Video Sequences" (Part 1) to "Distance Calculation Based on Text Features" (Part 4) are mostly based on our previous works. The main contributions, that is, the most important parts in the proposed system, particularly those in Part 5, are based on the following points.

1. Collaborative use of multimedia data: As shown in Parts 1 through 4, the proposed system can perform distance calculation of visual, audio, music, and text features, respectively. Because scenes within video contents simultaneously have such multiple data (i.e., multiple features), the final similarities or dissimilarities should be obtained on the basis of this characteristic. Therefore, in the proposed system, a multimodal approach is introduced into the retrieval of similar video contents. Specifically, the final distance between two different scenes is derived by merging distances calculated from available features. In this approach, the proposed system is equipped with a useful scheme for judging which features are available by monitoring the attribution probabilities of each shot belonging to shot classes Si, Sp, Mu, No, SpMu, and SpNo as shown in Parts 3 through 5, and distance calculation suitable for the target scenes becomes feasible. It should be noted that although the collaborative use of multimedia data is important for improving the performance of the proposed system, it is not an original contribution of this work. Multimodal approaches have been presented by several researchers, including Babaguchi et al. (2004), Bruno et al. (2008), Calic et al. (2005), and Sudha et al. (2008). Therefore, the multimodal approach in our system is an essential function to improve the performance of the video retrieval, but the basic concept is similar to those of some previous works. Its novelty is thus less than that of the following two points.
2. User's preference extraction: As shown in the first point, we can perform multimodal video retrieval by introducing the collaborative use of multimedia data into the proposed system. It should be noted that the proposed system merges the distances of different features and outputs the final distance between two different scenes. In this scheme, someone must provide weights of features for determining which features should be focused on to calculate the final distance. Generally, when viewing video contents, each user may focus on their visual data, audio data, music data, or semantic data, that is, the media focused on are different for each user. Therefore, in the proposed system, a user's preference

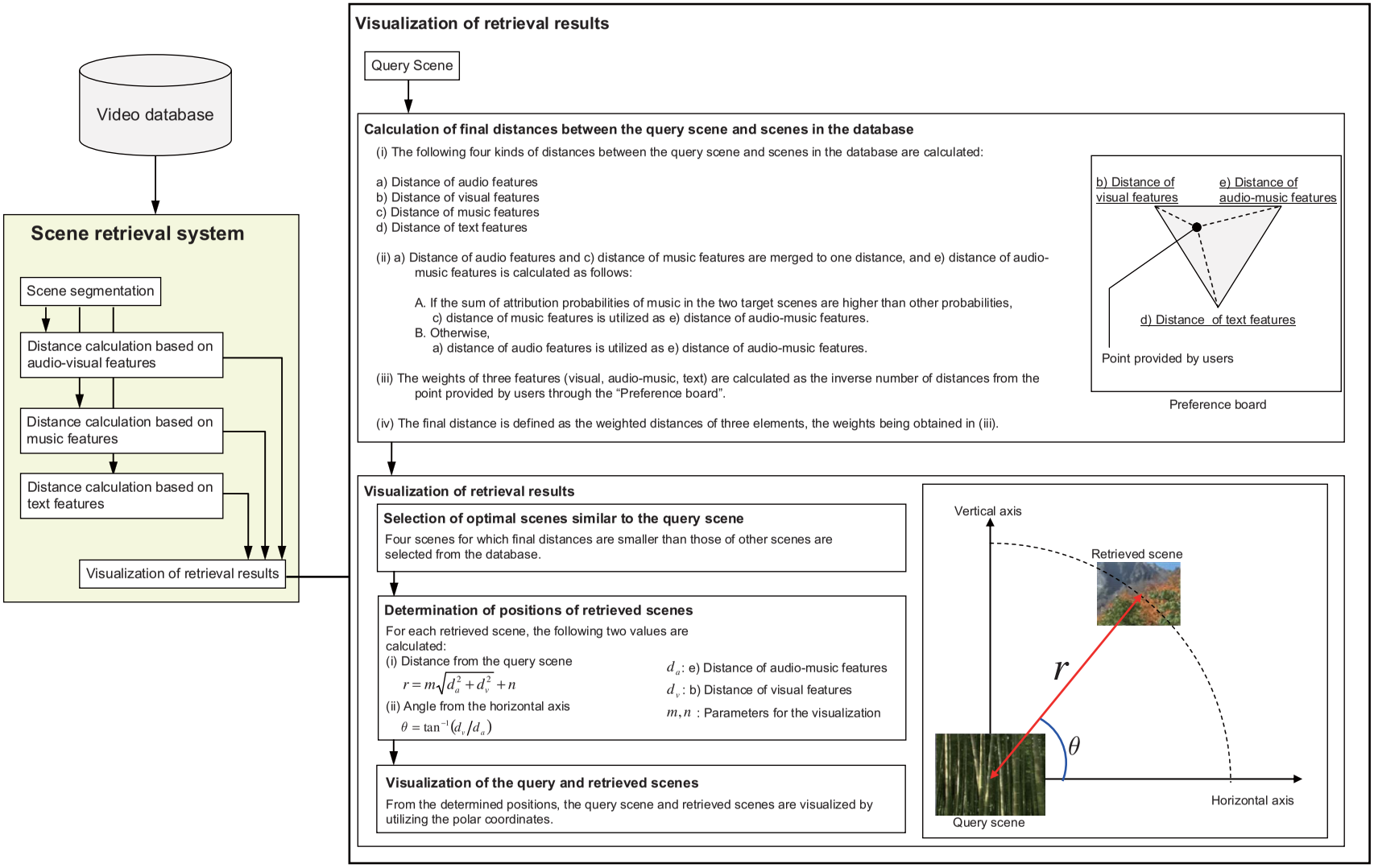

FIG. 2. Specific procedures for visualization of retrieval results. *Note.* The visualization step is broadly composed of two schemes, "Calculation of final distances between the query scene and scenes in the database" and "Visualization of retrieval results." In the first scheme, the final distance is determined from the results obtained by the Preference board. In the second scheme, the scenes are shown on the basis of the final distance calculated in the first scheme and distances of audio-music and visual features (color figure available online).

extraction scheme is introduced into the calculation of final distance between two video scenes. Then this can reflect the user's demands and enable successful video retrieval. It should be noted that in order to obtain the user's preference, the proposed system must be equipped with some functions. Thus, to realize this, we adopt the following novel approach.

3. Adaptive visualization: Adaptive visualization is adopted for realizing the following two points. First, to realize the user's preference extraction, some functions for connecting users and our system are necessary. Therefore, the proposed interface contains the function "preference board" to adaptively extract user's preferences. Second, because retrieval results have to be exhibited for users to be aware of their desired contents, the proposed system adopts the interface shown in Figure 1 to lead users to their desired contents.

4. EXPERIMENTAL RESULTS

To verify the effectiveness of the proposed system, results of quantitative evaluation are shown in this section. Video contents (320×240 pixels, 30 fps, 44.1 kHz, 6.99×10^3 s) that include 277 scenes of several TV programs were utilized in this experiment. We iteratively selected one query scene and retrieved similar ones by utilizing the conventional method (Babaguchi et al., 2004) and the proposed system. A multimodal approach using both visual and audio sequences is adopted in the conventional method, and it is therefore suitable for comparison with our system. In this experiment, we used video contents that were obtained from several news programs to make the evaluation easy. This means several scenes reporting the same topics exist for each query scene, and we can easily determine whether the retrieved results are correct. Thus, we can perform the leave-one-out-based evaluation. It should be noted that in this experiment, the scene segmentation shown in this article was adopted in both the proposed system and the conventional method. For each query, the proposed system and the conventional method calculate the distances from the other scenes and output the retrieval results based on the obtained distances. Note that we implemented the conventional method and conducted a comparison with the proposed system. Furthermore, to perform retrieval by the proposed system, the subject needs to operate the preference board. In this evaluation, one subject participated in the experiment. The determination of whether the retrieved scenes are correct can be easily performed because we use the video contents of news programs. However, the subject uses the preference board to obtain retrieved results, and we have to verify the performance of our system including such subjective parts. Therefore, we add new subjective evaluation using several subjects, and its details are shown later. In the proposed system, the subject had to choose the weights for visual, audio, music, and text in the preference board. The subject randomly operated the preference board, and then the retrieved results were changed. Furthermore, the weighting point in the preference board was determined in such a way that the subject could find the optimal results, that is, scenes reporting the same topics as those of the queries. Thus, from the earlier point, we can see the proposed system has the best retrieval performance because the subject can select the optimal results. This experiment compares the upper limit of the retrieval performance between the proposed system and the conventional method.

As previously described, for each query scene, several similar scenes reporting the same topics exist in the test dataset. Therefore, by only monitoring the retrieval results, we can perform the evaluation by using Recall, Precision, and F-measure. Specifically, we define Recall, Precision, and F-measure as follows:

$$\text{Recall} = \frac{\text{Number of correctly retrieved scenes}}{\text{Number of truly similar scenes}}$$

$$\text{Precesion} = \frac{\text{Number of correctly retrieved scenes}}{\text{Number of retrieved scenes}}$$

$$\text{F} - \text{measure} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Note that the results shown in Table 1 are the average values of the equations just shown. From this table, we can see that the proposed system can provide desired scenes more successfully than can the conventional method. Not only collaborative use of multimedia data but also user's preference extraction and adaptive visualization are adopted in the proposed navigation system. Therefore, more accurate video retrieval was realized as shown in Table 1. As just described, the important elements for retrieving video contents, such as visual data, audio data, music data, and text data, are different for each user. Thus, the conventional multimodal approaches tend not to be suitable for adaptive video content retrieval because their collaborative use of multimedia data is based on fixed procedures not adapting users' preferences. This means that the flexibility of retrieval in those conventional approaches is limited, and video contents provided for users are also quite limited. Furthermore, many attractive approaches to realize accurate image and video analysis for automatically extracting metadata from those contents have been proposed, and representative

TABLE 1
Performance Comparison Between the Conventional Method and the Proposed Method

	Babaguchi, Ishida, and Morisawa (2004)	Proposed System
Recall	0.55	1.0
Precision	0.38	0.89
F value	0.44	0.94

methods have been shown in TRECVID (National Institute of Standards and Technology, 2000). These attractive methods contribute to the improvement of accuracy of the image and video retrieval. On the other hand, we have to tackle another problem, that is, the realization of user-centric navigation. To achieve this application, we have to also concern how systems navigate users to their desired contents. Our system is a useful trial for realizing such human-centered navigation systems.

Next, we show another quantitative evaluation of the proposed system. In Lew et al. (2006), results of experiments in which systems were compared on the basis of content similarities and similarities of random content views are presented for realization of verification systems. We therefore performed similar experiments for verification of the proposed system. We also compared the performance of the proposed system and the performances of systems in two recent works (Bruno et al., 2008; Sudha et al., 2008). It should be noted that the first system in those works (Bruno et al., 2008) also adopts a multimodal retrieval approach. Furthermore, this method performs pseudo-relevance feedback by learning the queries provided from users based on the support vector machine (SVM). Therefore, because this conventional method adopts useful approaches similar to those of the proposed system, it is suitable for comparison in this experiment. The second system (Sudha et al., 2008) also performs relevance feedback and calculates the weights of multiple features from users. Furthermore, the Simultaneous Perturbation Stochastic Approximation technique is adopted for realizing the previous approach. Therefore, the method proposed by Bruno et al. and this method adopt similar schemes and are therefore for comparison with the proposed system.

In the experiments, we used quantitative evaluation based on grouping principles (Wertheimer, 1923). In Gestalt psychology, grouping principles describe the laws which let elements appear to be grouped together. Hence, these principles strongly influence the way in which the components of a networking drawing (i.e., the nodes and links) are organized and perceived as a whole. In Gestalt psychology, the following two laws are referred to as the main grouping principles: (i) The Law of Proximity, and (ii) The Law of Good Continuation.

The Law of Proximity means that elements which are near each other appear to be grouped together. The Law of Good Continuation means that points which result in straight or smoothly curving lines when connected are seen as if they belong together, and the lines tend to follow the smoothest path. The retrieval results gradually vary with repeated retrievals and tend to contain similar videos. Therefore, these videos appear to be grouped together in feature space. According to this tendency, we utilize the Law of Proximity and the Law of Good Continuation on the retrieval time axis to quantitatively evaluate the results provided by the video retrieval.

First, we denote K retrieved scenes in the t ($t = 1, 2, \ldots, T$)-th retrieval time as $f_{t,k}$ ($k = 1, 2, \ldots, K$). In this experiment, we selected one video scene from $K(= 4)$ retrieved results in each retrieval time. Note that these scenes are selected in such a way that the total sum of the distances for the selected scenes' features between neighboring retrieval times becomes minimum. For the following explanation, the set of selected scenes containing scenes $f_{T,k}$ at retrieval time T is denoted as $C_{T,k}$ ($k = 1, 2, \ldots, K$), and the total sum of distances calculated for the set $C_{T,k}$ is denoted by $e_d(T, k)$. The previous selection of the video scene in each retrieval time is performed to obtain retrieval results with grouping characteristics.

Next, we focus on the Law of Proximity, and a new evaluation value is defined as follows:

$$E_p(T) = \frac{1}{K} \sum_{k=1}^{K} e_d(T, k),$$

the value becoming smaller when the features of video scenes become similar between neighboring retrieval times. Therefore, we utilized the criterion just mentioned for the evaluation value representing the Law of Proximity. Furthermore, we focus on the Law of Good Continuation, and the evaluation value is defined as follows:

$$E_c(T) = \frac{1}{K} \sum_{k=1}^{K} e_c(T, k),$$

where $e_c(T, k)$ is the total sum of approximation errors between the video scenes in $C_{T,k}$ aligned at even intervals and their approximation curves. In this experiment, we utilized a second-order Bezier curve for its simplicity. The value just mentioned becomes smaller when the features of the video scenes between neighboring retrieval times vary gradually. Therefore, we utilized this criterion for the evaluation value representing the Law of Good Continuation.

In this experiment, we prepared 754 scenes and performed six tests. For each test, 500 scenes were randomly selected from the prepared 754 scenes, and we iteratively performed retrieval by our system, random content views like (Lew et al., 2006), and the two conventional methods. The two evaluation values $E_p(T)$ and $E_c(T)$ were calculated as shown in Figures 3 and 4. Note that in this experiment, we applied Multi Dimensional Scaling to features of video scenes in order to clearly show the difference between results obtained by using the proposed system and results obtained by using the conventional methods. The results obtained by using the proposed system, random content views, and the two conventional methods are shown in these figures. From the obtained results, we can see that the proposed system tends to have smaller values for these two criteria, and effective retrieval can thus be realized. We have thus confirmed the effectiveness of the proposed system. Furthermore, in this experiment, the conventional method in Sudha et al. (2008) also gave better results than the results of the other two conventional methods. This is related to the retrieval

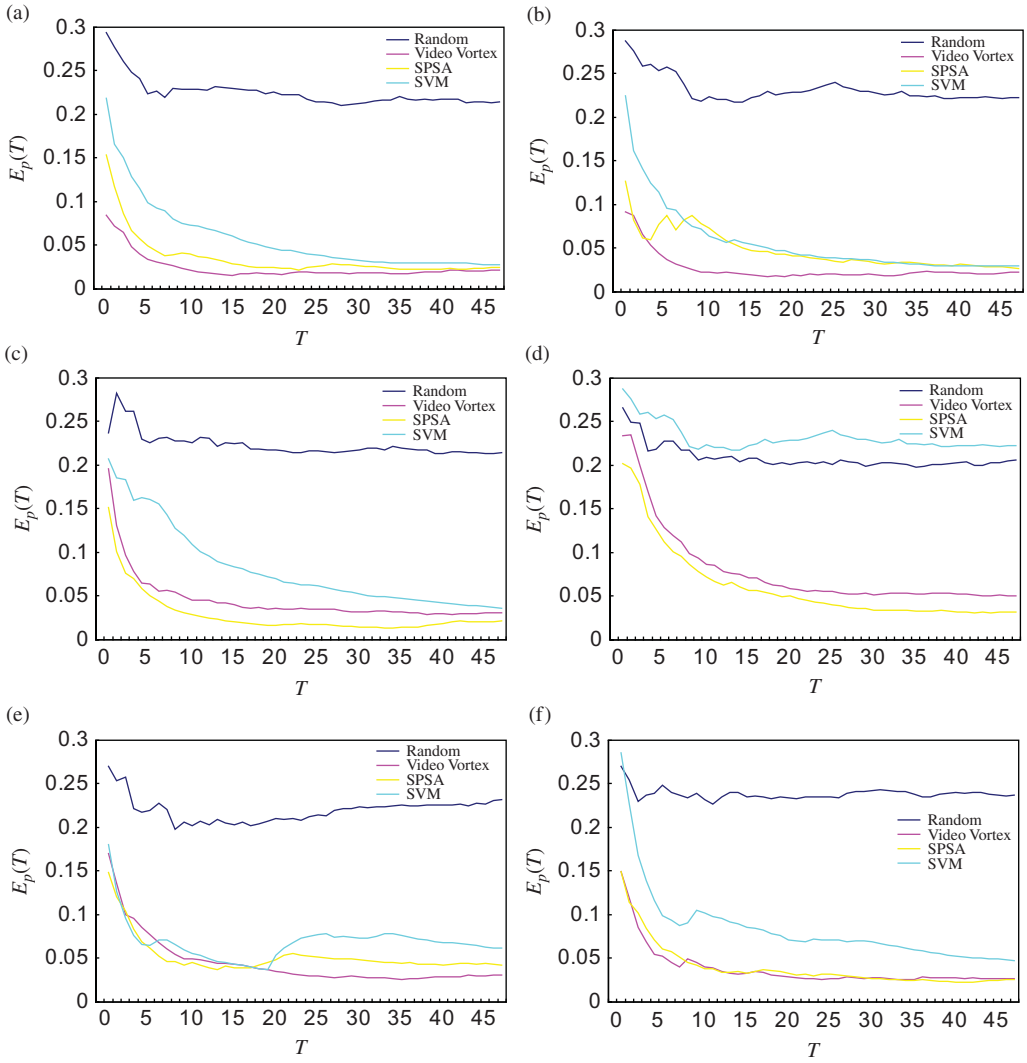

FIG. 3. Relationship between the criterion of the Law of Proximity and the retrieval iteration: (a) – (f) correspond to the results of Tests 1 to 6. *Note.* Video Vortex, Random, SVM, and SPSA, respectively, represent the proposed system, the random content views, and the previously reported two methods (Bruno et al., 2008; Sudha et al., 2008) (color figure available online).

time, and its details are shown in the following subjective evaluation.

From the obtained results, we can see that users can reach their desired contents by iterating the retrieval based on the proposed system. Specifically, from the Law of Proximity, the proposed system can provide similar video contents between neighboring retrieval times, that is, it enables accurate content retrieval. Then the proposed system can faithfully search neighboring contents in their feature space, and this is an essential characteristic for any video retrieval scheme. Furthermore, from the Law of Continuation, the contents provided by the proposed system gradually vary between neighboring retrieval times, and this enables users to steadily move toward their desired contents. Then, even if users do not have specific queries before starting retrieval, the proposed system gradually leads users to their desired contents by iterating the retrieval. Therefore, the proposed system provides a quite different scheme from those presented in the conventional approaches.

Because our work is about a human-centered multimedia navigation system, the effectiveness should be confirmed experimentally (Hartson, Andre, & Williges, 2003). We therefore performed a new experiment to investigate the performance of the proposed system and the performance of the conventional systems from users' verification (i.e., subjective evaluation). We performed subjective evaluation using 11 subjects (User1-User11). Table 2 shows the profiles of the subjects. We used the same video database as that used in the previous experiments. In this experiment, each subject performed video retrieval based on the proposed system and the conventional systems (random content views and the two previously reported methods; Bruno et al., 2008; Sudha et al., 2008), where the specific tasks for the evaluation are shown as follows:

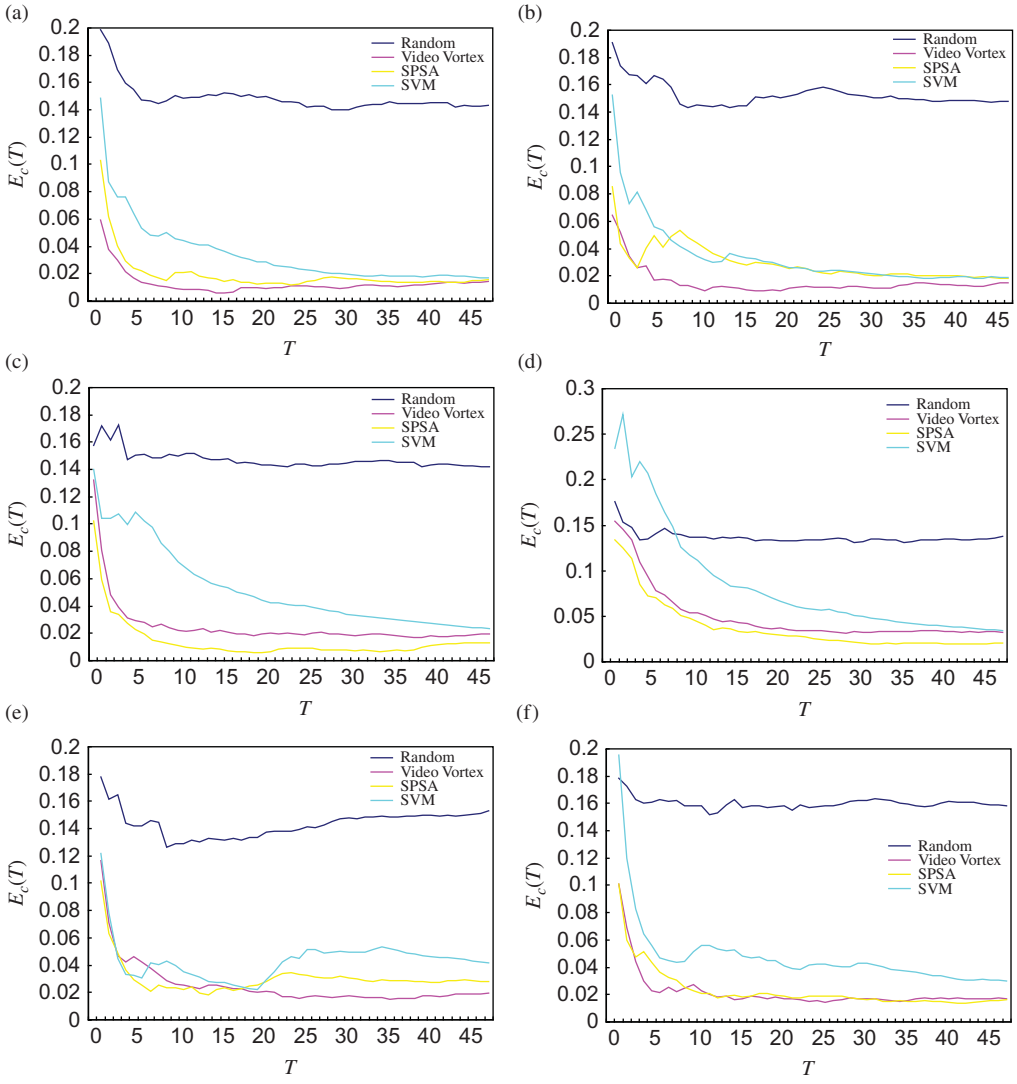

FIG. 4. Relationship between the criterion of the Law of Good Continuation and the retrieval iteration: (a) to (f) correspond to the results of Tests 1 to 6. *Note*. Video Vortex, Random, SVM, and SPSA, respectively, represent the proposed system, the random content views, and the previously reported two methods (Bruno et al., 2008; Sudha et al., 2008) (color figure available online).

TABLE 2
Profiles of the Subjects

Number of the subjects (male/female)	11 (10/1)
Nationality (number)	Japan (11)
Ages (years)	21–26

1. [Tasks for subjective evaluation] Goal video scenes were provided to the subjects.
2. The subjects tried to find the goal video scenes from arbitrary initial video scenes, that is, initial queries, by using the proposed system or the conventional systems.
3. Procedures 1 and 2 were repeated.

After these three tasks, the subjects answered the following questionnaires about each system:
[Q1] Effectiveness (six elements):

- [Q1–1] Could you effectively perform the retrieval? (1–7)
- [Q1–2] Could you obtain expected video scenes by the system? (1–7)
- [Q1–3] Were the retrieved results reasonable? (1–7)
- [Q1–4] Could the system lead you to the goal video scenes? (1–7)
- [Q1–5] Could you make a quick decision for finding new query scenes? (1–7)
- [Q1–6] Could you get a lot of information through the retrieval? (1–7)

TABLE 3
Details of the Results Obtained in the Subjective Evaluation

Criterion	Random Content Views	Proposed System (Video Vortex)	SVM (Bruno et al., 2008)	SPSA (Sudha et al., 2008)
Q1–1	2.45	5.09	5.14	4.82
Q1–2	3.00	**5.41**	4.73	4.77
Q1–3	2.91	**5.36**	4.73	4.68
Q1–4	2.32	5.18	**5.32**	5.05
Q1–5	4.91	**5.14**	4.64	4.45
Q1–6	3.27	3.64	**3.82**	3.64
Q1 (Average of Q1–1 . . . Q1–6)	3.14	**4.97**	4.73	4.57
Q2–1	4.50	**5.64**	5.05	4.73
Q2–2	3.00	**5.09**	4.77	4.14
Q2–3	2.68	4.36	**4.41**	4.09
Q2 (Average of Q2–1 . . . Q2–3)	3.39	**5.03**	4.74	4.32
Q3 (= Q3–1)	4.50	**4.82**	4.32	4.00
Q4–1	6.14	**6.23**	6.09	5.77
Q4–2	5.23	5.27	**5.36**	4.73
Q4–3	3.23	4.32	**5.05**	4.50
Q4–4	**5.50**	5.18	4.73	4.68
Q4 (Average of Q4–1 . . . Q4–4)	5.02	5.25	**5.31**	4.92

[Q2] Efficiency (three elements)

- [Q2–1] Could you intuitively perform the retrieval? (1–7)
- [Q2–2] Can the system be applied to other databases? (1–7)
- [Q2–3] Do you want to use this system for daily life? (1–7)

[Q3] Usability (one element)

- [Q3–1] Was the system useful? (1–7)

[Q4] Satisfaction (four elements)

- [Q4–1] Did the system work as you imagined? (1–7)
- [Q4–2] Was the system frustrating/relaxing? (1–7)
- [Q4–3] Was the system boring/exciting? (1–7)
- [Q4–4] Were you tired by using the system? (1–7)

Scores 1 to 7 in the questionnaires represent from bad to good (i.e., 1 is the worst and 7 is the best). The results of this experiment are shown in Table 3 and Figure 5. From the obtained results, we can see that the proposed system tends to have better performance in terms of effectiveness, efficiency, usability, and satisfaction than the previously reported methods. We also calculated quantitative values while the subjects performed the aforementioned tasks. The average times for finding a new query in each iteration were 8.6 s, 11.0 s, 15.4 s, and 8.3 s in the proposed system, the random content views, and the two previously reported methods (Bruno et al., 2008; Sudha et al., 2008), respectively. Therefore, from the obtained results, retrieval by

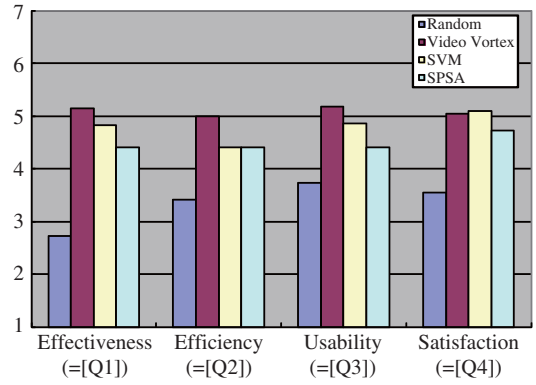

FIG. 5. Results (scores) of the subjective evaluation by the proposed system, the random content views, and the previously reported two methods (Bruno et al., 2008; Sudha et al., 2008). *Note.* Video Vortex, Random, SVM, and SPSA, respectively, represent the four methods mentioned here. Furthermore, the obtained results are average scores of the elements in effectiveness, efficiency, usability, and satisfactions shown in Table 3 (i.e., we show Q1, Q2, Q3, and Q4 of each system) (color figure available online).

the proposed system and the conventional method in Sudha et al. (2008) is faster than that by the other methods. On the other hand, as shown in Table 3 and Figure 5, the proposed method mostly gave better results than not only the random content views and the method in Bruno et al. (2008) but also the method in Sudha et al. (2008). Specifically, although the method in Sudha et al. (2008) perfoms fast retrieval, it cannot output better results of subjective evaluation. Therefore, from the previous discussion, the proposed system can averagely output better results than the conventional methods in terms of retrieval

time and subjective evaluation. Note that as shown in Table 3, it tends to be difficult for the proposed system to output better results in Q4 (Satisfaction). From the obtained results, we can see that because the subjects need to operate the preference board to get their goal scenes in the proposed system, they tend to feel difficulties in these operations. This is also confirmed from the results of the random content views. In this method, users do not perform any annoying operations, and the results therefore tend to be better than those of other questionnaires. Therefore, to solve this point in the proposed system, the user's preference extraction should be performed automatically from their viewing histories. This will be a subject of our future study.

4. CONCLUSIONS

A trial realization of human-centered multimedia navigation for video retrieval is presented in this article. Three functions are introduced into the proposed system in order to realize human-centered navigation. They provide a solution to the conventional problem of not being able to perform retrieval when users do not have specific queries and semantic gaps occur. Successful and efficient retrieval thus becomes feasible by using our system. Experimental results show the effectiveness of the proposed system.

In this article, we only focus on the retrieval of video contents, that is, new video contents are retrieved from other video contents as shown in Figure 1. It should be noted that in order to realize multimedia navigation, it is desirable that various kinds of multimedia contents be retrieved. Therefore, the retrieval of multiple kinds of contents over different media should be realized, and this will be a subject of our future work. Furthermore, user's preference extraction is performed by using operations based on the proposed interface. This approach enables users to reach their truly desired video contents. Nevertheless, because it is desirable that this information be extracted automatically (i.e., from only user's viewing histories), their preference extraction should be realized without using the preference board. This is also a subject of our future study.

REFERENCES

Babaguchi, N., Ishida, T., & Morisawa, K. (2004). Scene retrieval with sign sequence matching based on video and audio features. *IEEE International Conference on Multimedia and Expo (ICME2004)*. pp. 1107–1110.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). SURF: Speed up robust features. *Computer Vision and Image Understanding*, *220*, 346–359.

Bertino, E., Hacid, M.-S., & Toumani, F. (2005). Towards structure discovering in video data. *Journal of Experimental & Theoretical Artificial Intelligence*, *17*, 5–18.

Bruno, E., Moenne-Loccoz, N., & Marchand-Maillet, S. (2008). Design of multimodal dissimilarity spaces for retrieval of video documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*, 1520–1533.

Calic, J., Campbell, N., Dasiopoulou, S., & Kompatsiaris, Y. (2005). A survey on multimodal video representation for semantic retrieval. *Third International Conference on Computer as a Tool (Eurocon 2005)*. pp. 135–138.

Campbell, I. (1996). Applying ostensive functionalism in the place of descriptive proceduralism: The query is dead. *Proceedings of the Workshop on Information Retrieval and Human Computer Interaction*.

Chorianopoulos, K. (2008). User interface design principles for interactive television applications. *International Journal of Human-Computer Interaction*, *24*, 556–573.

Csurka, G., Dance, C. R., Fan, L., & Bray, C. (2004). Visual categorization with bags of Keypoints. *European Conference on Computer Vision*, 1–22.

Dalal, N., & Triggs, B. (2005). Histogram of oriented gradients for human detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 886–893.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., et al. (1995). Query by image and video content: the QBIC system. *Computer*, *28*, 23–32.

Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., et al. (2008). *The diverse and exploding digital universe* (IDC White Paper). Retrieved from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Gantz, J., & Reinsel, D. (2010). *The digital universe decade, are you ready?* IDC iView. Retrieved from http://www.emc.com/digital_universe

Greetha, P., & Narayanan, V. (2008). A survey of content-based video retrieval. *Journal of Computer Science*, *4*, 474–486.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria for evaluating usability evaluation methods. *International Journal of Human–Computer Interaction*, *15*, 145–181.

Huang, J., Liu, Z., & Wang, Y. (2002). Integration of audio and visual information for content-based video segmentation. *IEEE Transactions on Speech and Audio Processing*, *10*, 504–516.

Jaimesa, A., & Sebeb, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, *108*, 116–134.

Jain, R. (2008). EventWeb: Events and experiences in human centered computing. *IEEE Computer*.

Kobayashi, K., & Haseyama, M. (2007). A novel similarity measurement using melody lines for music retrieval. *International Conference on Kansei Engineering and Emotion Research 2007*, *B-15*.

Kooper, R., & MacIntyre, B. (2003). Browsing the real-World Wide Web: Maintaining awareness of virtual information in an AR information space. *International Journal of Human–Computer Interaction*, *16*, 425–446.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions of Multimedia Computing, Communications, and Applications*, *2*, 1–19.

Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing Systems*, *20*, 61–79.

Lowe, D. G. (1999). Object recognition from local scale invariant features. *IEEE International Conference on Computer Vision*, 1150–1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Proceedings of the International Journal of Computer Vision*, *60*, 91–110.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1615–1630.

National Institute of Standards and Technology. (2000). TREC Video Retrieval Evaluation: TRECVID. Retrieved from http://www-nlpir.nist.gov/projects/trecvid/

Nitanda, N., & Haseyama, M. (2007). Audio-based shot classification for audio-visual indexing using PCA, MGD and fuzzy algorithm. *IEICE Transactions on Fundamentals*, *E90-A*, 1542–1548.

Patel, N. V., & Sethi, I. K. (1996). Compressed video processing for cut detection. *IEE Proceedings of Visual, Image, and Signal Processing*, *143*, 315–323.

Picard, R. W. (2003). Applications of affective computing in human-computer interaction. *International Journal of Human-Computer Studies*, *59*, 55–64.

Pikrakis, A., & Kamarotos, D. (2003). Recognition of isolated musical patterns using context dependent dynamic time warping. *IEEE Transactions on Speech and Audio Processing*, *11*, 175–183.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*, 1–47.

Shneiderman, B. (1990). Future directions for human–computer interaction. *International Journal of Human–Computer Interaction*, *2*, *1*, 73–90.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 1349–1380.

Sudha, V., Shalabh, B., Basavaraja, V., & Sridhar, V. (2008). SPSA-based feature relevance estimation for video retrieval. *IEEE Workshop on Multimedia Signal Processing*, 598–603.

Takahashi, S., & Haseyama, M. (2007). Realization of personalized video recommendation based on audio-visual features. *International Conference on Kansei Engineering and Emotion Research 2007*, I-1.

Truong, B. T., Darai, C., & Venkatesh, S. (1996). Improved fade and dissolve detection for reliable video segmentation. *IEEE International Conference on Image Processing*, *3*, 961–964.

Wertheimer, M. (1923). Untersuchungen zur lehre von der gestalt II. *Psychologische Forschung*, *4*, 301–350.

Zhang, T., & Kudo, C. C. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, *9*, 441–457.

ABOUT THE AUTHORS

Miki Haseyama, Ph.D., joined Hokkaido University as an associate professor in 1994, and she is currently a professor at the university. Her research interests include image and video processing and its development into semantic analysis. She is a member of the IEEE, IEICE, ITE, and ASJ.

Takahiro Ogawa, Ph.D., is currently an assistant professor at the Graduate School of Information Science and Technology, Hokkaido University. His research interests are digital image processing and its applications. He is a member of the IEEE, EURASIP, IEICE, and ITE.