Interactive Video Retrieval System
Integrating Visual Search with Textual Search

Shuichi Shiitani, Takayuki Baba, Susumu Endo, Yusuke Uehara,
Daiki Masumoto and Shigemi Nagata

INFORMATION TECHNOLOGY MEDIA LABORATORIES, FUJITSU LABORATORIES LTD.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki City, Kanagawa 211-8588, Japan
{shiitani, baba-t, endou.susumu-02, yuehara, masumoto.daiki, nagata.shigemi}@jp.fujitsu.com

Abstract

In this paper, we propose a technique for efficient retrieval of videos and scenes in large digital video archives. This technique has two features. The first involves displaying several videos at the same time so that users can see many videos in a short time. The second involves detecting cut frame images from a video automatically and arranging them in a virtual 3D space in which similar cut frame images are located close to each other by using self organization map (SOM). Users can search for scenes intuitively and efficiently by only having to look in an area where images that look like the target image are gathered. We conducted experiments to confirm the effectiveness of the scene retrieval technique. The results show that retrieval using this technique is at least twice as fast as retrieval using fast-forwarding.

1. Introduction

The performance of computers continues to improve, and the broadband Internet is becoming increasingly popular. Concurrently, digital video contents are fast becoming the main contents transferred on the broadband Internet. Users can watch a variety of digital media: digital videos, DVD, digital television and so on. Moreover, the quantity of digital video contents increases exponentially.

Since users have to retrieve a video or a scene to edit or to view it, an efficient retrieval technique is necessary for large digital video archives. A general retrieval method is to enter keywords that can find the target video and scene. However, there are two problems with this method. The first is that all videos and scenes must be pre-annotated with keywords to enable efficient retrieval. At present, this has to be done manually. The second problem is that it is impossible to annotate all videos and scenes with keywords completely, because different annotators may use different keywords for the same video or scene. To solve these problems, techniques have been developed where keywords are automatically attached by recognizing speech, and/or telop characters are inserted in the video. However, speech recognition technologies are not yet fully developed, and don’t have enough accuracy to make keywords for retrieval. On the other hand, telop characters are not attached to all video scenes. So these techniques cannot provide efficient scene retrieval based on keywords.

To overcome the disadvantages noted above, we have developed a new multimedia information retrieval system called MIRACLES (Multimedia Information RetrievAl, CLassification, and Exploration System) (Endo et al. 2002), which retrieves multimedia contents using the characteristics of the different types of media. MIRACLES extracts visual features such as color and shape from images and arranges these images in a virtual 3D space. In this space, images that have similar visual features are gathered together. An efficient retrieval is possible, because users guess the area of the target image by seeing the space roughly, and then look for in detail by approaching the area.

This method enables computer-aided information retrieval that otherwise would be difficult to achieve with either machine or manual searching alone. In this paper, we describe the use of this method for video and scene retrieval and present a prototype system that can retrieve videos and scenes efficiently.

2. MIRACLES

Video retrieval techniques other than MIRACLES in which the user watches the content of the image have been researched (Lienhart et al. 1997) and (Boreczky et al. 2000). In these researches, cut frame images are displayed with fixed arrangement and users can select the target scene from them. There are some arrangement methods, order of time, relation graph, emphasizing important cut frame images and so on. These arrangements are very efficient. But users can not select the arrangement according to circumstances.

The characteristic functions of MIRACLES which does not exist in other researches are as follows. 1) Information is collected by a crawler. 2) The collected information is arranged based on similarity. 3) The search is narrowed down interactively. The following sections describe the implementation of each function in MIRACLES.

2.1 Collecting images

MIRACLES collects web pages through the Internet or Intranet by using a web crawler, which is a program that collects web pages.

The web crawler in MIRACLES can collect web pages by keyword. The web crawler passes a keyword to a search engine to retrieve web pages, and then gets the list of URLs, which are returned as search results from the search engine.

The web crawler accesses the web page of the each listed URL, and collects pairs of image and automatically extracts texts which explain the image by analyzing the tag of the HTML document. We call such text “explanatory text”.

The web crawler analyzes the anchor in the pages, follows the anchor, and analyzes the linked pages.

2.2 Arranging images based on similarity

MIRACLES extracts features from the images and the explanatory texts automatically. The system extracts color, shape and texture features from the images. As the color feature, HSI histogram is obtained by counting the number of pixels included in each block, which is the divided HSI color space. MIRACLES uses wavelet coefficients as the shape and texture features. Wavelet coefficients are obtained by breaking down each image into high-frequency elements (shape elements) and low-frequency elements (texture elements). The text feature is based on the frequency of each word in a text and represents the meaning of the text. Each feature is expressed as a vector.

After extracting features, MIRACLES arranges the collected images on a plane so that similar images are located close to each other. The self organization map (SOM) is used for this arrangement. The SOM is a kind of neural network based on unsupervised learning (Kohonen 2001). The SOM maps data in a high-dimensional space to a low-dimensional space while keeping the topological relations between the data in the high-dimensional space.

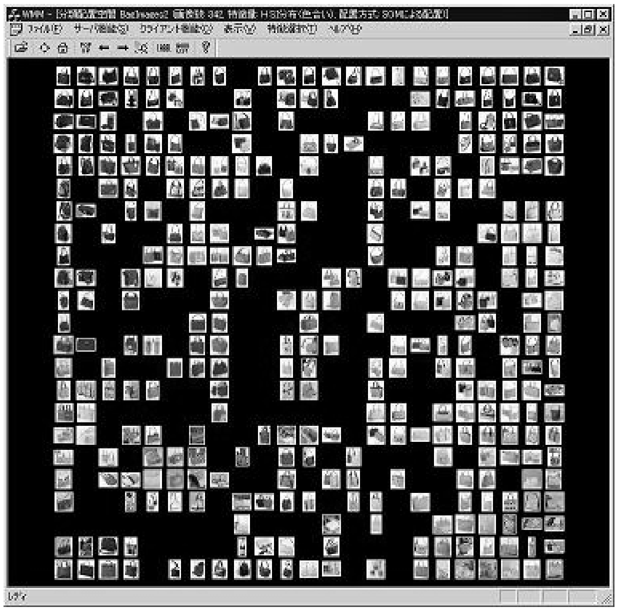

Figure 1: Arrangement of images based on the HSI histogram

The system maps features of images in a high-dimensional space into a 2D plane. The images are put in areas where the image features are mapped.

Figure 1 shows an example of arrangements based on the HSI histogram. It can be seen that bags of the same color are grouped together (e.g., yellow bags have been gathered in the middle of the plane).

2.3 Interactive narrowing down of a search

If the user is looking for a particular red bag out of many bags, as in Figure 1, he or she first finds the area where the red bags are gathered and then looks for the desired bag in that area. To view the red-bag area clearly, the user can fly-through in a virtual 3D space.

In addition, the user can choose the arrangement that is a most appropriate for his/her purpose by changing the features on which the arrangement is based, i.e., shape, texture, or text feature.

MIRACLES can narrow down the search if the user enters a keyword. It moves forward images associated with text that includes the keyword (Figure 2).

When the user selects an image, the system displays web pages that include the image.

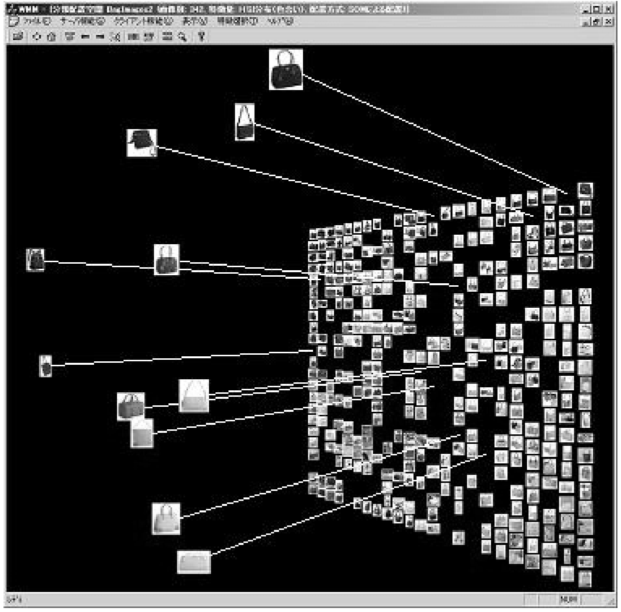

Figure 2: Narrowing down by keyword

3. Video retrieval

Video retrieval is divided into two steps: video content retrieval and scene retrieval. Figure 3 shows the flow of video retrieval. First, the user retrieves a video content from the video archive. Next, he or she retrieves a scene from the video content.

In the preceding section, we described a retrieval method that enables efficient retrieval by viewing a lot of images. This method can also be applied to video content retrieval and scene retrieval. In the following, we explain how this can be done.

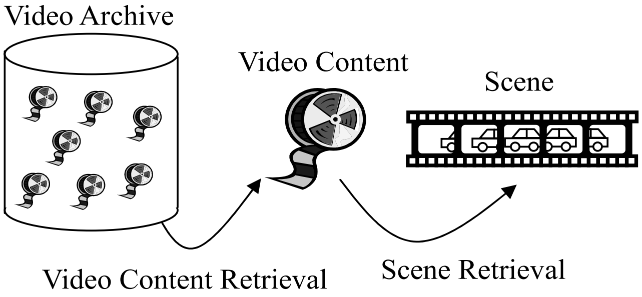

Figure 3: Flow of a video retrieval

3.1 Video content retrieval

MIRACLES displays a large number of images at the same time so that users can search for the images which they want. In a similar fashion, the system displays a large number of video contents at the same time.

We developed two methods for displaying a large number of video contents. The first method is to play several video contents at the same time. The second one is to display a lot of images that represent scenes of a video. In the following section, we describe each method in detail.

3.1.1 Simultaneous playing of video

The first method enables the user to watch several video contents at the same time. At first, the user has to select the videos to play. For example, in the case when all contents are classified into categories beforehand, the user selects a category of the database, and all video contents included in the category are played at the same time.

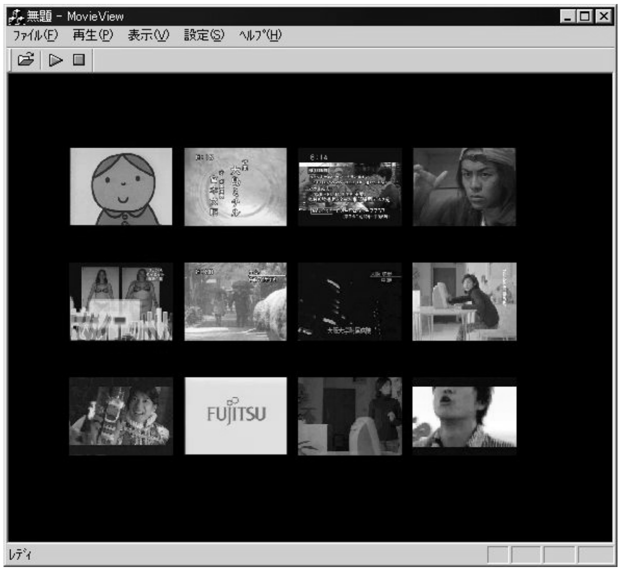

Figure 4: Matrix arrangement of video contents

Users can choose two possible arrangements to display multiple video contents being played. Figure 4 shows a display where different video contents are arranged as a matrix, while they are being played. Users can compare these video contents and select the one they are looking for. Figure 5 shows a display where video contents arranged in a spiral. Compared with the matrix, the spiral can simultaneously display more video contents. The position of each video content changes sequentially, with the images moving along the spiral and new (old) videos appear (disappear) one after another on the screen at the ends of the spiral.

Users can retrieve video content by using this method more quickly than by playing them one by one.

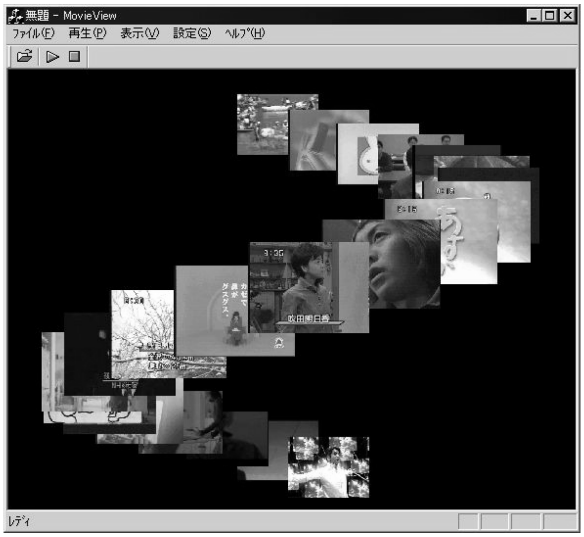

Figure 5: Spiral arrangement of video contents

3.1.2 Representative image

The second method enables users to understand the content of a video by displaying images that summarize the video, because images are suitable for showing to users a lot of information at once. We call such images "representative images".

It is possible to summarize the video content with an image related to it. For example, the poster and the pamphlet of a movie appropriately show the content of the movie.

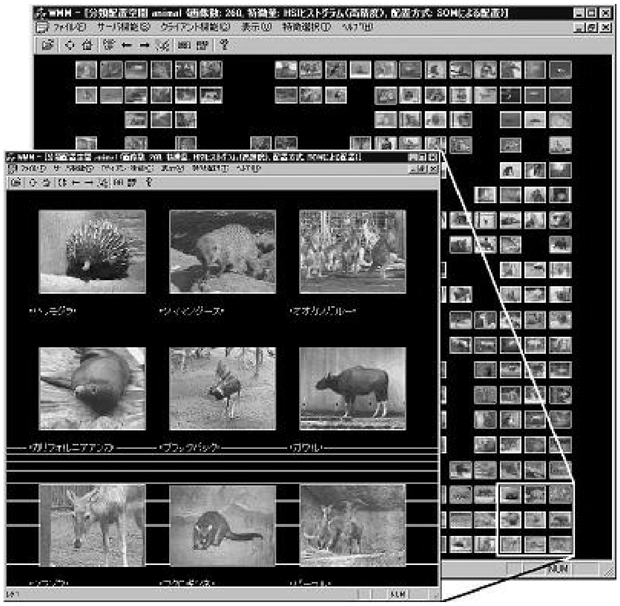

Figure 6: Arranging animal videos based on HSI histogram

Figure 6 is an example of arranging representative images taken from animal videos. Each animal image represents a video content. As representative images, we select the animal images that reflect the video content. These images are arranged by SOM based on the HSI histogram feature.

Figure 7 shows the arrangement of representative images by using SOM based on the text feature. In this case, only the habitat information is included in the explanatory text. Text labels indicate to which habitat the area grouping corresponds. For example, animals that live in the Arctic Ocean are gathered in the center.

For label keywords, MIRACLES chooses words based on the importance of each word. The importance is evaluated based on the frequency of the word. And each label is displayed in the area where the images including the label keyword in the explanatory text has gathered.

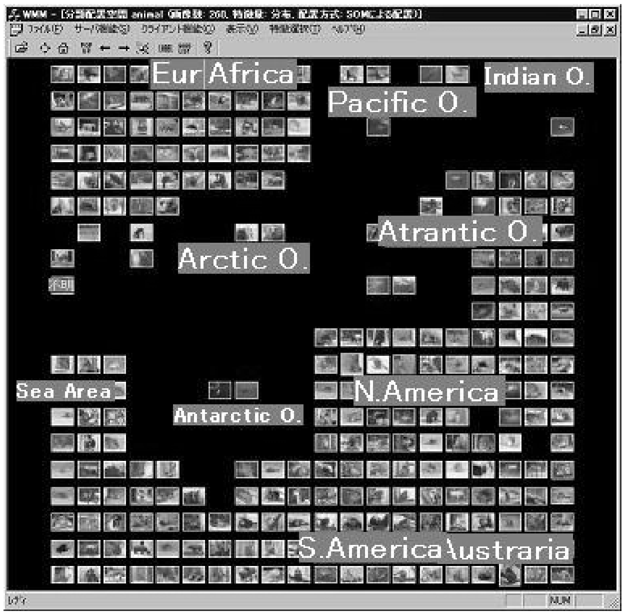

Figure 7: Arranging animal videos based on habitat information of the text

3.2 Scene Retrieval

We developed two methods of presenting scenes of a video to users. One method is to show cut frame images detected in the video content automatically; the other one is to show frame images at regular intervals.

In Figure 8, cut frame images automatically detected in a video content are arranged in order of time. The images were detected using the Chi-squared test of HSI histograms, which divides frame images into 4x4 pieces, calculates difference value between the HSI histograms of each piece of successive frame images, and uses the total of eight difference values from small one as an evaluation measure (Nagasaka et al. 1991). When cut frame images are arranged like this, a user can understand the content after one quick look. He or she can see what scenes compose the video content and in what order these scenes appear. The user does not have to play the video content to retrieve the target scene.

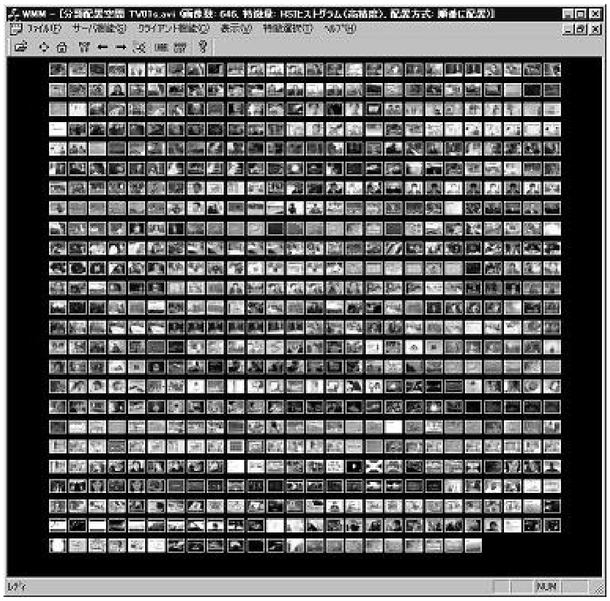

Figure 8: Arranging cut frame images in order of time

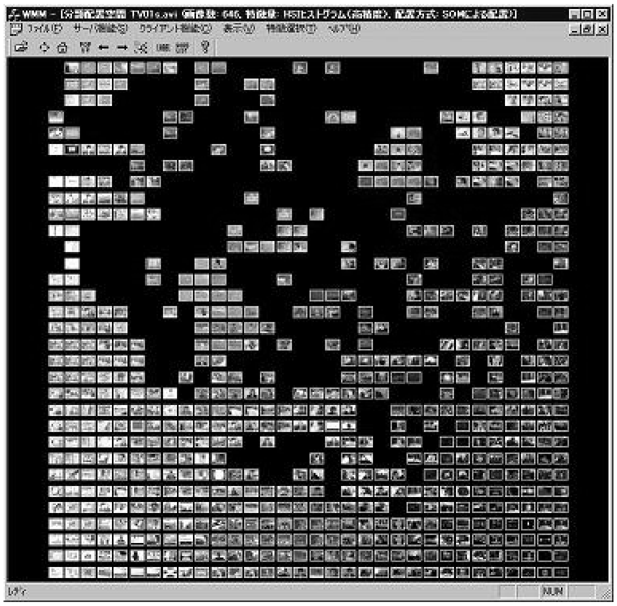

Figure 9: Arranging cut frame images based on HSI histogram

The system can also arrange cut frame images by using an SOM with the HSI histogram feature (Figure 9). In this arrangement, the area in which the desired scene is included can be distinguished.

Figure 10 shows an example of displaying frame images at a regular interval. In this case, the system presents these images like a film. The user can understand the contents of the video and the length of each scene by fly-through in this space. Displaying images in this way allows the user to understand the content of the video more easily than fast-forwarding, because the user can see long sections of the video. The user can efficiently retrieve scenes by choosing a suitable arrangement.

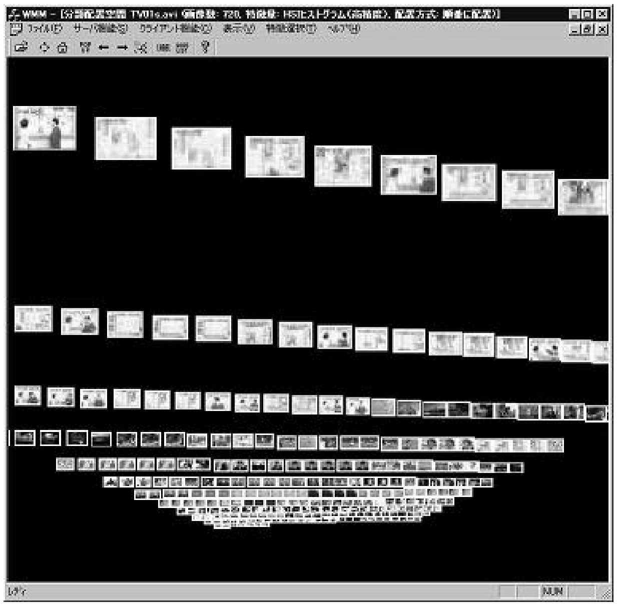
WMM – [分類配置空間 TV01s.avi (画像数: 720, 特徴量: HSIヒストグラム(高精度), 配置方式: 順番に配置)]
ファイル(F)
サーバ機能(S)
クライアント機能(C)
表示(V)
特徴量選択(T)
ヘルプ(H)
レディ
NUM

Figure 10: Arranging frame images at a regular intervals like a film

4. Experiment

We conducted two experiments to evaluate how effective the arrangement using image features are in video scene retrieval.

The first experiment involve searching for the specified scene from a video, supposing three typical situations when searching for videos. We compared our retrieval method of displaying all cut frame images arranged using image features with the retrieval method of playing video by fast-forward. The second experiment compared three arrangements, i.e., using the HSI histogram feature, time order, and random.

4.1. Comparing our method with fast-forward

4.1.1 Method

To confirm the effectiveness of our video scene retrieval technique, we measured the time spent for retrieving the specified scene from a video.

We used a one hour video that was captured from Japanese television broadcasting. It contained a weather forecast, news, and comparatively many commercial films. All cut frame images (646 images) of the video were arranged according to the HSI histogram feature.

The time taken to detect all cut frame images with the Pentium3 700MHz computer was about 14 minutes, the time taken to extract the HSI feature was about 3 minutes, and the SOM's arrangement calculation time was about 1 minute.

The scenes which users retrieve are the following three typical scenes.

A) A scene that can be remembered clearly

It is thought that it should be easier to search for video scene that can be clearly remembered and that have been viewed beforehand. Examples of this situation include ones where the user wants to see the scene of a favorite movie again or a scene of a home video he or she has shot. In the experiment, we showed a coffee commercial film to the users and then they started to retrieve it.

B) A scene of which the color or shape of objects can be remembered

Frequently, the users cannot clearly remember a scene image but can remember its color or shape. For example, a user might want to search for the result of last night's baseball game in a news program. In this case, the users only have to look for a green scene (the turf of the baseball park is green). In the experiment, the users searched for a scene showing 'Ichiro', who is a famous Major League Baseball player. The users could roughly imagine the scene they were looking for.

C) A scene of which the color and shape cannot be remembered

This is the case in which the user wants to see a scene of his or her favorite actor in a television program. However, the user likely doesn't know the scenes' background color or the color of the clothes worn by the actor. In the experiment, the users searched for a scene showing 'Ryoko Hirosue', who is a famous Japanese actress.

4.1.2 Result

Table 1 shows the time that each user spent for the retrieval. 0'51" means 0 minutes 51 seconds. U1 to U4 are the users and FF is an expected time when retrieving a scene from video of one hour by fast-forwarding, which is three times the normal rate (a half of 20 minutes). The reason for using this fast forward speed is that it might be the maximum speed at which a target scene will not be overlooked by viewing the screen. If the user scans a video at a speed faster than three times the normal rate, he or she may easily miss the target scene.

Table 1: Result of experiment 1

	A	B	C
U1	0'51"	1'28"	1'15"
U2	0'30"	4'19"	3'24"
U3	0'44"	2'03"	2'51"
U4	0'35"	4'08"	1'35"
average	0'40"	2'59"	2'16"
FF	10'00"	10'00"	10'00"

These results showed that our cut frame image arrangement method is more efficient than fast-forwarding.

In case A, all users retrieved the scene within a very short time (about 1/10 to 1/20 the time it took for fast-forward searching). The reason for this result is that the users only had to look around an area of the cut frame images. The users could concentrate on the images in the target's vicinity having colors similar to the target image .

In case B, users spent a very long time to retrieve. All users could find a cut frame image of baseball quickly by paying attention to the area where the green images were gathered. But none of these contained images of 'Ichiro,' so the users had to search through almost all the images in the space.

In case C, users spent about three times longer than case A. The users were not able to get an impression as to the target color. The users had to make repeated operations to get near the image where the target person seemed to be, before they could find a cut frame image of the target.

4.2 Comparing arrangements

4.2.1 Method

In the second experiment, we compared the retrieval time of three arrangement displays. We prepared the following three arrangements.

1) Arrangement in which similar color images are gathered (Figure 9)

2) Arrangement in order of time (Figure 8)

3) Random arrangement

We used the same video as the one in the first experiment. The users retrieved the target cut frame image while it was being displayed in another window.

We prepared three kinds of the target cut frame image (Figure 11). Type X image is mostly red. This image stands out and there are few such images. Type Y image is mostly white and there are many such images. Type Z image has various colors and a representative color can not be determined.

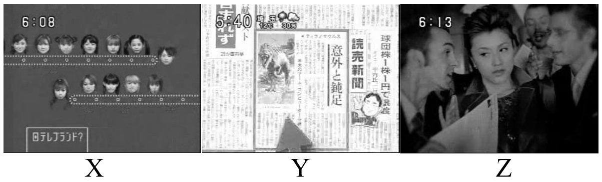

Figure 11: Three kinds of the target cut frame image

First, the users sequentially retrieved three cut frame images arranged using the HSI histogram feature. Next, the users retrieved images arranged in order of time, and retrieved images arranged randomly at last.

4.2.2 Result

Table 2 shows the time that each user spent to retrieve each image.

All users were able to find type X images in about ten seconds regardless of arrangement. This time included the time taken to get near the image and to confirm it. This result suggests that one can quickly find an image that stands out for any arrangement.

Retrieving type Y images spent more time than retrieving type X images. The reason was that the users would often get near an image only to find that it wasn't the target. Retrieving by the HSI histogram feature was the fastest in this case. This result can be seen by noting that images mistakenly thought to be the target will nonetheless be near the target.

The type Z image retrieval showed no clear advantage to using HSI histogram feature. It is thought that the cause of this result is the user's not being able to narrow down the target image by its color.

These results suggest that the arrangement using the HSI histogram feature is very effective when there are many images which have similar color with the target images.

Table 2: Result of Experiment 2

	HSI histogram		
	X	Y	Z
U1	0'07"	0'23"	1'02"
U2	0'14"	0'21"	0'22"
U3	0'07"	0'53"	0'42"
average	0'09"	0'32"	0'42"
	Order of time		
	X	Y	Z
U1	0'06"	1'05"	0'53"
U2	0'16"	2'03"	0'42"
U3	0'07"	0'23"	0'41"
average	0'10"	1'10"	0'45"
	Random		
	X	Y	Z
U1	0'08"	0'46"	1'15"
U2	0'10"	0'22"	0'49"
U3	0'07"	0'59"	0'26"
average	0'08"	0'42"	0'50"

5. Conclusion

We described a method of retrieving videos by simultaneously displaying all videos or all cut frame images of a video. We conducted experiments of retrieving specified scenes from a video and confirmed that our retrieval technique is efficient and practical. In the future, we will have to conduct remaining experiments. One of the experiments is a scene retrieval from very long video. Moreover, we plan to develop a system that analyzes contents of video automatically and calculates semantic features, and we will try to find a new arrangement that is more efficient than the current ones.

References

Boreczky, J.; Girgensohn, A.; Goloychinsky, G. and Uchihashi, S. 2000. An Interactive Comic Book Presentation for Exploring Video, In *Proc. of Conference on Human Factors in Computing Systems (CHI2002)*: 185-192. ACM.

Endo, S.; Shiitani, S; Uehara, Y.; Masumoto, D. and Nagata, S. 2002. MIRACLES: Multimedia Information RetrievAl, CLassification, and Exploration System. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME2002).*

Kohonen, T. 2001. *Self-Organizing Maps*, Springer-Verlag.

Lienhart, R.; Pfeiffer, S. and Effelsberg, W. 1997. VIDEO ABSTRACTING. *J. of Communication of the ACM* 40(12):55-62.

Nagasaka, A. and Tanaka, Y. 1991. Automatic Video Indexing and Full-Video Search for Object Appearances, In *Proc. of 2nd Working Conference on Visual Database Systems*: 119-133. IFIP.