I-041

# Semantic Retrieval for large-scale Video Database by Integrating Image and Text Features

Chuan Yu† Hiroshi Mo‡ Norio Katayama‡ Shin`ichi Satoh‡ Shoichiro Asano†‡

wusen,mo,kayatama,satoh,asano@nii.ac.jp

## 1． Introduction

So far robust content-based image retrieval in the large-scale video/image database is still a challenging problem. In this paper, we explore how user involvement with an image retrieval system using various features of this system can improve performance. As a fundamental job of our proposed news video digest system that gives the ability to retrieve correct parts of news video from a large-scale video database to where user shows the interest, we investigate the effect of feature integration, the keywords in news video, and the assistant of user input.

[1] demonstrates how to structure and retrieve TV news data by producing automatically a table of contents and indices from multimedia data. It consists of audio processing, video processing and information integration. A high-rate retrieval result can be achieved by using information integration which means a kind of method of text keyword based video retrieval on the condition of the system generates a parsed speech database. Different from the way [1] employed in information integration, only nouns in the speeches are extracted and stored as text queries in a relatively simple method. Besides the text keyword query, the image features are get involved in our information integration to achieve the semantic retrieval and improve its retrieval accuracy. We calculate different features of image for boosting retrieval performance in advance at the first step. Then the keyword input plays a great role on image and shots extracting with relative topic. Then the user input indicates which object on the certain image is most interested, and the combination of keywords and features retrieval will be performed to create the search result. Our research tries to give a new answer to solve the problem by integrating image feature and text information.

The rest of paper is organized as follows. Section 2 details the structure of system. Section 3 describes the detail of images features and keyword and how to integrate them to retrieve image. Section4 shows the search result respectively. The last section outlines the method and the future work.

† University of Tokyo,UT

‡ National Institute of Informatics,NII

## 2． Structure of System

The system work flow shows in the figure 1. We first segment news video into video shots and select representative frame for each video, in the meantime, the video keywords for each shots are captured too. Then image features of all image frames are calculated in advance offline before retrieval. Thus the original image database including image, keywords, and features is created.

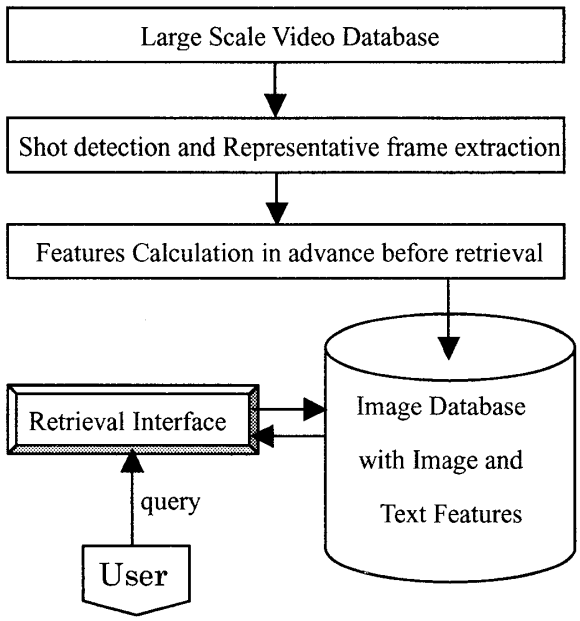


**Figure 1 system work flow**

The retrieval interface accepts the individual or multiple keywords inputted by user and sends the query to image database and returns the retrieved result back to user. User can also choose any interest image randomly selected by interface and retrieve global similar images through single or integrated image features. In addition, the effect of supplying more than one query image is considered. On the condition that user choose more than one image to perform more accuracy retrieval, the distance between image results and any selected image is computed and the result of image cluster is reconstructed.

## 3. Individual and Integrated Features

We employ two kinds of color histogram and edge histogram features to calculate global feature composition of an image.

### 3.1 Image Features

- Color Histogram

The traditional color histogram is very useful because it is invariant to translation and rotation of the images and normalizing the histogram leads to scale invariance. Here we use compute color histogram features in two different ways:

1. The image is divided into $2\times2$ sub-images evenly, and the histogram of RGB of each sub-image is calculated on the axis of one dimensional RGB space that is divided into 16 partitions respectively. Thus there are $16\times3\times4=192$ bins. The color histogram feature is designated as $w_1$.
2. The image is divided into $2\times2$ sub-images too, and the joint histogram of every sub-image is calculated in the three dimensional RGB spaces which is divided into $4\times4\times4$ sub-cube space. Thus there are $64\times4=256$ bins. The cube histogram feature is designated as $w_2$

- Edge Histogram

Edge in the image is considered an important feature to represent the content of the image. Human eyes are known to be sensitive to edge features for image perception. The normative part of the edge histogram descriptor consists of 80 local edge histogram bins[2].

There are five different types of edge in the image, four directional edges and a non-directional edge. Four directional edges include vertical, horizontal, 45 degree, and 135 degree diagonal edges (1)~(5). To detect each of them, we can use five edge filters corrsponing to different edge.

$$Her\_filter = \begin{vmatrix} 1 & -1 \\ 1 & -1 \end{vmatrix} \quad (1)$$

$$Hor\_filter = \begin{vmatrix} 1 & 1 \\ -1 & -1 \end{vmatrix} \quad (2)$$

$$Dia45^\circ\_filter = \begin{vmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{2} \end{vmatrix} \quad (3)$$

$$Dia135^\circ\_filter = \begin{vmatrix} 0 & \sqrt{2} \\ -\sqrt{2} & 0 \end{vmatrix} \quad (4)$$

$$None\_filter = \begin{vmatrix} 2 & -2 \\ -2 & 2 \end{vmatrix} \quad (5)$$

Usually, the image is splited into $4\times4=16$ sub-images, and every sub-image is divided the sub-image into a fixed number of image-blocks. That is, the size of the image-block is propotional to the size of original image to deal with the images with different resolutions.Equations (6) and (7) are used to divide sub-images.

$$x=\sqrt{\frac{subimagewidth\times subimageheight}{desired\ number}} \quad (6)$$

$$block\_size=\left\lfloor\frac{x}{2}\right\rfloor\times2 \quad (7)$$

Here, the image-block is further divided into four sub-blocks. Then, the luminance mean values for the four sub-blocks are used for the edge detection. The next convolution equation (8) is used to extract the number of bin existing in the each sub-image.

$$edge(i,j)=\left|\sum_{k=0}^{3}L_k\times edge\_filter\right| \quad (8)$$

(i,j) stands for the number of image block, $L_k$ is the average of luminance of image block. Using different edge filter, the edge bin can be computed. Since there are 16 sub images and 5 types of edge, therefore the edge histogram has 80 bins. The edge histogram feature is designated as $w_3$.

### 3.2 Keywords

The keywords are related to every shots of news video data. Those keywords are not an absolute copy of announcers voice in news video. The keyword database is automatically generated by taking advantage of JUMAN[3] what composing of all the extracted nouns in caption text of character broadcasting. A typical keywords relating to a shot looks like:

68740;それ:1,今夜談話:1,;藤井総裁:1,;発表:1,;

The keyword composes of time, keywords and the frequency of keywords appear. The query of keyword input by user can generate relative image cluster about certain topic, which can be people, places, or sports and political affairs.

### 3.3 Integrating Features

The keyword composes of time, keywords and the frequency of keywords appear. The ability to extract and describe distinct objects in a complex scene is crucial for image understanding, in particularly it is difficult when deal with a large scale image database captured from the very complex news video data.

The color histogram is suitable for global recognition of image, and the edge histogram is suitable for structure analysis. The aim is to explore if feature integration using (i) color histogram (ii) edge histogram results in better performance than using those features individually. We employ simpler feedback that users

get involved in the retrieval process too to choose one or more query images to improve retrieval accuracy.
The retrieval process by integrated features shows in figure 2.

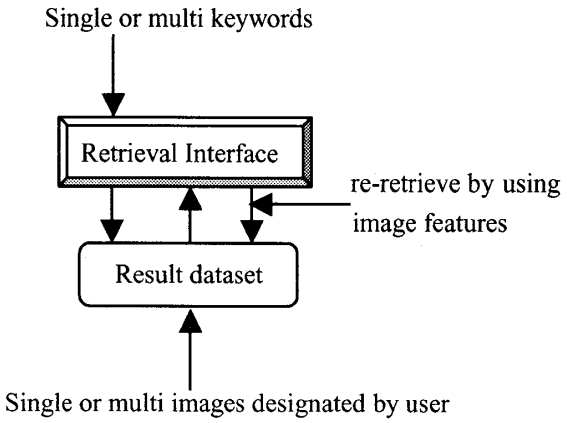


**Figure 2 Integrating Retrieval**

First of all, Interface returns search results according to the inputted single or multiple keywords. Then user can designate one or more image queries to reconstruct the new result dataset by using integrated image features.
The simplest way is user input keyword as search query to retrieve image. User can input more than one keyword to retrieve image, $k_i$ is the result retrieved by $i$th keyword. $\alpha$ is the constant coefficient to control whether the features is to be integrated.

$$W = \prod_{i=1}^{m} k_i \alpha_i \quad (\alpha_i = 0,1) \qquad (9)$$

We use a linear combination (10) of features to complete retrieval via integration of features. $W$ stands for the final retrieved result.

$$W = \sum_{i=1}^{n} w_i \alpha_i \quad (\alpha_i = 0,1) \qquad (10)$$

$w_i$ is the results retrieved by using traditional color histogram, cube histogram and edge histogram respectively, $\alpha$ has the same meaning as above to control how many text queries are involved.
The integration of features is not only among image features or text query respectively, but also can be expand to among image features and text queries (11).

$$W = \prod_{j=1}^{m} k_j \beta_j \ \wedge \ \sum_{i=1}^{n} (w_i \alpha_i) \ (\alpha_i, \beta_j = 0,1) \qquad (11)$$

Since our text database is not absolute one on one map, it means sometimes the retrieved image does not include the object designated by query, in the case, the user assistance is necessary and valuable. User can specify more than one image and retrieve either in individual way or in integrating way. System retrieves dataset firstly by text retrieval, and then compute distance between every image in the dataset and the every query image by using individual or integrated image feature retrieval. Thereafter, the final dataset is created by the completion of iteration of computation of every query image.

## 4 Experiment and discussion

Our image database composes of 45.000 images, keywords and features data calculated in advance. The size of image is $352 \times 240$. The images include all representative frames of 30 minutes video data of news 7 of NHK of half year from 10,2003 to 3,2004.

**Table 1 features computing time**

| Features compute | 2003 10 | 2003 11 | 2003 12 | 2004 01 | 2004 02 | 2004 03 |
|---|---|---|---|---|---|---|
| Time(s) | 603 | 569 | 631 | 660 | 465 | 706 |

The retrieval speed is relatively high since all the images features are calculated in advance. In generally, it takes around 8 seconds to retrieve similar images from 45.000 images using any of images features, and it takes about 2 seconds to retrieve by one keyword input and 4 seconds by 2 keywords input and 6 seconds by 3 keywords.

**figure 3 multi keywords retrieval result**

The figure 3 shows the result in 2 keyword inputs. There are 69 results corresponding to keywords "小泉" and "イラク". And there would be 282 results if we only inputs "小泉".
The figure 4 shows the retrieval result in joint keywords "松井","大リーグ". User maybe not be satisfied with the result since there are too many irrelevant results. The figure 5 and 6 shows the result of integrating image and text features. In the figure 5, user select top left image as query to retrieve similar image by integrating all 3 types of image features, and figure 6 shows the result user select top left 2 images as query

to retrieve similar image by only using cube color histogram. The latter has the better performance than the former, we can point out that it is not for sure integrating more feature means higher accuracy. To different kind of objects or proposes, we should select different strategy of integrating features. But integrating image and text features with the assistance of user input can obviously improve the image retrieval performance, and it makes us closer to do content-based retrieval. If we add object extraction in the image to let user has the capacity to feedback the object in the image, then the retrieval process will be more truly semantic and higher accuracy.

## 5 Summary and future work

Automatic analysis and retrieval of images from a database is a challenging task. The difficulty arises from a number of issues, such as the complex mix of manmade and natural objects in an image. We studied the features and how to integrate them to improve performance, the edge histogram are suitable for object detection and histogram is easy to use in global comparison. It was observed that feature integration enhanced retrieval accuracy in general. And the user assistant also proved valuable. The method provides a useful tool to achieve content-based image retrieval.

The goal is not only to retrieve similar image, underlying this work, we will add precise object extraction on the system, and even to analyze video shot itself. We want not only to achieve image retrieval, but also content-based video retrieval.

## Reference


1. Yasuo Ariki,"Multimedia Technologies for Structuring and Retrieval of TV news, "New Generation Computering, 18 (4), pp341-357,2000
2. S. J. Park, D. K. Park, C. S. Won, "Core experiments on MPEG-7 edge histogram descriptor," *MPEG document M5984*, Geneva, May, 2000.
3. 黒橋 禎夫, 長尾 真,"日本語形態素解析システム JUMAN Version 3.61," 京都大学大学院情報科学研究科, 1999.


**figure 4 retrieval results in 2 keywords**

**figure 5 integrating all 3 image features**

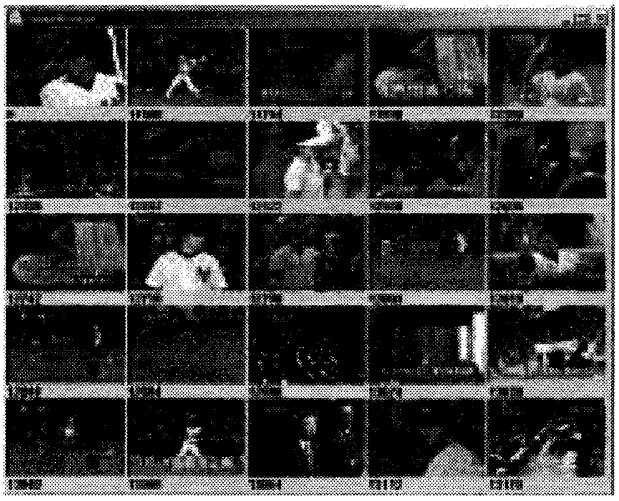

**figure 6 integrating 2 image retrieval**