Relating Graphical Features with Concept Classes for Automatic News Video Indexing

Ichiro IDE and **Reiko HAMADA** and **Shuichi SAKAI** and **Hidehiko TANAKA**
Graduate School of Electrical Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
E-mail: {ide, reiko, sakai, tanaka}@mtl.t.u-tokyo.ac.jp
TEL: +81-3-5841-7413, FAX: +81-3-5800-6922

Abstract

Automatic indexing of video data, especially news videos, is in strong demand considering their contents' importance and value. Various attempts have been made to index news videos automatically in order to cope with this demand, including recent challenges that utilize accompanying textual information. However, most of these methods tend to be textual information driven, which do not thoroughly consider the image contents. We will propose an indexing method, which considers image contents together with textual information, to ensure the consistency of the video contents and the indexes. This is enabled by first, acquiring the relations between graphical features and textual concepts from a large volume of training video data. Next, indexing to incoming video is performed by assuming their contents from the acquired relations, referring to the graphical features. In this paper, we will discuss about relating graphical features with concept classes, which is the key technology to enable such indexing.

1 Introduction

The demand for automatic indexing to video data is becoming stronger in proportion to the increase of the amount. The demand arises from both quantitative and qualitative limits of manual indexing. Especially, news videos consist of important and valuable information that require prompt *i.e.* automatic indexing for recycling and retrieval.

We are trying to perform automatic indexing by integrative use of image data and accompanying textual information. Similar approaches have recently been made by various groups, as prominent in Informedia project's News-on-Demand system [Hauptmann and Witbrock, 1997]. However, they tend to be textual information driven, which do not thoroughly consider the image contents, such as tagging keywords where they appeared, without confirming whether they reflect *'what is actually happening'* in the image. Taking this issue in account, we have proposed an indexing method, which indexes appropriate keywords to news video shots classified into semantically typical shot classes, to ensure the consistency of the contents and the indexes [Ide *et al.*, 1999]. Nakamura and Kanade have also proposed a similar method, which relates image clues and language clues for video contents extraction [Nakamura and Kanade, 1997]. However, both methods have a common problem that the number of classes tend to be small since the classification rules need to be manually described.

Therefore, we are proposing an automatic indexing method based on relations between graphical features and concept classes. The relations are acquired from sample video data, and will be used as clues to estimate the contents' semantics of incoming video. It could not only be used in a similar way as the two 'indexing by shot classification' methods, but also as a clue to roughly estimate the genre of image that lack textual information.

In this paper, we will introduce the overall indexing scheme in Section 2, describe the relating process in Section 3, and conclude the paper in Section 4.

Since term definition vary among researchers, terms related to video structure are defined as shown in Figure 1.

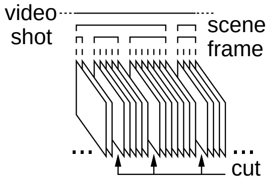

Figure 1: Term definition related to video structure.

- **Frame:** still images that constitute a movie
- **Shot:** group of graphically continuous frames
- **Scene:** group of shots with graphically and/or semantically similar contents
- **Cut:** discontinuous point between adjoining shots

2 Overall Indexing Scheme

The basic idea of the proposed indexing scheme is to tag textual information that matches the graphical contents. This is to ensure the consistency of the contents

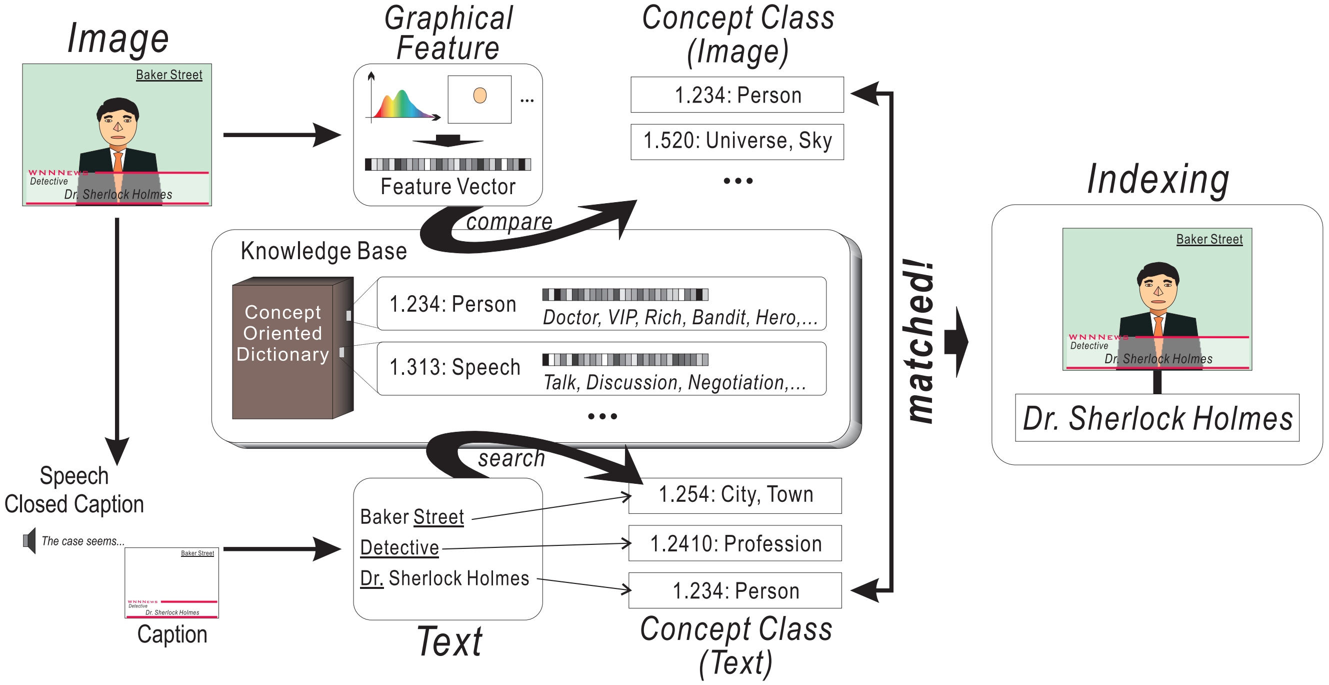

Figure 2: Overall scheme of indexing based on relations between graphical features and concept classes. Concept analysis is performed both to the image and the texts to tag appropriate keywords by selecting texts with concept classes that match that of the image.

and the tagged indexes. To minimize the transition of image contents, concept analysis is performed to shots; minimum units of graphically continuous frames.

Among various textual information sources available from a news video, we employed (open) captions as keyword candidates. This is based on the belief that captions depict most important matters in digestive forms. This frees us from dealing with keyword extraction from redundant texts derived from other sources such as main audio and closed caption. There may be an argument on the frequency and the variety of captions, but most Asian TV news programs have quite a few number of captions with various contents, enough for applying to indexing. As a matter of fact, we counted four to five captions per minute in the sample news programs. Nonetheless, when applying to news programs with less availability of captions, once keyword candidates are extracted from other textual information sources, the same scheme could be applied.

The overall indexing scheme is depicted in Figure 2. Graphical features are extracted from the image, and stored in the form of a vector. The feature vector of an incoming shot is compared with the representative vectors for each concept class to assume the resemblance. Concept classes with high resemblance are assumed to indicate the contents of the incoming shot.

On the other hand, concept classes for (noun phrase) captions are analyzed according to the class which the last noun belongs to. This is done based on the characteristic in Japanese and other major Asian languages, that the last noun tends to represent the semantics of the entire noun phrase.

After these concept analyses, concept classes of image and text are compared. Captions with concept classes that match those of the image are tagged as indexes. Indexing ensuring the consistency of the video contents and the indexes is thus realized, by integrating both graphical and textual information in the concept class level.

3 Relating Graphical Features with Concept Classes

As shown in Figure 2, the proposed indexing method makes use of a knowledge-base consisting of relations between graphical features and concept classes. The relations are not necessarily trivial, thus the key of this method lies in acquiring distinct relations. In this Section, we will describe the relating process, together with the employed graphical features.

3.1 Related Works

As for acquiring relations between graphical features and concept classes, several attempts have been previously made.

Pioneering works have been made in the field of *kansei* engineering or human interfaces, such as Kurita and Kato's ART MUSEUM system [Kurita and Kato, 1993], which relates graphical features with personal visual impressions (mostly expressed by adjectives). Since they deal with personal impressions, the relations are opti-

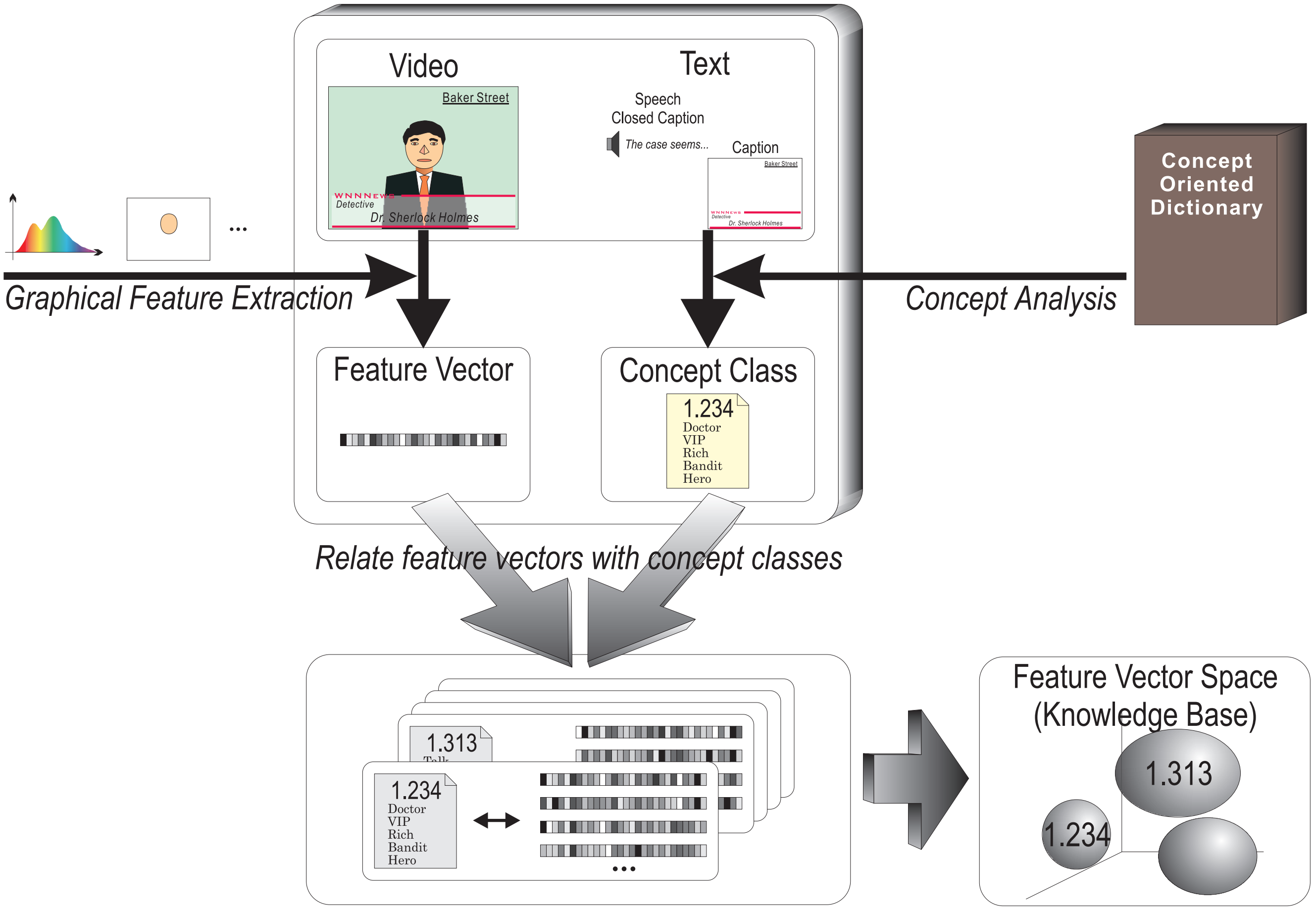

Figure 3: Acquisition of the relations between graphical features and concept classes: For each shot, graphical features are extracted and concept classes are determined by text analyses. After applying the process to a large number of shots, a knowledge base that contains relations between graphical features and concept classes is acquired.

mized for each user, and moreover they are limited to a group of certain adjectives.

As more related works, image and video shot classification has recently been focused upon by various groups. Satou and Sakauchi have developed a typical shot recognition model framework named GOLS [Satou and Sakauchi, 1996]. The framework employs a descriptive rule for shot recognition in order to classify news video, which requires manual description by users, and accordingly limits the classes to specific domains. On the other hand, Huang *et al.* have introduced a hierarchical image classification scheme based on relations between graphical features extracted from collection of images and their titles [Huang *et al.*, 1998]. Likewise, Mo *et al.* have proposed a video shot classification system based on statistically acquired models [Mo *et al.*, 1996]. Both methods are similar to our method from the point of view of automatic classification model acquisition. Nonetheless, they require manual supervision or pre-indexing for the naming of classes, and therefore the granularity of the classes become comparatively vague, since precise manual supervision is burdensome.

As the last and most related work, Mori *et al.* are proposing a text-image combined dual clustering method [Mori *et al.*, 1998]. This method relates a graphical feature vector with explanatory texts from an encyclopedia, so that the input of an unknown image returns texts that explain the contents of a similar image. The idea of creating relations between graphical feature space and textual concepts is very close to our method, but the point that they do not generalize textual concepts makes it somewhat different.

Compared with these related works, the proposed automatic classification model —or, graphical feature - concept class relation— acquisition method is superior in terms of (1)handling concrete contents (expressed by nouns), (2)fine granularity of classes, and (3)automatic naming to the classes.

3.2 Acquiring Relations

Figure 3 depicts the acquisition process. Shots with captions are used for the task. Graphical features are extracted from each shot, and contents of the shot are analyzed by analysis of captions based on the structure of a concept oriented dictionary. After applying the process to a large volume of video, there will be groups of feature vectors (*i.e.* clusters) related with various concept classes. Each cluster will then be statistically analyzed, and mapped on to the feature vector space. Thus is acquired the relations between graphical features with concept classes.

Although acquiring appropriate relations is the key for this method, it would generally be considered quite difficult to relate primitive graphical features with concept classes. Nonetheless, most parts of a news video consist of typical shots for similar topics [Ide *et al.*, 1999], where common graphical features could be expected to be acquired, ignoring the circumstantial diversity among individual shots.

3.3 Graphical Features

Here, we will introduce the graphical features used to relate with concept classes. These features are comparatively primitive, which enables rapid feature extraction.

Color features

As features related to color, (1)histogram, (2)correlogram, and (3)intensity are employed.

A color histogram represents the overall color tone of the image, but does not preserve spatial information at all. On the other hand, a color correlogram preserves local spatial information in a rather abstract way. For example, color correlogram could distinguish an image with a big red circle and another one with red polka-dots, which a histogram could not distinguish if the total red colored areas are equivalent. Consequently, a correlogram and a histogram could be considered as micro and macro color features, respectively. Intensity could be considered as the gray-scaled value of a color. The average intensity of the entire image is used to represent the feature.

In the following definitions, a frame sized m pixels wide and n pixels high is represented as $F(m,n)$, the color step as c_{max}, and the color of pixel p as $c(p)$.

- **Definition of color histogram**
  A color histogram $H(c_i)(c_i = 1, 2, ..., c_{max})$ consists of probabilities of a pixel $p \in F$ to be colored c_i. It is defined as the following equation, and represented as a c_{max} dimension vector.
  $$\begin{aligned} H(c_i) &\equiv Pr\{p \mid p \in F, c(p) = c_i\} \\ &= \frac{\mid \{p \mid p \in F, c(p) = c_i\} \mid}{mn} \end{aligned}$$
- **Definition of color correlogram** [Huang *et al.*, 1997]
  A color correlogram $C_d(c_i, c_j)(c_i, c_j = 1, 2, ..., c_{max})$ consists of probabilities of two pixels $p_a, p_b \in F$ at an interval of d to be colored c_i and c_j, respectively. It is defined as the following equation, and represented as a c_{max}^2 sized two dimension array.
  $$\begin{aligned} C_d(c_i, c_j) &\equiv Pr\left\{ (p_a, p_b) \left| \begin{array}{l} p_a, p_b \in F, \\ \|p_a - p_b\| = d, \\ c(p_a) = c_i, c(p_b) = c_j \end{array} \right. \right\} \\ &= \frac{\left| \left\{ (p_a, p_b) \left| \begin{array}{l} p_a, p_b \in F, \\ \|p_a - p_b\| = d, \\ c(p_a) = c_i, c(p_b) = c_j \end{array} \right. \right\} \right|}{H(c_i)N(d)} \end{aligned}$$
  , where $N(d)$ is the number of pixels at an interval of d from a certain pixel. Here, distance is measured by the 8-neighbor distance, or namely the chess board distance, which makes $N(d) = 8d$.
- **Definition of intensity**
  Intensity I is one of the three axes of the HSI color system. We consider the average intensity of all the pixels in the frame as the overall intensity feature I_F of an image. Intensity c_I of pixel p is derived from the c_R, c_G and c_B values of the RGB color system as follows:
  $$\begin{aligned} c_I(p) &= \max(c_R(p), c_G(p), c_B(p)) \\ I_F &= \frac{\sum_{p \in F} c_I(p)}{mn} \end{aligned}$$

We are currently using a $(m, n) = (320, 240)$ sized 24bit RGB image. The actual size $m \times n$, color step c_{max} and correlation distance d used to extract each feature is shown in Table 1.

Feature	Size	Step	Distance	Dimension
	$(m \times n)$	(c_{max})	(d)	
Histogram	320×240	32	——	32
Correlogram	80×60	16	1, 2, 3, 4	1,024
Intensity	320×240	32	——	1
Total				1,057

Table 1: Facts on the employed color features.

Edge features

As features related to edge, (1)complexity and (2)linearity are employed. These features are extracted from the output of conventional edge detection and linear segment extraction methods. Edge complexity reflects the complexity of the image. It is defined as the ratio of the number of boundary pixels to the number of all the pixels *i.e.* $m \times n$. On the other hand, linearity reflects the overall linearity of the edge segments. It is defined as the number of linear segments with certain lengths in an image. Linearity should appear prominently in images with buildings and other artificial objects.

Facial features

As most news topics deal with human activities, features related to facial region are important. Various methods have been developed to detect facial regions, but primitive ones would be sufficient since facial regions in news videos are usually taken from the front under good lighting condition.

As features related to facial region, (1)number, (2)size, and (3)position are employed. Facial features show different characteristics and play important roles depending on image contents as shown in [Ide *et al.*, 1999].

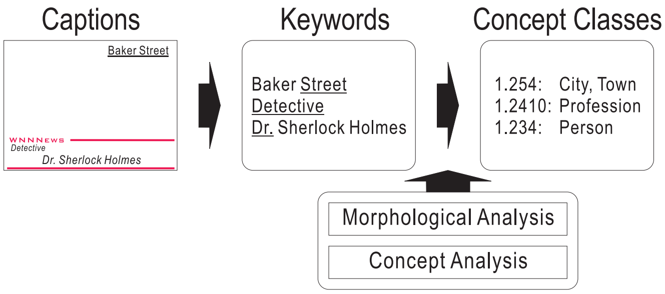

Figure 4: Concept classification according to captions: Caption character recognition is currently not implemented.

3.4 Concept Classification According to Captions

The concept classes of the contents of a shot are classified according to captions that appear in a shot as shown in Figure 4. Contents are classified into one or more concept classes defined in a concept oriented dictionary. Concept analysis to captions is performed in the following manner, based on the fact that in Japanese and other major Asian languages, generally the semantics of a noun phrase is determined by the last noun [Ide and Tanaka, 1998].

1. **Apply morphological analysis to a caption.** Japanese morphological analysis system JUMAN [Matsumoto *et al.*, 1997] was employed for the task. Since Japanese is an agglutinative language, JUMAN's main role for this application is to separate the last noun from the rest part of the noun phrase.
2. **Search for the noun in the concept dictionary.** "The Classified Lexical Table" or "*Bunrui-Goi-Hyo*" [NLRI, 1993]; a Japanese concept oriented dictionary, is employed for the task. This concept dictionary consists of 36,780 words classified in 798 concept classes.

Although it would generally be considered better to give classification standards manually, captions are used as the standard to automate the process. Inappropriate captions are expected to become statistical noise among most of those that depict image contents properly.

4 Conclusion

In this paper, (1)we introduced an automatic news video indexing method, which considers both image and textual contents, and (2)showed the scheme of the relation acquisition process, which is the key technology for realizing such indexing.

We are currently examining methods to analyze and extract characteristics from the feature vectors related to each concept class, and the relation acquisition process by feature vectors extracted from actual news videos.

References

[Hauptmann and Witbrock, 1997] A. G. Hauptmann and M. J. Witbrock. Informedia news-on-demand: Using speech recognition to create a digital video library. In *AAAI'97 Spring Symp. on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, pages 120–126, 1997.

[Huang *et al.*, 1997] J. Huang, S. R. Kumar, and M. Mitra. Combining supervised learning with color correlograms for content-based image retrieval. In *Fifth ACM Intl. Multimedia Conf.*, pages 325–334, 1997.

[Huang *et al.*, 1998] J. Huang, S. R. Kumar, and R. Zabih. An automatic hierarchical image classification scheme. In *Sixth ACM Intl. Multimedia Conf.*, pages 219–228, 1998.

[Ide and Tanaka, 1998] I. Ide and H. Tanaka. Automatic semantic analysis of television news captions. In *IRAL'98 Third Intl. Workshop on Information Retrieval with Asian Languages*, pages 56–60, 1998.

[Ide *et al.*, 1999] I. Ide, K. Yamamoto, and H. Tanaka. *Automatic Video Indexing Based on Shot Classification*, pages 87–102. Number 1554 in Lecture Notes in Computer Science. Springer-Verlag, 1999.

[Kurita and Kato, 1993] T. Kurita and T. Kato. Learning of personal visual impression for image database systems. In *Second Intl. Conf. on Document Analysis and Recognition*, pages 547–552, 1993.

[Matsumoto *et al.*, 1997] Y. Matsumoto, S. Kurohashi, H. Taeki, and M. Nagao. Japanese morphological analysis system JUMAN, 1997. Downloaded from http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html.

[Mo *et al.*, 1996] H. Mo, S. Satoh, and M. Sakauchi. A new type of video scene classification system based on typical model database. In *MVA'96 IAPR Workshop on Machine Video Applications*, pages 329–332, 1996.

[Mori *et al.*, 1998] Y. Mori, H. Takahashi, and R. Oka. Image understanding based on two database composed of images and words allocated in spaces. In *IEICE Fourth Symp. on Intelligent Information Media*, pages 127–132, 1998. (in Japanese).

[Nakamura and Kanade, 1997] Y. Nakamura and T. Kanade. Semantic analysis for video contents extraction —spotting by association in news video—. In *Fifth ACM Intl. Multimedia Conf.*, pages 393–402, 1997.

[NLRI, 1993] National Language Research Institute of Japan NLRI, editor. *Classified Lexical Table [Floppy Disk Edition]*. Number 5 in NLRI Language Processing Data Collection. Shuei Publishers, 1993. (in Japanese).

[Satou and Sakauchi, 1996] T. Satou and M. Sakauchi. A software multimedia platform with real-time video manipulation capability. In *Real-Time Imaging*, volume 2, pages 153–162. Academic Press Ltd., 1996.