Identification of Scenes in News Video from Image Features of Background Region

Ichiro Ide† Reiko Hamada† Shuichi Sakai† Hidehiko Tanaka†

† Graduate School of Electrical Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
TEL: +81-3-5841-7413 FAX: +81-3-5800-6922
{ide,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

Abstract

In this paper, we will introduce a content identification method based on character region segmentation, which will be used for automatic news video indexing. As a result of a preliminary scene identification experiment based on background region image feature analysis, the proposed method showed higher relativeness between the evaluation data and the most related training data.

1 Introduction

As the amount of broadcast video data increases, it is becoming more and more important to store them in a well organized manner, considering their recycling and retrieval. Above all, television news programs are worthwhile indexing considering the importance and usefulness. Currently this process is mostly done manually, but automatic indexing is in big demand both to cope with the increasing amount and to achieve sufficient precision for detailed retrieval.

We are trying to accomplish this task by referring to both video data and accompanying textual data of television news programs. In order to enable high quality indexing for detailed retrieval, simple tagging of keywords as seen in many conventional methods are insufficient, and semantic analysis of keyword candidates is indispensable.

There have been various attempts to automatically index television news video from this approach as prominent in the Informedia project's [14] News-on-Demand system [15]. Nonetheless, although they do satisfy the demand for automatic indexing to a certain extent, their indexing strategies are mostly based on statistics, or just simple occurrence of words and phrases. These strategies do not necessarily ensure the correspondence between the indices and the image contents. Although they may satisfy demands for retrieving a whole news topic, it is not sufficient for retrieving short video segments, where keyword candidates may exist in adjacent segments. Moreover, ensuring such correspondence is crucial, in order to provide indices that depict certain topics occurring in such short segments.

Reflecting this issue, we are currently working on an automatic video indexing system, which performs indexing considering such correspondences. In the image processing part of this system, identification of image contents is required. This paper describes an image feature based video content identification method and show the result of a preliminary scene identification experiment.

1.1 Indexing Reflecting Image Contents

To realize an indexing method that reflects image contents, we are proposing a system that matches keyword candidates and video segments by checking the correspondences of attributes derived from each media, as shown in Figure 1. Since it is extremely difficult to check general correspondences, the attributes are limited to the so-called 4W *i.e.* (1) Who, (2) Where, (3) When, and (4) What. These are surely not sufficient for general video, but would serve for news video where it is natural to expect a query such as, 'I want to see *someone* doing *something* at *somewhere sometime*.'

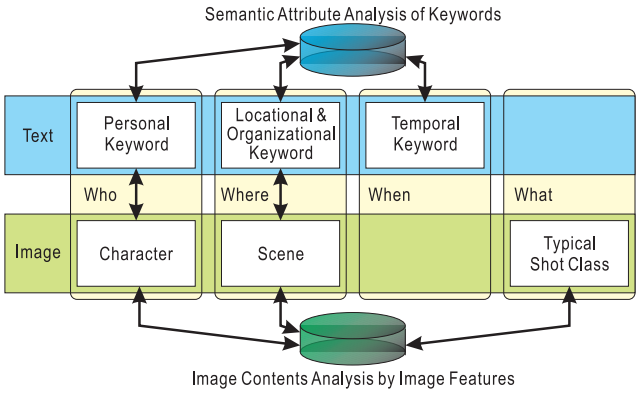

Figure 1: Indexing system overview.

In order to consider such correspondences, it is necessary to employ knowledge on relations between image features and contents. Among the four attributes, we will exclude (3) from the following discussion, since it does not have much to do with image features. Regarding (1), Satoh *et al.* have realized an indexing considering the 'personal keyword - character' relation in the Name-It system [12]. For (4), Nakamura and Kanade [10] and Ide *et al.* [6] have realized shot classification based indexing methods.

Thus, in this paper we will concentrate on knowledge acquisition and image content identification required for considering (2), *i.e.* 'locational / organizational keyword - scene' relation. Although such acquisition and identification would be considered extremely difficult in general, it will be realized by taking advantage of the characteristics of news video, that they tend to be taken under similar conditions for typical contents, *i.e.* location, camera position, angle, and so on.

As for the text analysis, the source text will be provided from (open) captions. Among the captions, index candidates will be limited to noun phrases, and semantically analyzed [4]. We will not discuss further about text analysis in this paper, but analysis applied to 2,546 captions that appeared in 370 minutes of actual news video showed precision and recall of 72.47%, 82.35% for personal keywords, 54.77%, 88.47% for locational/organizational keywords, 41.93%, 93.50% for temporal keywords, respectively.

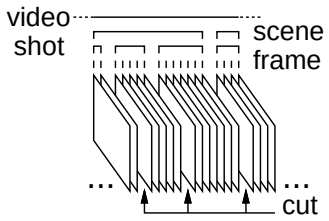

Figure 2: Hierarchical Structure of Video and Term Definition.

1.2 Term Definition

Figure 2 describes the hierarchical structure and term definitions of video. A video consists of still images called *frames*, and a sequence of graphically continuous frames is called a *shot*. The incontinuous point between shots is called a *cut*. The indexing method introduced in 1.1 aims to provide indices to shots. A sequence of graphically and/or semantically similar shots is called a scene. The latter is equivalent to a news topic in the case of a news video.

2 Video Content Identification by Image Feature Analysis

In order to realize an indexing system as described in 1.1, content identification from image features is necessary. In this Section, we will introduce an identification method based on character region segmentation, after citing several related works.

2.1 Related Works: Relating Image Features with Image Contents

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 [7], which relates graphical features with personal visual impressions (mostly expressed by adjectives). Since they deal with personal impressions, the relations are optimized 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 [13]. The framework employs a descriptive rule for shot recognition in order to classify news video, which requires manual classification rule description by users.

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 [3]. Likewise, Mo *et al.* have proposed a video shot classification system based on statistically acquired models [8]. We have also tried to acquire general relations between conceptual classes of (open) captions and image features [5], which showed limited ability. These methods are similar to the current task from the point of view of automatic classification model acquisition based on relatively simple image features. Nonetheless, they seem to be applicable to vague but not concrete classifications, without manual supervision.

As the last and most related work, Mori *et al.* are proposing a text-image combined dual clustering method [9]. 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.

2.2 Content Identification by Character Region Segmentation

As a characteristic of news video, most topics are related to human activities. On this account, there would most likely be a facial closeup of some person in the image. On the other hand, in case of a frequent topic, the background scene (location) of the people tend to be common, even if the people vary from time to time.

Considering such characteristics, it could be possible to identify scenes referring to the image features of background regions, after excluding personal body region. This would be considered extremely difficult in general, but it should be possible to some extent, since news videos tend to cover similar topics under similar circumstances.

As shown in Figure 3, we propose a method that identifies contents based on character region segmentation. The method segments a character region from the background region, in order to identify the individual content, *i.e.* who the character is, where the scene is, and what the whole image is about. Among these identifications, we will focus on identifying the scene in this paper, as previously noted.

Preprocess

The following preprocesses were performed before the identification.

- Digitization of Video
  Digitization of video was done under the following condition:

Spatial Resolution:	320 × 240 pixels
Color Resolution:	24bits (RGB 8bits each) 16,777,216 colors
Temporal Resolution:	15 frames/second
Frame Compression:	JPEG
Video Compression:	none

- Cut Detection
  Among various methods introduced to detect shot boundaries, or cuts, we employed the discrete cosine transformation (DCT) feature method [1]. The method showed 59.78% precision and 92.05% recall to a 370 minute news video with 1,541 cuts. We then manually corrected the automatically detected cuts, in order to evaluate the scene identification method independently.

Region Segmentation

After the shot segmentation in the preprocess, character region segmentation from the background region is performed to the frames in the shots. Although it is desirable to perform the region segmentation to all the frames in a shot, we selected the top one as the representative frame of the shot, to reduce computational costs.

We assumed that a character in news video will be taken from the front in a good lighting condition, so that a fixed model as shown in Figure 5 would detect a character region based on the location and the size of a facial region. As for facial region detection, we employed Rowley *et al.*'s neural network-based tool, namely *face detector* [11]. The method detected every single anchor person in a total of 173 studio shots without mis-detection, and the character regions

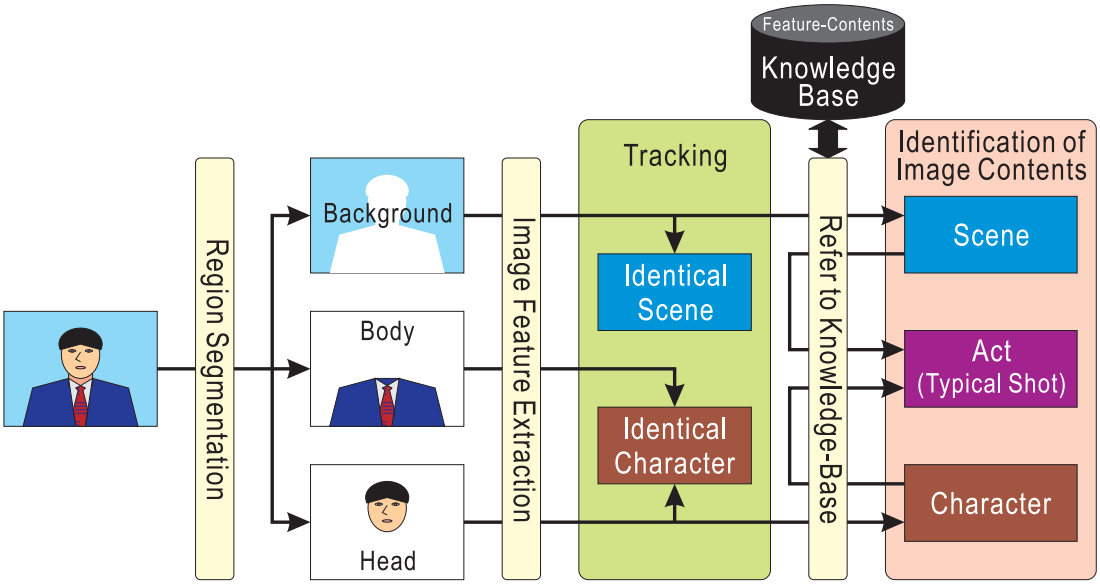

Figure 3: Content Identification by Character Region Segmentation.

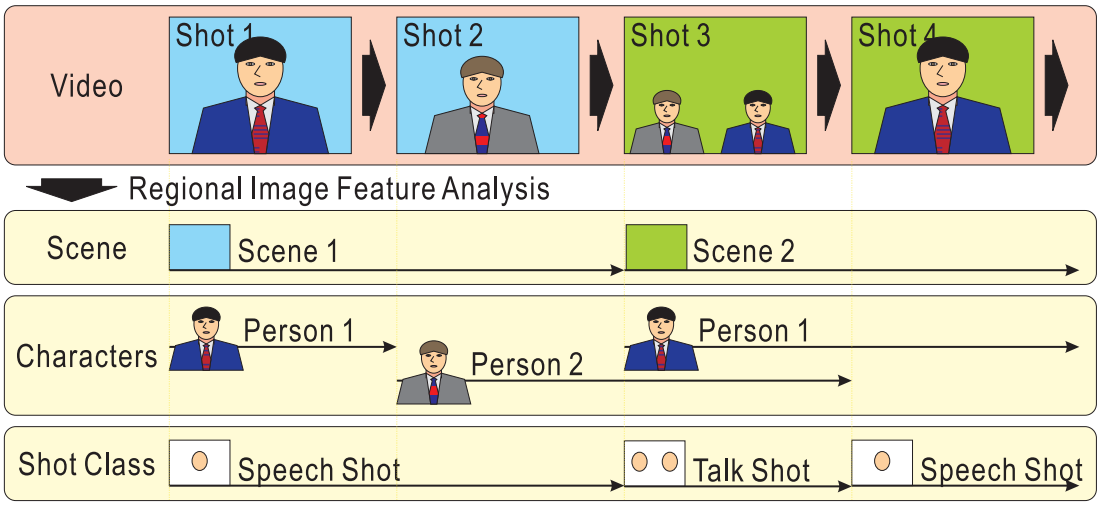

Figure 4: Tracking of Identical Content.

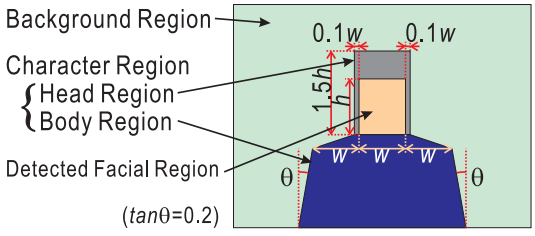

Figure 5: Character Region Estimation from Facial Region.

detected according to the model in Figure 5, seemed to be accurate. Nonetheless, contrary to our assumption, a considerable number of facial regions were either not detected or mis-detected, due to poor lighting condition and diagonal shooting.

We will consider other facial region detection methods and contour extraction methods to improve the ability of automatic character region segmentation in the future. For the meantime, again we manually segmented the character regions from the images used for the experiment.

Image Feature Extraction

After region segmentation comes the image feature extraction. Although image feature of any abstract level could be employed, we will choose relatively simple ones both to acquire general knowledge for robust identification, and to decrease computational costs.

Appropriate features required for identification should differ among regions. We used color related features for background scene identification in the experiment. For facial identification, relative locations between facial parts might be used, or eigen faces may be used as in Name-It [12].

Content Identification Referring to Knowledge Base

Contents of each region will be identified by comparing the feature vector of the region to the vectors in the knowl-

edge base, and evaluating their relativeness. Act identification is somewhat different then the other two. It will be identified by the combination of the identified scene and character, similarly to the previously introduced shot classification methods [6, 10].

Knowledge (*i.e.* relations between image contents and features) could be described by small numbers of representative vectors, but since we were not sure if the same content would form a dense cluster in the feature space, we decided to employ a case based reasoning-like identification method.

Tracking of Identical Contents

As shown in Figure 4, identical contents (scenes and/or characters) may exist across several shots. In order to define an indexing section, tracking of identical contents is necessary. This will be realized by evaluating relativeness of each segmented region among shots. The relativeness will be evaluated as it is done in the identification process.

3 Scene Identification Experiment

In order to evaluate the reality of the proposed method, we performed an background scene identification experiment.

3.1 Image Features Used for the Experiment

The following two color related features were used separately in the experiment.

Color Histogram

Color histogram $H(c_i)$ is the probability of a pixel to be colored in c_i, and is defined as follows:

$$H(c_i) \equiv \frac{\text{Numbers of pixels colored in } c_i}{\text{Total numbers of pixels}} \qquad (i = 1, 2, ..., 64)$$

Color Correlogram

Color correlogram $C(c_{j1}, c_{j2}, d)$ is the probability of two pixels at a distance of d to be colored in c_{j1} and c_{j2}, and is defined as follows:

$$C(c_{j1}, c_{j2}, d) \equiv \frac{\text{No. of pixel pairs at a distance of } d \text{ colored in } c_{j1}, c_{j2}}{H(c_{j1}) \times 8d} \qquad (j1, j2 = 1, 2, ..., 16; d = 1, 2, 3, 4)$$

In contrast with color histogram that represents macro color characteristics, color correlogram represents micro color characteristics. For example, as exemplified in Figure 6, a color correlogram can distinguish a large circle from a group of polka dots of the same area in total, where a color histogram can not.

The maximum values of color resolutions $(i, j1, j2)$ and distance (d) denoted in the equations are those used in the experiment. As a result, (1) 64 dimensional color histogram, and (2) 1,024 dimensional color correlogram vectors were used separately in the experiment. Note that colors were defined by linearly segmenting the RGB color space, and that distance was measured by the chess board (8 neighbor) distance, for convenience.

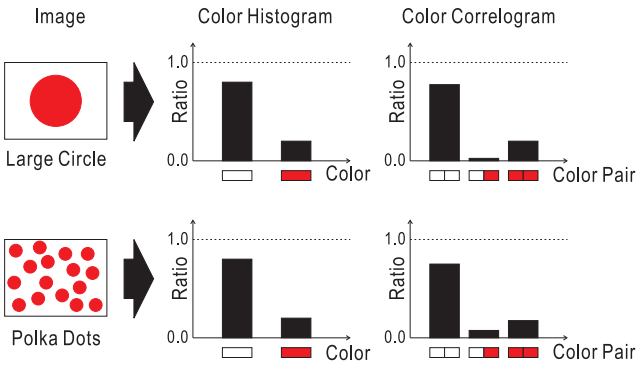

Figure 6: Characteristics of Color Histogram and Correlogram.

3.2 Conditions

Characteristics of Experimental Data

Fifteen 15 minutes news video, a total of 225 minutes was used for the experiment. Although, there were 1,542 shots in the video, we limited the data for the experiment to domestic political shots, since a certain number of data were needed for training. Concretely, (1) Cabinet meeting (27 shots), (2) Parliament (20 shots), and (3) Press briefing (16 shots) were chosen from the source, and one shot from each class was randomly chosen as an evaluation image, which makes the size of training image set to 90 shots.

Relativeness Measure

The following equation was used in order to measure the relativeness f_r of feature vectors between an evaluation image $(\vec{F_e})$ and a training image $(\vec{F_t})$:

$$f_r \equiv \cos\theta = \frac{\vec{F_e} \cdot \vec{F_t}}{|\vec{F_e}||\vec{F_t}|}$$

The equation denotes the cosine of the angle θ between the two vectors $(0 \leq \cos\theta \leq 1)$.

3.3 Result

The result of background scene identification following the conditions is shown in Table 1. There were images with and without large character regions among the training images. In the former case, character regions are segmented from the background region, and in the latter case, the entire image is considered as background region. The evaluation image was selected among the former images. The relativeness was evaluated with and without segmentation to see the difference. Note that (2) Parliament did not have enough sample for performing background region segmentation. Figure 7 shows different types of relativeness comparison performed in the experiment and shown in the result.

As for scene identification from relativeness to the training images, majority of the k-nearest neighbor was inferred as correct. The result shows that especially in the case of (1) Cabinet meeting, comparison with segmentation shows better performance than without segmentation, and correlogram better than histogram. Although we could not observe clear difference in other cases, the value of relativeness measure (f_r) shows clear difference between comparison with and without segmentation.

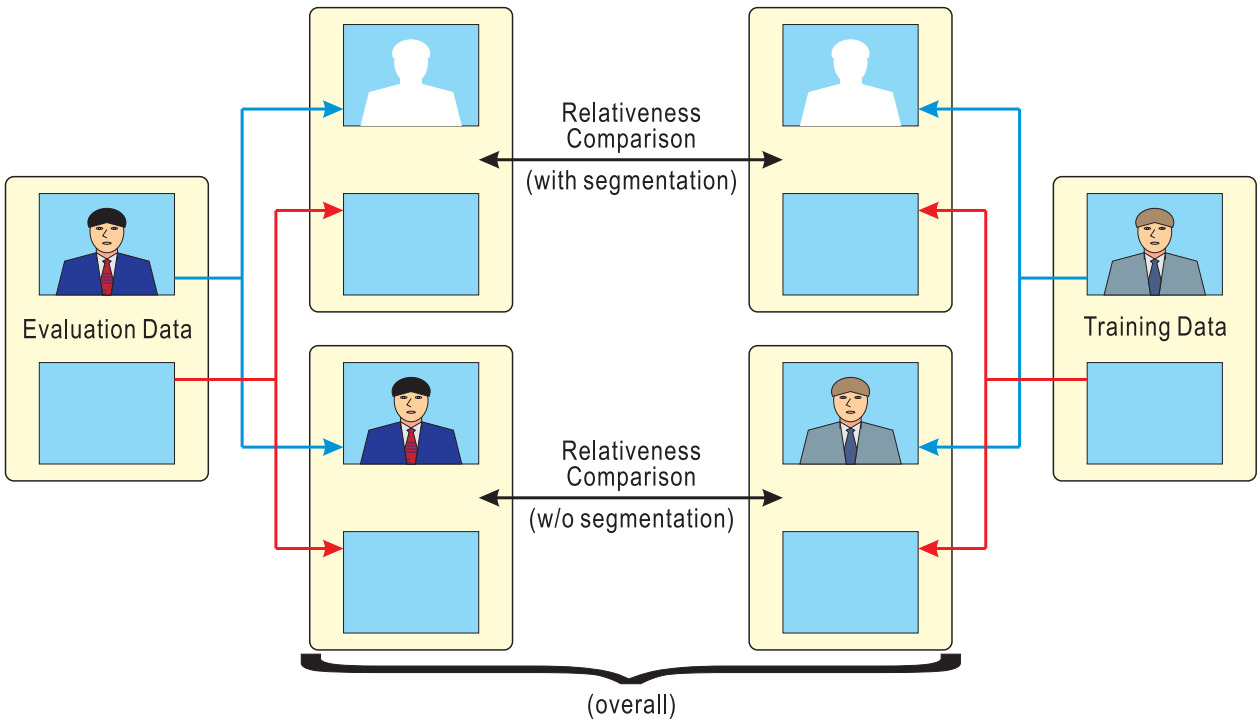

Figure 7: Evaluation of Background Region Resemblance.

Table 1: Result of Scene Identification Experiment: (f_r) shows the relativeness between the evaluation image and the top related training image.

	Scene Classes of Evaluation Images		Number of Training Data	Numbers of Correct Scenes Among Top k Related Images					
				Color Histogram			Color Correlogram		
				k=1 (f_r)	k=3	k=10	k=1 (f_r)	k=3	k=10
(1)	Cabinet meeting	(with segmentation)	17	1/ 1 (0.957)	3/ 3	9/10	1/ 1 (0.936)	3/ 3	8/10
		(w/o segmentation)	17	1/ 1 (0.909)	3/ 3	7/10	1/ 1 (0.905)	3/ 3	6/10
		(overall)	26	1/ 1 (0.957)	3/ 3	10/10	1/ 1 (0.936)	3/ 3	8/10
(2)	Parliament	(overall)	19	1/ 1 (0.988)	3/ 3	9/10	1/ 1 (0.987)	3/ 3	10/10
(3)	Press briefing	(with segmentation)	10	1/ 1 (0.947)	3/ 3	8/10	1/ 1 (0.967)	3/ 3	8/10
		(w/o segmentation)	10	1/ 1 (0.900)	3/ 3	7/10	1/ 1 (0.895)	3/ 3	8/10
		(overall)	15	1/ 1 (0.947)	3/ 3	7/10	1/ 1 (0.967)	3/ 3	8/10

4 Conclusion

We have introduced an indexing method that considers image contents, and proposed a content identification method based on character region segmentation. Among various contents, we focused on background scene identification, and performed a preliminary identification experiment, which showed limited but promising results. As the result of the experiment, although there was not much difference in identification with and without segmentation, relativeness between feature vectors showed higher value with segmentation.

Although there were not much difference between the two image features used in the experiment, it would generally be considered that appropriate features for identification vary among contents. Thus, we will try to weigh features according to their contribution to the contents in the future. We will also employ other features and scenes to evaluate the method under more general condition.

Acknowledgment

We would like to thank Dr. Henry D. Rowley for the permission to use his neural network based face detecting software *face detector* [11].

References

[1] Ariki, Y., Saito, Y.: "Extraction of TV News Articles Based on Scene Cut Detection Using DCT Clustering ", *Proc. 1996 Intl. Conf. on Image Processing*, pp.847-850, Sep 1996.

[2] Huang, J., Kumar S. R., Mitra, M.: "Combining Supervised Learning with Color Correlograms for Content-Based Image Retrieval", *Proc. 5th ACM Intl. Multimedia Conf.*, pp.325-334, Nov 1997.

[3] Huang, J., Kumar S. R., Zabih, R.: "An Automatic Hierarchical Image Classification Scheme", *Proc. 6th ACM Intl. Multimedia Conf.*, pp.219-228, Sep 1998.

[4] Ide, I., Hamada, R., Sakai, S., Tanaka, H.: "Semantic Analysis of Television News Captions Referring to Suffixes", *Proc. 4th Intl. Workshop on Information Retrieval with Asian Languages*, to appear in Nov 1999.

[5] Ide, I., Hamada, R., Tanaka, H., Sakai, S.: "News Video Classification Based on Semantic Attributes of Captions", *Proc. 6th ACM Intl. Multimedia Conf. -Art Demos, Technical Demos, Poster Papers-*, pp.60-61, Sep 1998.

[6] Ide, I., Yamamoto, K., Tanaka, H.: "Automatic Video Indexing Based on Shot Classification", *Advanced Multimedia Content Processing –1st Intl. Conf. AMCP'98, Osaka, Japan–*, Nishio, S., Kishino, F. eds., LNCS Vol.1554, pp.87-102, Springer-Verlag, Mar 1999.

[7] Kurita, T., Kato, T.: "Learning of Personal Visual Impression for Image Database Systems", *Proc. 2nd Intl. Conf. on Document Analysis and Recognition*, pp.547-552, Oct 1993.

[8] Mo, H., Satoh, S., Sakauchi, M.: "A New Type of Video Scene Classification System Based on Typical Model Database", *Proc. IAPR Workshop on Machine Video Applications*, pp.329-332, 1996.

[9] Mori, Y., Takahashi, H., Oka, R.: "Image Understanding Based on Two Database Composed of Images and Words Allocated in Spaces", *Proc. IEICE 4th Symp. on Intelligent Information Media*, pp.127-132, Dec 1998 (in Japanese).

[10] Nakamura, Y., Kanade, T.: "Semantic Analysis for Video Contents Extraction –Spotting by Association in News Video–", *Proc. 5th ACM Intl. Multimedia Conf.*, pp.393-402, Nov 1997.

[11] Rowley, H. D., Baluja, S., Kanade, T.: "Neural Network-Based Face Detection", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.20, No.1, pp.23-38, Jan 1998.

[12] Satoh, S., Nakamura, Y., Kanade, T.: "Name-It: Naming and Detecting Faces in News Video", *IEEE Multimedia*, Vol.6, No.1, pp.22-35, Mar 1999.

[13] Satou, T., Sakauchi, M.: "A Software Multimedia Platform with Real-Time Video Manipulation Capability", *Real-Time Imaging*, Vol.2, pp.153-162, Academic Press Ltd., 1996.

[14] Wactler, H. D., Christel, M. G., Gong, Y., Hauptmann, A. G.: "Lessons learned from Building a Terabyte Digital Video Library", *IEEE Computer* Vol.32, No.2, pp.66-73, Feb 1999.

[15] Wactler, H. D., Hauptmann, A. G., Witbrock, M. J.: "Informedia News-on-Demand: Using speech recognition to create a digital video library", *CMU Tech. Rep.* CMU-CS-98-109, Mar 1998.