Scene Identification in News Video by Character Region Segmentation

Ichiro IDE
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430 Japan
TEL:+81-3-4212-2585 FAX:+81-3-3556-1916
ide@nii.ac.jp

Reiko HAMADA, Shuichi SAKAI,
and Hidehiko TANAKA
Graduate School of Engineering,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
reiko,sakai,tanaka@mtl.t.u-tokyo.ac.jp

ABSTRACT

Reflecting the demand for recycling and retrieval of video, we are proposing an automatic indexing system for news video that considers correspondences between textual indices and image contents. In this paper, we focus on the background image content (*i.e.* scene) identification portion of the system. The analysis is performed by segmenting (human) character region from background region, and was applied to actual news video for evaluation. The overall result showed the effectiveness of the proposed method by 7 to 8%, and indicated that character existence itself is an important feature. Individual observation among various scenes indicated that multiple features should be combinatorily used according to each scene, and that the data set should be exponentially extended for higher performance.

1. INTRODUCTION

1.1 Background

Reflecting the demand for recycling and retrieval of video data, automatic video indexing has become a major research topic as prominent in the News-on-Demand system [9] in the Informedia project. Nonetheless, most general indexing methods simply utilize indices extracted from speech or text in the video, and correspondence between index and image content are rarely concerned except in special applications, such as facial image and human name.

1.2 Indexing considering image contents

Considering the above mentioned issues, we have been proposing an automatic indexing system that considers correspondences between indices derived from texts in the video and image contents. News video is chosen as an application, since it is a rich and valuable information source.

The actual indexing concept is shown in Figure 1. First, alignment of information derived from both text and image in the video are analyzed in four attributes; the so called '4W', *i.e.* 'When', 'Where', 'Who', and 'What'. Such restrictions may seem quite limited for general applications, but could be considered appropriate when expecting queries to news video. Next, correspondence in each alignment is checked to ensure correct indexing. Thus, indexing considering correspondence between textual index and image content will be realized.

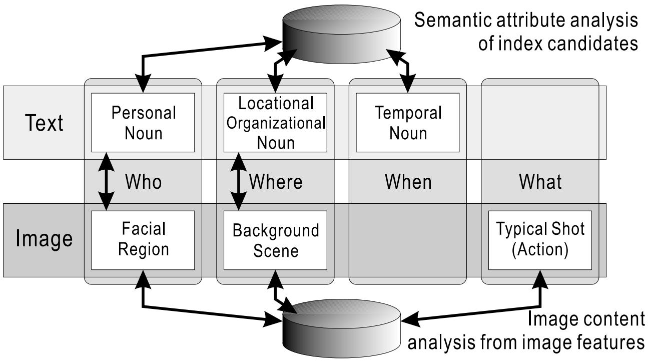
Semantic attribute analysis of index candidates
Text
Personal Noun
Locational Organizational Noun
Temporal Noun
Who
Where
When
What
Image
Facial Region
Background Scene
Typical Shot (Action)
Image content analysis from image features

Figure 1: Concept of the indexing system that considers correspondences between indices and image contents.

In this paper, we will focus on the image content analysis portion of the system, by first introducing the method and next applying it to actual news video for evaluation. The analysis is performed by segmenting (human) character region, if any, and background region, in order to handle them separately.

1.3 Term definition

Figure 2 describes the hierarchical structure and term definition of video components. A video consists of still images called *frames*, and a sequence of graphically continuous frames is called a *shot*. The incontinuous gap between shots is called a *cut*. The proposed indexing system aims to provide indices to shots as minimum units. A sequence of graphically and/or semantically continuous shots is called a *scene*. The latter is mostly equivalent to a topic in the case of news video.

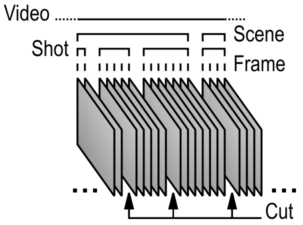

Figure 2: Hierarchical structure and term definition of video.

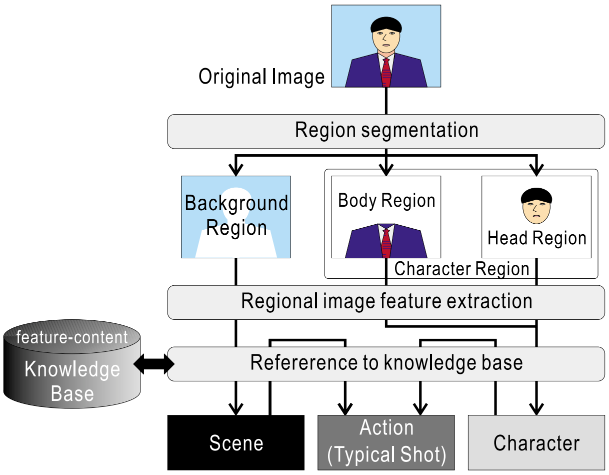

Figure 3: Image contents analysis by character region segmentation.

2. IMAGE CONTENT ANALYSIS BY CHARACTER REGION SEGMENTATION

2.1 Overview

There have been image content analysis methods based on segmentation; mostly segmentation to fixed rectangular blocks in order to apply to general images and objects. However, as news programs focus on providing mostly important information on human activities, there is a big chance that a considerably large (human) character appears in the image. This characteristic provokes the importance to take special consideration in character existence when analyzing news video.

There are image content analysis methods based on typical shot classification that refer to existence, location, and size of characters in news video [2, 6]. We do use this sort of analysis in the proposed indexing system when identifying the action ('What'), but in order to analyze the character ('Who') and the scene ('Where'), it is necessary to segment and separately analyze character region and background region. Moreover, such separate analyses may lead to more detailed analysis of the entire image compared to conventional methods without segmentations or with fixed segmentation.

As shown in Figure 3, we will segment and separately analyze *character region* (consists of *head/facial region* and *body region*) and *background region*, when there is a facial region of certain size and direction. Among the individual analyses, 'Who' is related to various works done in the face recognition field, and moreover there is an automatic news video indexing method that relates facial images and personal names derived from texts in the video [8]. As for 'What', there are rough analysis methods as previously mentioned [2, 6]. Thus we will concentrate on analyzing 'Where' in this paper; first introduce a scene identification method based on image features of background region, and next show an evaluation experiment applied to actual news video. This approach to scene identification should improve the precision by eliminating mal-effects caused by the mixture of image features of character and background regions as in conventional methods.

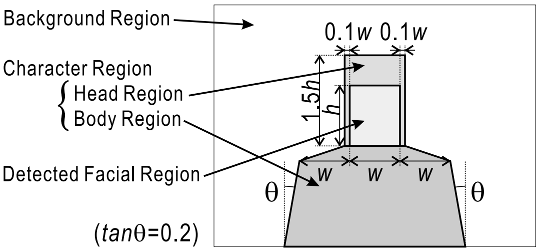

Figure 4: Template for character region estimation from facial region.

2.2 Character region segmentation

In order to realize the analysis as described in **2.1**, character regions should be segmented from the background region.

We applied a template as shown in Figure 4 to determine head and body regions from a facial region. This simple segmentation method was chosen based on the estimation that characters in news video tend to be taken (1)under good lighting condition, and (2)from the front. As for facial region extraction, we utilized a neural-network based tool; 'the Face Detector' [7] developed at CMU.

Although the estimations were applicable to a certain extent, not all situations allowed their appliance (especially estimation (2)). On the other hand, in the case of anchor people in a studio, the estimations were completely applicable, which resulted in complete detection of their facial regions and character region segmentation based on the template. Thus in the following experiment, character regions were manually segmented, except for anchor people in a studio.

3. SCENE IDENTIFICATION REFERRING TO IMAGE FEATURES OF BACKGROUND REGION

After character region segmentation, content analysis, *i.e.* scene identification of background region is done referring to knowledge on relations between contents (scenes) and image features. In order to realize the identification, such knowledge should be acquired beforehand.

It may seem extremely difficult to acquire such knowledge from general images. Nonetheless, since news video tend to focus on specific events periodically or intensively, if a data set of a certain size is available, it should be realistic

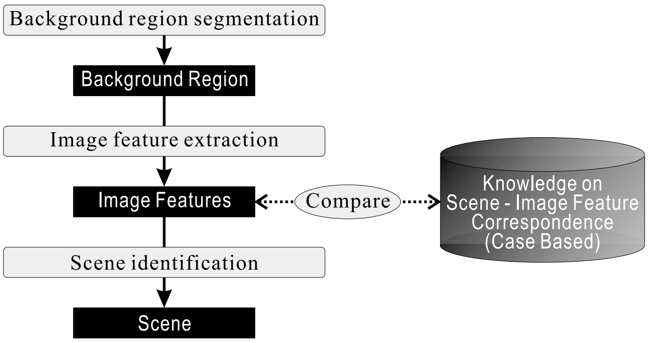
Background region segmentation
Background Region
Image feature extraction
Image Features
Compare
Knowledge on Scene - Image Feature Correspondence (Case Based)
Scene identification
Scene

Figure 5: Scene identification process.

to acquire knowledge on relations between frequent scenes and their image features. Under such expectation, in this work, scenes are identified by a case-based approach where cases are correspondences between scene names and image features. The simple case-based approach was taken in this case since the size of the data set was relatively small. Alternative approaches should be taken when applying to a larger data set.

3.1 Related works: Content analysis based on image features

Before going into details of the method, we will introduce related works on content analysis based on image features.

As early works, Kurita *et al.* created a correspondence map between impressional adjectives and image features statistically acquired from user studies for an art museum database [3]. Such methods are often used in the field of cognitive engineering, but problems and their solutions differ in the case of acquiring relations with concrete objects, which is necessary to handle news video.

Conventionally, when handling concrete objects, the relation was described by a precise (sometimes 3 dimensional) graphical model of an object, and/or by extremely limiting the object domain. As a fairly general model, Mori *et al.* are proposing a method that relates annotations and image features of pictures provided from an encyclopedia [5]. They first cluster words according to general co-occurrences, and next form image clusters in an image feature space based on the similarity of annotations derived from an encyclopedia. This enables the retrieval of annotations and terms related to a query image, but the target is too general to achieve realistic performance.

On the other hand, Mo *et al.* are proposing a sports video classification method based on relations between clustered training images and sports genres [4]. Although their aim and case-based approach is similar to the proposed method, they identify the genres from the entire image and do not look into the semantic structure of images, which may lead to decrease in performance.

3.2 Scene identification process

Scene identification is performed following the process described in Figure 5:

Table 1: Numbers of images in the pre-classified scenes.

Pre-classified Scenes		Character Region		
		With	Without	Total
1)	Cabinet meeting	22	10	32
2)	Parliament	31	21	52
3)	Press conference	11	6	17
4)	Court	6	23	29
5)	Studio	231	0	231
	Others	124	332	456
Total		425	392	817

1. *Extract character region:*
   Extract character regions if characters with certain size and face direction are found in the image.

2. *Extract image features:*
   Segment the extracted character regions apart and extract image features from the background region. When there is no character region, treat the whole image as background.

3. *Measure similarity:*
   Measure the similarity of the features between the test data and all the other pre-indexed images (training data) in the database.

4. *Identify scenes:*
   Identify scenes referring to the indices annotated to the top rated similar images.

4. SCENE IDENTIFICATION EXPERIMENT

4.1 Conditions

Following are the conditions under which the scene identification experiment was performed.

4.1.1 Image source

The first frames from 817 shots selected from twenty 15 minutes news video recorded in two different periods through a year were used in the experiment. Several frequent scenes were selected in order to see the identification ability depending on different situations. Correct scenes were manually given beforehand in order, and only, to evaluate the results. The selected frequent scenes were the following four domestic scenes:

1. *Cabinet meeting*
2. *Parliament*
3. *Press conference*
4. *Court*

and

5. *Studio*

as a special scene.

Table 1 shows the numbers of cases in the data set that belong to each frequent scene. Note that when less than

Table 2: Conditions of video capturing.

Spatial resolution	320 pixels wide × 240 pixels high
Color resolution	24 bits (8 bits each for R, G, B)
Temporal resolution	15 frames per second
Frame compression	JPEG
Video compression	None

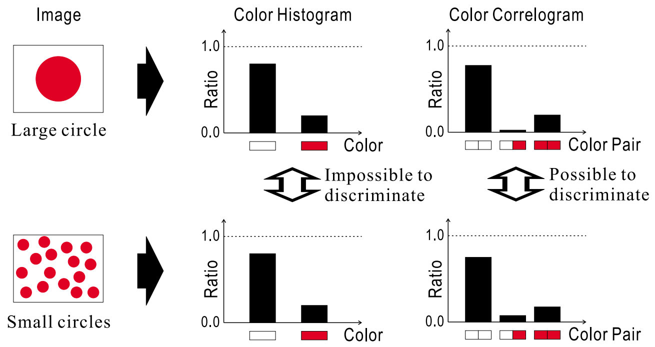

Figure 6: Difference between color histogram and correlogram.

three large and frontfaced characters existed, character region segmentation was applied (Indicated as 'With'). None or more than four characters were considered as part of the background region (Indicated as 'Without').

The images were captured from an VHS recorded analog broadcast video, and digitized at the lowest JPEG compression rate, frame by frame. Detailed capturing conditions are specified in Table 2.

4.1.2 Image features

Among various image features, the followings were used:

1. *Color histogram:*
   Normalized distribution of color frequency in an image. In this experiment, we quantized the RGB represented color space into 64 blocks linearly. Thus the histogram is represented as a 64 dimensional vector.

2. *Color correlogram*[1]*:*
   Normalized distribution of color pair frequency of pixels in certain distance. In this experiment, we quantized the RGB represented color space into 16 blocks linearly and set the distance from 1 to 4. Thus the correlogram is represented as a 1,024 (= 16 × 16 × 4) dimensional vector.

These features have different characteristics as shown in Figure 6. Histograms reflect macro color features and correlograms reflect micro features. We chose these features since it seems that color characteristics play an important role when human beings identify scenes roughly, and also to compare the behavior of two different features between different scenes.

In the experiment, each feature was used independently to see the difference in their performance, but integrative use of features including other features is necessary in order to discriminate more scenes.

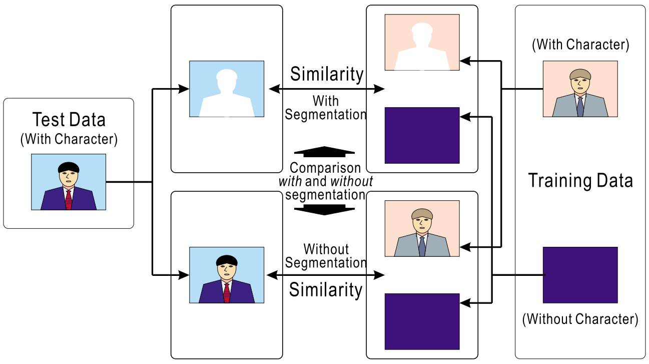

Figure 7: Similarity comparison with and without segmentation.

4.1.3 Similarity measure

As a measure to evaluate the similarity between image feature vectors, we used the cosine of the angle θ between two vectors, $\vec{F_1}$ and $\vec{F_2}$, formulated as follows (ranges between 0 and 1):

$$\cos\theta = \frac{\vec{F_1} \cdot \vec{F_2}}{\mid \vec{F_1} \mid\mid \vec{F_2} \mid}$$

4.1.4 Evaluation measure

In order to evaluate the scene identification ability, we used the following measure:

- *Percentage of images that majority of the top n similar ones were correct (n = 1, 3, 5, 7, 9):*
  We considered that if the majority (*i.e.* more than half) of the scenes of the top n similar images were identical to the correct scene (This was manually given beforehand for evaluation purpose) of the test data. Note that when $n = 1$, the results show the percentages of images that the scenes of the most similar ones were correct.

4.2 Procedure

Following the procedure and conditions described in **3.2** and **4.1**, the actual experiment was performed as follows:

1. *Extract character region:*
   Extract character region manually if characters with certain size and face direction are found in the image. As an exception, anchor people in a studio were completely automatically extracted using *the Face Detector* and the template shown in Figure 4.

2. *Extract image feature:*
   Create both color histogram and correlogram of the segmented background region. When there is no character region, treat the whole image as background.

3. *Measure similarity:*
   Measure similarity between the test data and all other images in the data set.

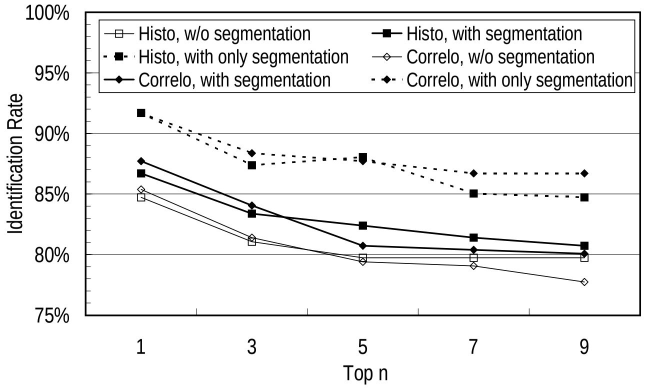

Figure 8: Overall identification rate.

4. *Identify scenes:*
   If the majority of the scenes of the top n similar images were identical to the correct scene of the test image, consider it as correctly identified.
   Apply 3. and 4. to all the images in the data set, and evaluate in a cross validation manner.

Note that as shown in Figure 7, in order to evaluate the effect of the proposed character region segmentation, we performed steps 2. to 4. without any segmentation even to images with character regions, along with the steps shown above.

Thus the experiment was performed under three different similarity comparison conditions and two different features:

- *Similarity comparison:*
  - Comparison without any character region segmentation (Conventional method)
  - Comparison with character region segmentation including images without characters in the training data set (Proposed method)
  - Comparison with character region segmentation excluding images without characters in the training data set (Proposed method)
- *Image feature:*
  - Color histogram
  - Color correlogram

4.3 Results and discussions

Figure 8 shows the overall identification rate of scenes 1) through 5) classified in Table 1. It shows that,

- Correlogram performs slightly better than histograms when n is small, and vice versa when n gets larger. This indicates that correlograms are good at extracting extremely similar images but sensitive to slight differences, and vice versa in the case of histograms. This could be explained by the micro and macro characteristics of both features described in *4.1.2*.

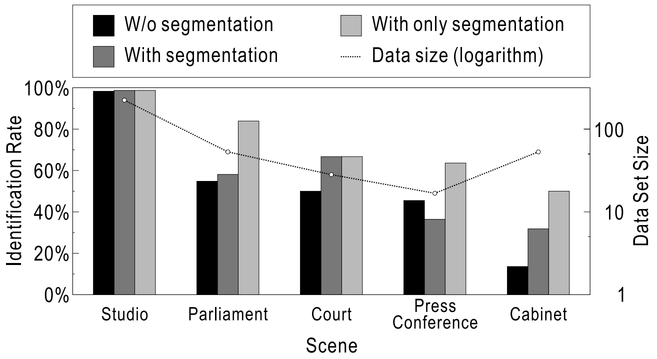

Figure 9: Identification rate (n=1, histogram).

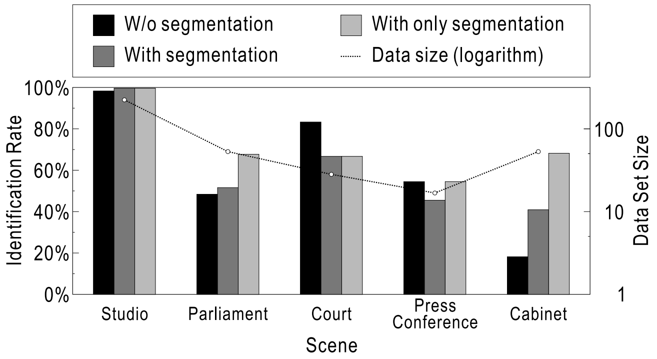

Figure 10: Identification rate (n=1, correlogram).

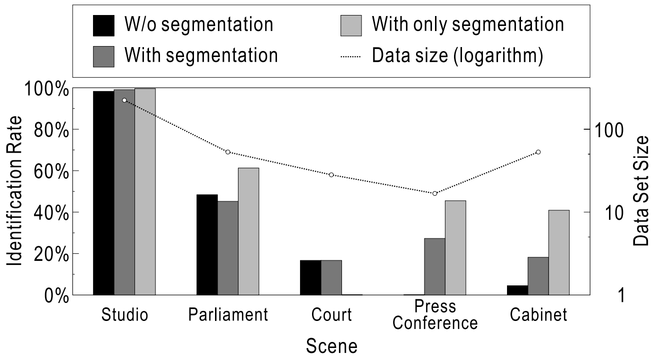

Figure 11: Identification rate (n=3, histogram).

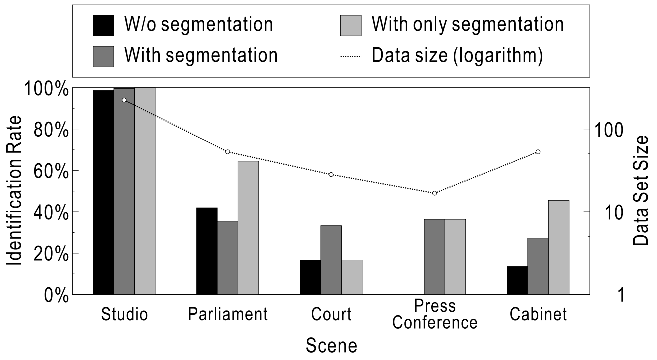

Figure 12: Identification rate (n=3, correlogram).

- The proposed character region segmentation is effective by 2 to 3%. Moreover, similarity comparison excluding images without characters in the training data set increases the effect up to 7 to 8%. This indicates that considering character existence itself as an image feature is important.

Next, Figures 9 through 12 show the identification rates of each pre-classified scene, using histogram or correlogram as image feature and evaluated by $n = 1$ and 3, respectively. The data set size of each pre-classified scene is also shown as a line graph in logarithmic scale. These results show that,

- The proposed method is basically effective as discussed in the overall result, but the effectiveness varies among scenes.
- The priority between histogram and correlogram differs among scenes.

  From these observations, the effectiveness and the appropriateness of the segmentation and the image features vary among scenes due to their characteristics, including restrictions to camera related parameters (location, angle, *etc.*). Even among frequent scenes, the graphical typicality varies to some extent due to camera related parameters and other variables.
- The identification rate is roughly proportional to the logarithm of the data set size. This is prominent when $n = 3$ (Figures 11 and 12), since identification is done based on the majority of the top n similar shots. Since the larger the training data set size is, the more likely similar scenes exist in it, it would be difficult to have the majority of most similar images with correct scenes from a small data set without much variations.

5. CONCLUSION

We introduced a method to identify scenes by segmenting character regions from background region, as a part of the image analysis portion of an automatic news video indexing system. The method takes advantage of news video that some scenes appear frequently, so that similarity measurement with pre-indexed scenes enables such identification.

The method was applied to 817 shots derived from actual news video for evaluation. The overall result showed realistic performance, and indicated that not only the segmentation, but also considering character existence itself as an image feature is important. Although character region segmentation was mostly done manually, automation was possible under certain ideal circumstances. Further study in this field should lead to automation under more general conditions.

On the other hand, independent results among pre-defined scenes indicated several issues that must be considered in future studies:

- Since priority of image features vary among scenes, multiple features should be combinatorily used with weights. Weights may be acquired from training data sets.
- Since the identification rate is roughly proportional to the logarithm of the data set size, in order to cover various situations, the size of training data set should be increased exponentially.

  However, the discrimination ability will decrease in proportion to the increase of scene variations. This should be solved by employing various features in the above mentioned way.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Henry Rowley for kindly allowing us to use ‘the Face detector’[7].

7. REFERENCES

[1] J. Huang, S. R. Kumar, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ’97*, pages 762–768, June 1997.

[2] I. Ide, K. Yamamoto, R. Hamada, and H. Tanaka. *Advanced Multimedia Content Processing –First International Conference AMCP’98, Osaka, Japan–*, volume 1554 of *Lecture Notes in Computer Science*, chapter Automatic Video Indexing Based on Shot Classification, pages 87–102. Springer-Verlag, March 1999.

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

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

[5] Y. Mori, H. Takahashi, and R. Oka. Image understanding based on two database composed of images and words allocated in spaces. *Proceedings of the Fourth Symposium on Intelligent Information Media (Tokyo, Japan)*, pages 127–132, December 1998. (in Japanese).

[6] Y. Nakamura and T. Kanade. Semantic analysis for video contents extraction –spotting by association in news video–. In *Proceedings of the Fourth International Multimedia Conference, ACM Multimedia’97 (Seattle WA, USA)*, pages 393–402, November 1997.

[7] H. D. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[8] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, March 1999.

[9] H. D. Wactler, A. G. Hauptmann, M. G. Christel, R. A. Houghton, and A. M. Olligschlaeger. Complementary video and audio analysis for broadcast news archives. *Communications of the ACM*, 43(2):42–47, February 2000.