An attribute based news video indexing

Ichiro IDE
Nat'l 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
Grad school of engineering
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656, Japan
reiko@mtl.t.u-tokyo.ac.jp

Shuichi SAKAI
Hidehiko TANAKA
Grad school of info sci & tech
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-0033, Japan
sakai,tanaka@mtl.t.
u-tokyo.ac.jp

ABSTRACT

We propose an attribute based news video indexing method. Semantic attributes of both text and image are analyzed, and correspondences within each attribute is considered when indexing, to ensure correspondences between indices and image contents. In this paper, both text and image attribute analysis is briefly introduced, and the result of the final indexing is displayed. The overall result showed comparable performance to existing systems in the case of personal name - character indexing.

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 [10] in the Informedia project. Nonetheless, most general indexing methods utilize indices simply extracted from transcript of speech or caption in the video, and correspondence between index and image content are rarely concerned except in special applications, such as facial image and human name [8].

1.2 Indexing considering image contents

Considering the above mentioned issues, we are proposing an automatic news video indexing system that considers correspondences between indices derived from texts in the video and image contents. In order to realize this, we consider the correspondences separately according to the so called *4W* attributes (*i.e. When*, *Where*, *Who*, and *What*) .

This paper will focus especially on the attribute based indexing between, personal noun - character region, and locational / organizational noun - background scene, which could be considered as the main interest in news video.

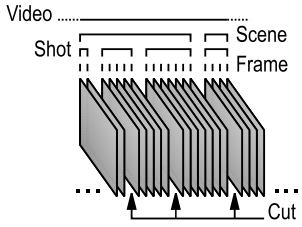

Figure 1: Hierarchical structure and term definition of video.

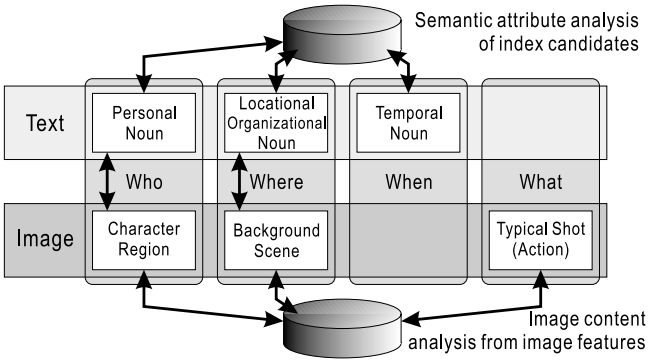

Figure 2: Concept of the indexing system that considers correspondences between indices and image contents.

1.3 Term definition

Figure 1 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.

2. ATTRIBUTE BASED INDEXING

2.1 Overview

The attribute based indexing concept is shown in Figure 2. First, alignment of information derived from both text

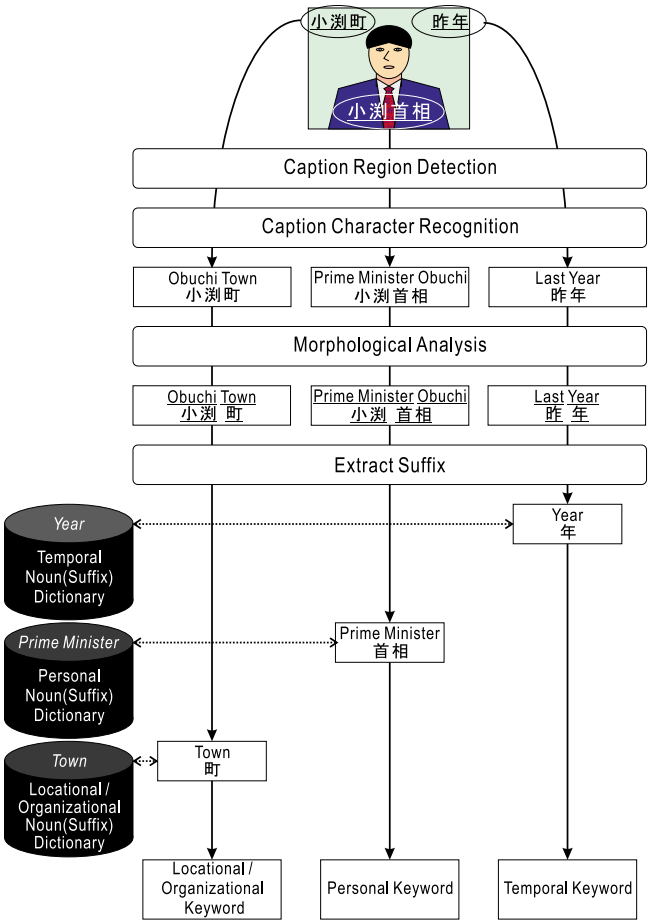

Figure 3: Text attribute analysis referring to suffix.

and image in the video are analyzed in four, so called *4W* attributes. 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, especially between personal noun - character region, and locational / organizational noun - background scene, is checked to ensure correct indexing. Thus, indexing considering correspondence between textual index and image content will be realized.

In the next two sections, attribute analysis of text and image will be briefly introduced.

2.2 Text attribute analysis

(Open) captions are used as the source of textual indexing, among various textual data accompanying a video. Caption was chosen as an indexing source, since they indicate important information briefly, especially what is actually in the image. According to our statistics, 43% of the captions that appeared in news videos had the above mentioned three attributes, which appeared approximately 3 times per minute.

As shown in Figure 2, semantic attribute analysis of keyword candidates derived from captions is required. the task is mostly identical to classification of noun phrase into four classes; namely, (1)personal, (2)locational / organizational, (3)temporal, and others. This task is similar to the named entity task as defined in [9]. Nonetheless, although most solutions for [9] refers to contexts derived from surrounding sentences, it is almost impossible to achieve such information in the case of captions, since they tend to appear as individual noun phrases.

Table 1: Result of caption analysis.

Attribute	Precision	Recall
Personal	72.47%	82.13%
Locational / Organizational	54.76%	88.38%
Temporal	41.93%	93.50%

We have developed a suffix based semantic attribute analysis method especially customized for news video captions [2], taking advantage of Japanese noun phrases, that the suffix tends to determine the semantic attribute of the phrase. Noun (suffix) dictionaries necessary for the analysis was created automatically from a large news paper text corpus under certain conditions, and later expanded by a thesaurus. An example of the analysis is shown in Figure 3; the attributes of three different noun phrases (captions) are determined according to suffixes, 'town', 'prime minister', and 'year'. The determination is done by referring to the noun (suffix) dictionaries.

The result of the analysis applied to 2,549 captions that appeared in 370 minutes of Japanese news video, is shown in Table 1. The result showed high recall and low precision. Nonetheless, low precision could be accepted, since inappropriate captions could be eliminated later when checking the correspondences with the image. Note that captions were written down manually for the experiment, but OCR technologies specialized for caption recognition such as [5] may be used in case of automation.

2.3 Image attribute analysis

As news programs focus on providing mostly important information on human activities, there is a good 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 [4, 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 and background regions.

As shown in Figure 4, we proposed an image attribute analysis method that segment and separately analyze character and background regions, when there is a facial region of certain size and direction [3]. First, background and character region is segmented, and image features of each region are extracted. Next, content analysis is performed referring to pre-constructed knowledge base on image feature and content relations.

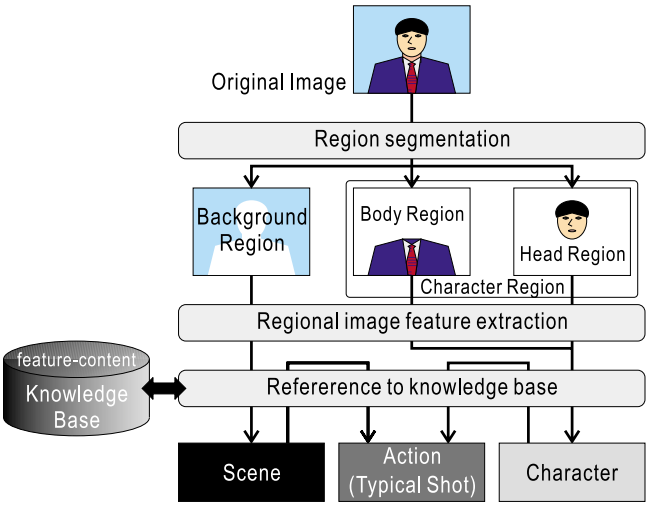

Figure 4: Image contents analysis by character region segmentation.

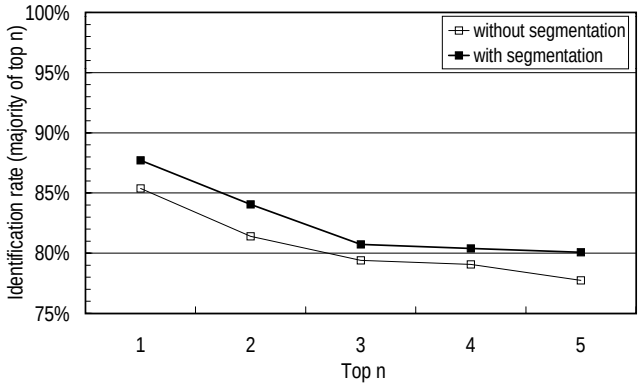

Figure 5: Overall scene identification rate.

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 [4, 6].

The result of a scene (*Where*) identification experiment applied to 817 shots, is shown in Figure 5. Face detection and character region segmentation was mostly done manually, and color correlogram [1] was used as an image feature. The graph shows the overall result of a supervised classification to five scene classes that frequently appear in domestic news topics; namely, (1)cabinet meeting, (2)parliament, (3)press conference, (4)court, and (5)studio. The identification rate shows the percentage of correctly identified scenes, in the sense that majority of top n scenes were correctly identified. The scene identification showed better performance when character region segmentation was done, compared to non-segmented identification.

Since the performance of individual shot classes was roughly proportional to the logarithm of the size of the training data set size (number of shots used for training), the performance should improve if the training data set were to be enlarged exponentially. Generally, such enlarged training should be laborious. However, shot classes as those used in the experiment could be easily gathered during a relatively short time period of broadcast news video, since they appear very frequently.

Although character region segmentation was done manually in the experiment, it could be performed automatically under certain conditions; good lighting condition and frontal portrait. We experimentally applied an automatic face detection method based on neural network training [7] for face detection, and a template relative to the detected facial region in order to extract the character region. As a result, not all characters were extracted correctly, but 100% of anchor persons in a studio that satisfies the conditions mentioned above, were correctly detected and extracted.

Details on the image attribute analysis, especially on background region (*Where*) analysis is described in [3].

3. INDEXING

Following the text and the image analyses, indexing was performed integrating the results. The correspondence between (1)personal noun - character region (*Who*), and (2)locational / organizational noun - background scene (*Where*) was considered when indexing, as shown in Figure 2.

The same 817 shots derived from Japanese news videos used in **2.3** were used in the following experiments, and the evaluation was done manually based on human marking.

3.1 Personal noun - character region indexing

194 shots out of the 817 shots, with considerably large human figures were used for this experiment. The indexing was done following the following procedure:

1. Detect topic boundaries referring to typical anchor shots with title captions.
2. Find inside the same topic, graphically similar shots (similarity over 90%) to the shot to be indexed.
3. Select among captions in the similar shots, those that have a personal noun attribute.
4. Sort and list the selected captions according to the graphical similarity. The higher ones will be the indices to the shot to be indexed, and the similarity is referred as the appropriateness.

3.2 Locational / organizational noun - background scene indexing

130 shots out of the 817 shots, which were easy to determine the location of the scene manually, were used for this experiment.

The indexing was done following the following procedure:

1. Detect topic boundaries referring to typical anchor shots with title captions.

Table 2: Result of indexing.

Attribute	Accuracy
Personal - Character	77.6%
Locational / Organizational - Scene	23.8%

2. Find among all shots, graphically similar shots (similarity over 90%) to the shot to be indexed.

3. Find inside the same topic with the similar shot, graphically similar shots (similarity over 90%) to the similar shot.

4. Select among captions in the similar shots, those that have a locational / organizational noun attribute.

5. Sort and list the selected captions according to the graphical similarity. The higher ones will be the indices to the shot to be indexed, and the similarity is referred as the appropriateness.

3.3 Results and discussions

Table 2 shows the accuracy of the indices. Here, accuracy is defined as the percentage of correctly indexed shots, in the sense that the index with the highest appropriateness was correct. The personal - character indexing shows a good result, comparable to the ability of the Name-It system [8]. On the other hand, locational / organizational - scene does not show a good result.

Part of the failures could be blamed on the text analysis result; the accuracy goes up to 82.2% and 35.4%, respectively, when manually corrected attributes are used. On the other hand, scene identification ability by image analysis, as shown in Figure 5 could also be blamed for some part of the failure. But, the largest reason for the failure (especially the locational / organizational - scene indexing) is that some frequent contents do not have captions that are too obvious to the viewers.

To improve personal - character indexing, face recognition might be a solution. On the other hand, introducing larger volume of video data so that there would be a good chance that a caption describing the content appears, would be a solution to overcome the failures for locational / organizational - scene indexing.

4. CONCLUSIONS

We have introduced a news video indexing system that considers correspondences between indices and image contents. A brief overview and result of text and image analyses was displayed, and the result of the indexing was also shown. The indexing result for personal - character indexing showed good performance, comparable to existing systems. Although locational / organizational - scene indexing requires improvement for practical application, countermeasures for improving the indexing accuracy was considered for future study.

5. 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, R. Hamada, S. Sakai, and H. Tanaka. Semantic analysis of television news captions referring to suffixes. In *Proceedings of the Fourth International Workshop on Information Retrieval with Asian Languages, IRAL'99 (Taipei, Taiwan)*, pages 37–42, November 1999.

[3] I. Ide, R. Hamada, S. Sakai, and H. Tanaka. Scene analysis in news video by character region segmentation. In *Proceedings of ACM Multimedia 2000 Workshops (Marina del Rey CA, USA)*, pages 195–200, November 2000.

[4] 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.

[5] H. Kuwano, S. Kurakake, and K. Odaka. Telop character extraction from video data. In *Proceedings of the Workshop on Document Image Analysis '97*, pages 82–88, June 1997.

[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] B. M. Sundheim. Named entity task definition, version 2.1. In *Proceedings of the Sixth Message Understanding Conference, MUC-6*, pages 317–332, November 1995.

[10] 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.