Automatic Video Indexing Based on Shot Classification

Ichiro Ide, Koji Yamamoto*, and Hidehiko Tanaka

{ ide | kyama | tanaka } @mtl.t.u-tokyo.ac.jp

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 recycling and searching. 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 to cope with the increasing amount and to achieve sufficient precision for detailed searching.

We are trying to accomplish this task by referring to both video data and accompanying natural language data in Japanese television news video. There are several notable attempts to automatically index television news video from this approach. Most of their indexing strategies are based on frequency or just simple occurrence of words or phrases. On the other hand, others search for words in a full text searching manner. These methods are relatively simple, and in that sense quite practical approaches, but the critical point is that they do not necessarily ensure the correspondence of the image contents and the index.

Reflecting these background issues, in this paper we will propose and evaluate an indexing method, which indexes keywords with appropriate semantic attributes to classes of shots with graphically typical feature. The base of this

*

method lies in the characteristics of television news programs; graphically similar (in a certain perspective) shots contain semantically similar contents. Since keywords are tagged selectively according to the contents of each typical shot class, the correspondence of image contents and keywords is guaranteed to a certain extent.

We will first take an overview of the characteristics specific to television news video and related works in the next Sect., and then introduce the proposed method in Sect. 3. The succeeding Sects. 4, 5 and 6 discuss in-depth matters of the method, and Sect. 7 concludes the paper.

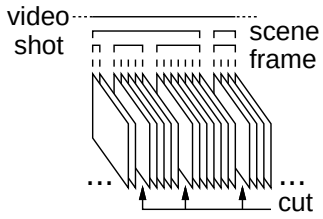
2 Indexing to Television News

First, we will overview the characteristics specific to television news video, and next introduce several related works by other people.

2.1 Structure of Television News Video

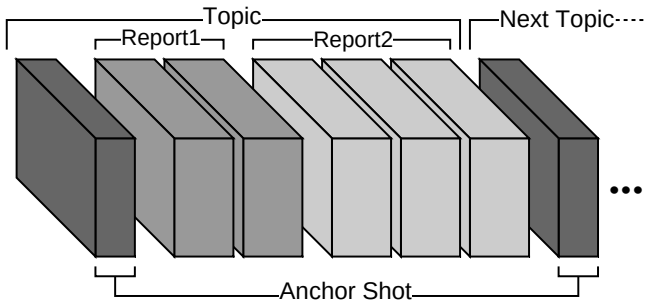
News videos have both graphical and semantic structures as described here.

Graphical Structure. The graphical structure that television news videos have are not specific to the genre. As shown in Fig. 1, they are generally hierarchically structured.



Semantic Structure. As shown in Fig. 2, the semantic structure of television news video is quite unique.

It is very important to detect the boundaries of topics and to grasp the semantic structure of the video before indexing, so this structure could be used as an opportune key. As mentioned in Sect. 1, since news video tend to be taken in similar situations, most shots could be classified to several typical shot classes referring to graphical feature.



2.2 Natural Language Data in Video

Variety of Natural Language Data Source. There are various natural language data sources accompanying the video; main audio, sub audio, closed caption (mostly same contents as the main audio) and caption. Particularly, captions are usually used to describe important matters in a digestive form, so they could be considered as adequate keyword candidates for indexing. According to our statistics, they appear approximately once every 15 seconds in news programs, which is a moderate frequency for finding keyword candidates. Since main audio (or closed caption) require complicated process to be used as a keyword extraction source, captions that have these characteristics are employed in the proposed system.

Characteristics of Captions. Captions have specific characteristics that differ from normal texts, which makes the analysis employing conventional natural language processing methods difficult. This problem is solved to a certain extent in this application by the method described in Sect. 5.

On the other hand, semantic characteristics of captions could be classified as shown in Tab. 1. Among these types, (a) and (b), which consist about half of the captions, represent the contents in the image directly. This allows (a) and (b) to almost directly become keyword candidates. Although (c) does not necessarily reflect the graphical contents of the video, it is an important information which explains the topic. Thus, nearly 60% of the captions; (a), (b) and (c), could be directly used as keywords.

2.3 Related Works

As a general video database creating and browsing system, Informedia project [8, 10, 13, 21] at CMU is the most significant work in this field. They have created

an automatic archiving and presenting system for CNN (Cable News Network) news video. It automatically recognizes main audio speech and extract keywords from the text deduced from it by evaluating the rarity of words by the TF-IDF (Term Frequency Inverse Document Frequency) method. Although such statistic approaches are relatively simple, and in that sense quite practical, they do not necessarily ensure the correspondence of image contents and keywords, which is essential for video database.

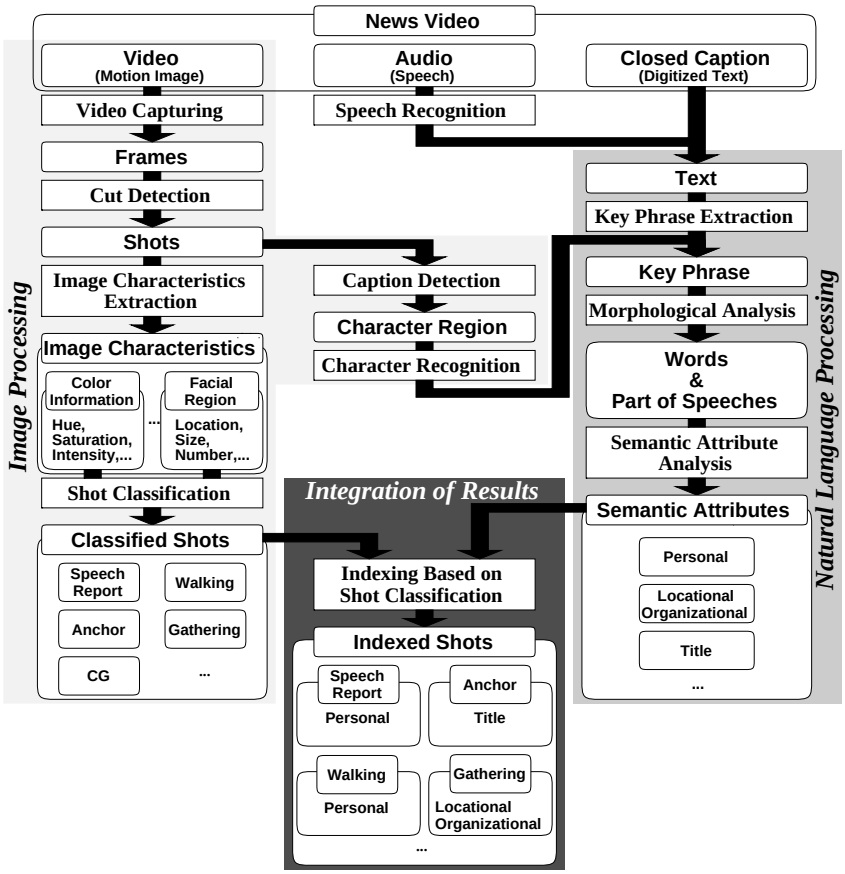
On the other hand, Nakamura and Kanade [5] have proposed an automatic indexing method that classifies shots into several typical classes, and tag key sentences derived from syntactic and semantic analyses of the closed caption. Although the basic idea of classifying shots into typical classes is similar to our approach, the point that they utilize closed captions and that they execute key sentence extraction, differentiates the two methods.

Similar to this approach, Satoh *et al.* [6] proposed an automatic facial image and personal name associating method that associates facial regions extracted from the image and personal names derived from closed captions analysis. Although this is completely automated and performs fairly well, it concentrates on associating personal faces and names, which is an acceptable limitation, but not sufficient for news video database.

3 Indexing Based on Shot Classification

Considering the issues discussed in Sect. 2, we will propose an automatic indexing method based on shot classification. The basic idea of the method is based on the characteristic specific to television news video; graphically similar images contain similar contents. Based on this assumption, graphically typical shots will be indexed keywords with certain attributes.

The overall indexing scheme is shown in Fig. 3. A simple overview of each phase is introduced in this Sect. In-depth description and evaluation on shot classification, caption analysis, and indexing are discussed in the succeeding Sects.



Although we do not currently employ main/sub audio and closed caption data, the scheme could be easily extended to handle them as shown in the figure. This may also seem as a mere combination of conventional algorithms, but the overall ability of indexing by integrating video and natural language data is far superior than a simple combination.

3.1 Image Processing Phase

Video Capturing. First, analog video is digitized by an image capture board connected to a PC at a sampling rate of 5 frames per second. We consider this rate is sufficient for indexing, although the original frame rate of the NTSC standard is approximately 30 frames per second.

Cut Detection. Next, cuts are detected, since shots are the most primary unit to handle video data. Cut detection has been challenged in two approaches; detecting from non-compressed video, and from compressed video. The former approach detects cuts by analyzing the graphical similarity of adjoining frames [12, 17]. The latter takes advantage of the compression algorithm. Especially MPEG video is fit for the purpose due to its compression algorithm based on graphical correlation between frames [3].

Among these methods, we chose the Nagasaka-Tanaka (segmented χ^2 examination) algorithm [17], which is very simple but quite effective. The Nagasaka-Tanaka method evaluates the similarity of color histograms of corresponding equally segmented blocks of adjoining frames, applying the χ^2 examination function. The function is defined as follows:

$$\chi^2(i) = \sum_{c=0}^{c_{max}} \frac{(H_i^n(c) - H_i^{n+1}(c))^2}{H_i^n(c)} \tag{1}$$

where $H_i^n(c)$ and $H_i^{n+1}(c)$ represent the color histogram of the ith block of two adjoining frames n and $n+1$ respectively. When more than half of the values of $\chi^2(i)$ exceeds the threshold, a cut is detected in between the frames.

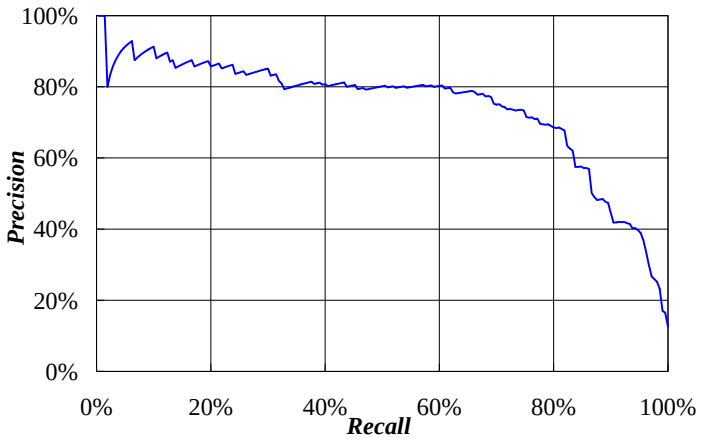


Figure 4 shows the relation between recall and precision of cut detection, when the method was applied to 30 minutes of actual news video with 210 cuts.

After these pre-processes, extraction of graphical feature is performed before the shot classification.

Caption Recognition. As a tributary to the main-stream image processing, caption detection and character recognition should be performed. This is not currently implemented, and captions are written down manually. Although OCR (Optical Character Recognition) softwares with high recognition rates do exist, the resolution of TV captions are limited due to their size and the number of scanning lines (525 lines per frame in the case of NTSC broadcasting standard). Fragments of background image filtering through the characters should be eliminated before character recognition, which is also a difficult process. These restrict the application of conventional OCR techniques to television caption character recognition, especially to complicated Japanese characters. Nonetheless, several attempts are made to accomplish the task [9, 14], although their recognition rates have still room for improvement. We may hope for digitized caption texts to be broadcast along with the video, following the future digitalization of television broadcasting.

3.2 Natural Language Processing Phase

First, morphological analysis to digitized texts derived from captions are performed using the Japanese morphological analysis system JUMAN [20]. This is a pre-process for analyzing the semantic attribute of the entire caption, which refers to suffixable nouns.

3.3 Integration Phase

After image classification and caption analysis of the shot are done, the integration phase indexes shots with captions with appropriate semantic attributes for the typical shot class; *i.e.* when the shot is a 'speech shot', the speaker's name is an appropriate keyword. When it is a 'gathering shot', the name of the gathering, say a conference, is considered appropriate.

Such indexing scheme ensures the correspondence of image content and keyword, which is essential for video database.

4 Shot Classification

After the pre-process, each shot is classified based on its graphical feature. Note that the classification rules are based on combinations of relatively simple graphical feature extraction process, which makes the method applicable to large amount of incoming video data.

We have defined five shot classes:

- Speech/Report
- Anchor
- Walking
- Gathering
- Computer Graphics (CG)

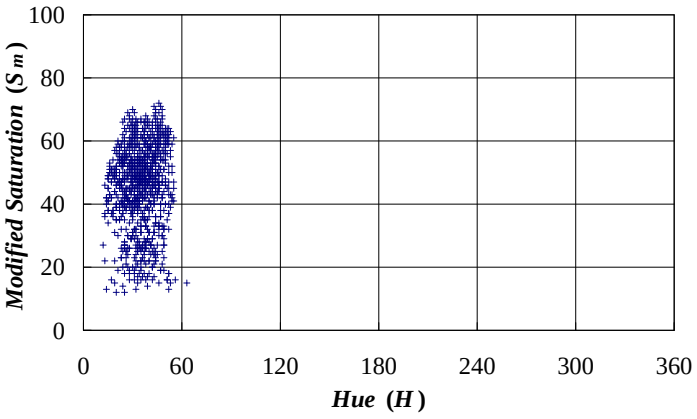
These classes covered 57% of the entire news video that we used for experiment. Details on each shot class are described in this Sect.

4.1 Speech/Report Shot

When a person is addressing a speech, or a reporter is reporting from a relay spot, there is usually one person speaking in the middle of a frame. In order to detect such a shot, (1) human face, and (2) lip movement should be detected. Condition (2) is employed to avoid detecting a portrait picture, or a video with a person just standing in the middle of a frame without speaking.

An anchor shot is also detected from these conditions, but they will be separated later.

Face Detection. Face detection is a very popular research field that has developed various algorithms, but the following method is considered sufficient and simple enough to serve our purpose.



H S_m

1. **Skin colored region extraction**
   The modified HSI color system is used to detect skin colored regions [16]. I (Intensity) is used only for excluding dark regions. A certain rectangular region in the H (Hue) - S_m (Modified Saturation) plane was defined as skin color. The distribution of sample skin colored regions on the H - S_m plane is shown in Fig. 5. Pixels whose H and S_m exist in this region are determined as skin colored. Small regions that consist of less than a certain number of pixels are deleted, and adjoining isolated regions are merged by spreading out their boundaries for a few pixels.
2. **Template matching**
   Template matching is performed to exclude hands, walls, desks and so on

that were extracted as skin colored regions. Average faces in several different resolutions are prepared from the I of various facial regions, and are selectively used for matching, according to the size of the extracted skin colored region. In order to decrease the influence of optical states, the I of the extracted region is regularized by the overall I of the frame.

Lip Movement Detection. Once a facial region is detected, it is easy to estimate the mouth location. This is because faces in speech/report shots and anchor shots are usually full faces. If the temporal change of the area around the estimated mouth location is relatively bigger than the change of the entire facial region during a shot, lip movement is detected.

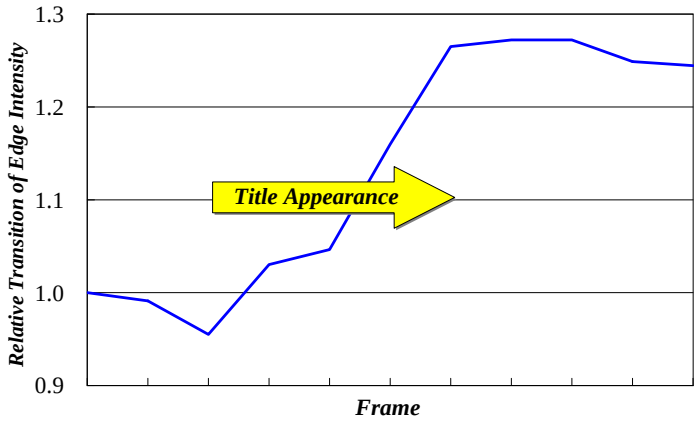
4.2 Anchor Shot

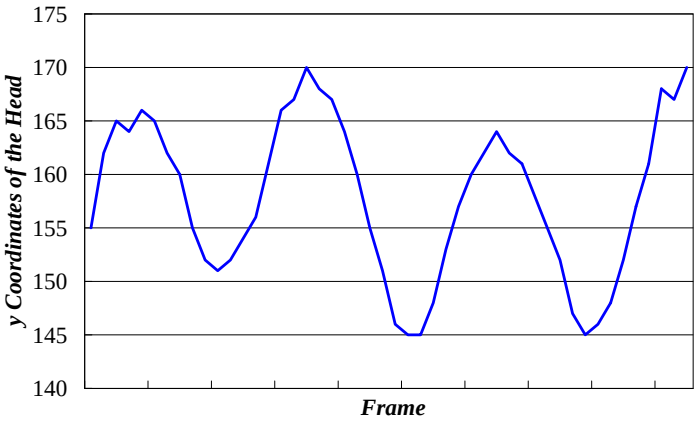
Separation of Anchor Shots Among Speech/Report Shots. Anchor shots initially satisfy the classification conditions for the speech/report shot. One distinctive feature of anchor shots is that they are graphically extremely similar among themselves, and also frequent compared to other speech/report shots. Therefore, after speech/report shot classification, anchor shots are separated by clustering the speech/report shots by evaluating the mutual similarity between all the classified shots. The largest and most dense cluster would be the anchor shots. The similarity is evaluated by the comparison of color histograms applying the segmented χ^2 examination previously used for cut detection. Each shot is regularized so that the facial region should be located in the same position.

Detection of Topic Boundary. The main purpose for detecting anchor shots is to detect boundaries of news topics. However, since anchor shots may appear in the middle of a long topic, just separating them from speech/report shots is not sufficient to fulfill this purpose. A distinctive feature of anchor shots in the beginning of a new topic is the presence of a title caption. As shown in Fig. 6, this could be detected by observing the transition of the overall edge intensity of frames, caused by the superimposion of the title caption CG. Therefore, anchor shots with prominent edge intensity transitions are used to detect topic boundaries.

4.3 Walking Shot

When a person is walking, the upper half of the body oscillates up and down following the steps. A television camera is usually stabilized on a tripod and does not oscillate along the vertical axis. Facial region detection is performed likewise the speech/report shot to the yet unclassified shots. As shown in Fig. 7, a walking shot is classified by detecting the up and down oscillation of the bottom of a facial region.





4.4 Gathering Shot

When many people are gathering, there are usually more than two similar sized people in a frame. Facial region detection is performed likewise the speech/report shot to the yet unclassified shots. A gathering shot is classified by detecting more than two similar sized facial regions in a frame.

4.5 Computer Graphics (CG) Shot

CG shots are quite tricky, since they may contain a referential portrait picture, or have explanatory figures and texts, which may cause mal-effects to shot clas-

sification and caption analysis. Therefore, CG shots need to be detected and excluded from the indexing scheme. A distinctive feature of CG shots is that they are usually motionless. Nevertheless, there are occasional movements, for example in an explanatory flowchart. So, CG shots are classified by the total duration of motionless frames, not by the overall motionlessness. In the following experiment, the duration was set to one second.

4.6 Classification Experiment

Table 2 shows the result of the shot classification applied to 75 minutes of news video. The numbers of true answers ($= N_{Classified} + N_{Unclassified}$) were determined and counted manually for evaluation. Recall and precision are defined as follows:

$$Recall = \frac{N_{Classified}}{N_{Classified} + N_{Unclassified}} \tag{2}$$

$$Precision = \frac{N_{Classified}}{N_{Classified} + N_{Misclassified}} \tag{3}$$

	$N_{Classified}$ $N_{Misclassified}$ $N_{Unclassified}$	

Major reasons for misclassification and unclassification were:

- Lip movement could not be detected in a speech/report shot, since the face was not a full face.
- Topic boundaries were not detected for some very short topics, which did not have title captions in the beginning.
- Facial regions were not correctly detected in a gathering shot, since the faces were too small and/or hidden by hair. This is often the case with indoor meetings, when a camera shoots from the rear.
- Up and down oscillation of a head could not be detected in a walking shot, since the person was walking too far away from the camera.
- A completely still image of an object was misclassified as a CG shot.

5 Caption Analysis

Caption analysis is necessary to index typical shots with appropriate keywords reflecting their typical contents. As shown in Tab. 1, captions that have (1) personal and (2) locational/organizational attributes are adequate keyword candidates. They are mostly noun phrases (often simple arrays of nouns). In Japanese, the utmost tail nouns, *i.e.* suffixable nouns, define their attributes[2].

As related research on semantic disambiguation of nouns, several methods do exist. Nasukawa [7] proposed a method that determines semantic attributes of proper nouns (*i.e.* whether a proper noun indicates a place or a person) referring to the context of neighboring sentences. On the other hand, Watanabe *et al.* [11] proposed a method that analyzes television news captions by referring to both locations and grammatical characteristics as keys.

Although these methods perform fairly well, the former method is difficult to serve our purpose since captions do not have enough neighboring information to analyze contexts, and also since it is purposed to handle only proper nouns. The latter method is originally designated to serve similar purpose to ours, but is not generally applicable to various news programs, which have different designing policies where layouts of captions vary.

Similarly to our task, the Named Entity task defined for the Message Understanding Conference (MUC) [15] assigns the participants to classify personal, organizational, locational, temporal and numerical phrases. The difference is that our aim is not limited to proper nouns, where the Named Entity task limits the tagging to personal, organizational and locational phrases to proper names.

Considering these issues, we decided to analyze captions on their own by referring to suffixable nouns.

5.1 Collecting Suffixable Nouns

To enable caption analysis based on suffixable nouns, first such suffixable nouns, *i.e.* (1) personal nouns, and (2) locational/organizational nouns, should be collected. These were collected according to certain conditions from two text corpora that consist of newspaper articles [18, 19]. These were manually morphological analyzed beforehand, which ensures the basic reliability of the collection process.

Details are discussed elsewhere [1, 4] since they meddle with language specific issues, but as a result, 3,793 nouns were collected as personal nouns, and 11,166 as locational/organizational nouns. Note that the collected suffixable nouns include those that represent people or locations/organizations alone, such as 'volunteer' and 'kitchen', but does not include proper names.

[2]

5.2 Semantic Analysis Experiment

Table 3 shows the result of the caption analysis applied to the same 75 minutes of news video used for the shot classification experiment. The numbers of true answers ($= N_{Classified} + N_{Unclassified}$) were determined and counted manually by a third person, and recall and precision are defined as noted in formulae (2) and (3), respectively.

	$N_{Classified}$ $N_{Misclassified}$ $N_{Unclassified}$	

Major reasons for misclassification and unclassification were:

- Some nouns were essentially applicable to both categories (Semantic diversity).
- Some nouns in the collected dictionary were inappropriate (Noise).
- The collected nouns were insufficient (Lack of vocabulary).

The latter two could be solved by further improvement of the collection scheme, but the semantic diversity is an essential issue when dealing with semantics of words.

The reason for the low rates of locational/organizational captions is due to the loose conditions of the collection rule. It is difficult to tighten the rule without more precise grammatical information tagged beforehand in the corpora.

6 Indexing to Classified Shots

Following shot classification and caption analysis, appropriate keywords are tagged to each classified shots according to their classes.

6.1 Indexing Scheme

Appropriate semantic attributes of keywords are defined to each typical shot class as shown in Tab. 4. Anchor shots with captions are referred to detect boundaries of topics for this process. On the other hand, CG shots are excluded from indexing, considering their uniqueness.

Following these principles, indexing is performed to all the classified shots except intermediate anchor shots and CG shots according to the following procedure:

1. Search for a caption with an appropriate attribute inside the shot, and index it if found.
2. If not found, search for it in graphically similar shots inside the topic, and index it.

6.2 Indexing Experiment

Table 5 shows the result of indexing applied to the same 75 minutes of news video used in previous experiments. In order to evaluate "indexing based on shot classification" independently, true answers of semantic attributes of captions given by a third person were used.

The result does not necessarily show practical performance as a whole (/All), but as an evaluation of the proposed method itself (/Indexable), all classes showed more than 75% of recognition rate. The overall performance should improve by employing other natural language source as shown in Fig. 3.

7 Conclusion

In this paper, we have proposed and evaluated an indexing method that indexes television news video, considering the correspondence of image contents and keywords. The overall result is not necessarily practical, but the performance of the method itself is quite promising. Although the techniques applied for shot classification and pre-process for caption analysis were conventional, the effectiveness of integrating image and language information was also shown through the indexing.

The problem is that the numbers of typical shot classes and caption attributes are relatively small, since the classification rules were given in a top-down manner. We are currently examining an automatic classification rule acquisition method based on statistic relations of graphical feature and semantic attributes of captions [2]. The result of a preliminary experiment showed promising performance, although the amount of data was not large enough to discuss statistics. We will apply the acquired classification rules to enable better automatic indexing based on shot classification in the near future.

Acknowledgments

The Japanese Morphological Analysis System JUMAN is a free software developed at and distributed by Nagao Lab., Kyoto University and Matsumoto Lab., Nara Institute of Science and Technology. The Kyoto University Text Corpus is a product of Nagao Lab., Kyoto University. RWC text database is a product of Real World Computing Partnership (RWCP), and is used by the authors under a licensed agreement.

References

Proc. 3rd Intl. Workshop on Information Retrieval with Asian Languages to appear

Proc. 6th ACM Intl. Multimedia Conf. -Art Demos-Techinical Demos-Poster Papers-

Proc. 14th Intl. Conf. on Pattern Recognition

Trans. IPS Japan *in Japanese*

Proc. 5th ACM Intl. Multimedia Conf.

Proc. IJCAI'97

Proc. 11th Annual Conf. JSAI *in Japanese*

Proc. AAAI'97 Spring Symp. on Intelligent Integration and Use of Text, Image, Video and Audio Corpora

SPIE Proc. of Storage and Retrieval for Image and Video Database V

CMU Tech. Rep.

Tech. Rep. IPS Japan *in Japanese*

Proc. 1996 Intl. Conf. on Image Processing

IEEE Computer

Tech. Report IEICE *in Japanese*

Proc. 6th Message Understanding Conference

Canadian Conf. on Elec. and Comp. Eng. '95

IFIP Trans.

http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

http://www.informedia.cs.cmu.edu/