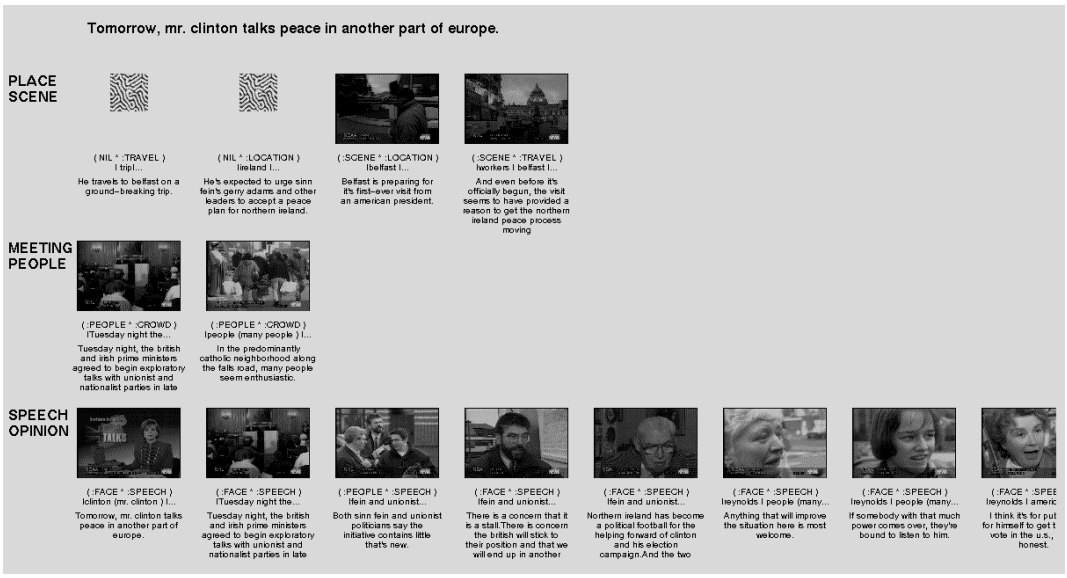

Figure 7: News video TOPIC EXPLAINER

Table 13: Spotting result 1 (six 30-minute videos)

type	all A	matched B	correct C	miss D	wrong E
speech	292	226	178	40	48
meeting	47	26	19	18	7
crowd	63	35	26	19	9
travel	15	8	7	6	1
location	76	34	27	32	7
face	472	217	173	0	44
people	220	84	63	0	21
scene	168	25	21	0	4

A is the total number of *key-data*, B is the number of *key-data* for which inter-modal correspondences are found, C1 is the number of *key-data* associated with correct correspondences, D is the number of missing association, that is *clues* for which association is failed in spite of having real correspondences, E is the number of wrong association, *i.e.* mismatching.

Table 14: Spotting result 2

	face	people	scene
speech	199/165	24/12	2/1
meeting	9/6	15/12	1/1
crowd	5/1	28/25	1/0
visit	1/0	4/4	3/3
location	3/1	13/10	18/16

Each figure (X/Y) in the following table shows, the number of found correspondences (X) and the number of correct correspondences (Y).

Figure 8: Details in TOPIC EXPLAINER

References

Hauptmann, A. and Smith, M. Video Segmentation in the Informedia Project. In *IJCAI-95, Workshop on Intelligent Multimedia Information Retrieval*, 1995.

Miller, G. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, Vol. 3, No. 4, 1990.

Rowley, H., Baluja, A. and Kanade, T. Neural Network-Based Face Detection. *Image Understanding Workshop*, 1996.

Smith, M. and Hauptmann, A. Text, Speech, and Vision for Video Segmentation: The Informedia Project. *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, 1995.

Sleator, D. and Temperley, D. Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies*, 1993.

Wactlar, H., Kanade, T., Smith, M. and Stevens, S. Intelligent Access to Digital Video: The Informedia Project. *IEEE Computer*, Vol. 29, No. 5, 1996.

Table 12: Matching evaluation for type combinations

	speech	meeting	crowd	visit	location
face	1.0	0.25	0.25	0.25	0.0
people	0.75	1.0	1.0	0.5	0.5
outdoor scene	0.0	0.25	0.25	1.0	1.0

Cost of Matching ($Match$): The evaluation of correspondences is calculated by the following formula.

$$Match(i,j) = M_{\text{time}}(i,j) \cdot M_{\text{type}}(i,j) \qquad (2)$$

where M_{time} is the duration compatibility between an image and a sentence. The more their durations have overlap, the less the penalty becomes.

A *key-image*'s duration (d_i) is the duration of the cut from which the *key-image* is taken; the starting and ending time of a sentence in the speech is used for *key-sentence* duration (d_s). In the case where the exact speech time is difficult to obtain, it is substituted by the time when closed-caption appears.

The actual values for M_{type} are shown in Table 12. They are roughly determined by the number of correspondences in our sample videos.

Experiments

We chose 6 CNN Headline News videos from the Informedia testbed. Each video is 30 minutes in length.

Results

Fig.6 shows the association results by DP. The columns show the *key-sentence*s and the rows show *key-image*s. The correspondences are calculated from the paths' cost, as shown in the figure. In this example, 167 *key-image*s, 122 *key-sentence*s are detected; 69 correspondences are successfully obtained.

Total numbers of matched and unmatched *key-data* in 6 news videos are shown in Table 13. Details are in Table 14.

As shown in the above example, the accuracy of the association process is good enough to assist manual tagging. About 70 segments are spotted for each video, and around 50 of them are correct. Although there are many unmatched *key-image*s, most unmatched *key-image*s are taken from commercial messages for which corresponding *key-sentence*s do not exist. However, there are still considerable number of association failures. They are mainly caused by the following factors:

- Errors of *key-image* or *key-sentence* detection
- Time lag between closed-caption and actual speech
- Irregular usage of *clues*. For example, an audience's face close-up rather than the speaker's in a speech or talk situation.

Usage of the Results

Given the spotting results, the following usage can be considered.

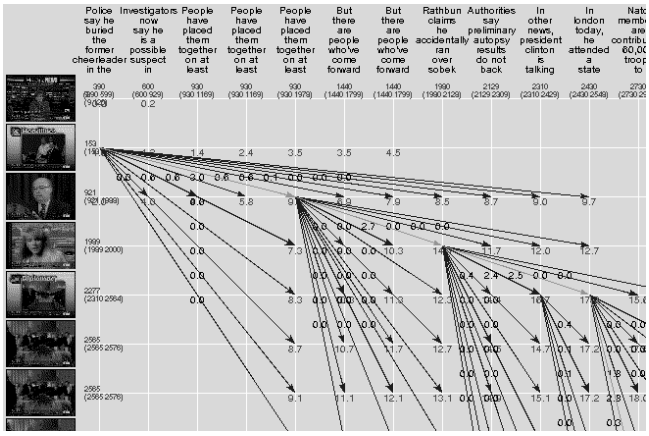

Figure 6: Correspondence between sentences and images

1. Summarization and presentation tool:
   Around 70 segments are spotted for each 30-minute news video. This means an average of 3 segments in a minute. If a topic is not too long, we can place all of the segments in one topic into one window. This view could be a good presentation of a topic as well as a good summarization tool. An example is shown in Fig.7 and Fig.8. Each row shows segments about location, those for meetings/crowds, and those for speech/opinion, respectively: The first row shows Mr. Clinton's visit to Ireland and the preparation for him in Belfast; the second row explains about politicians and people in that country; the third row shows each speech or opinion about Ireland peace.

2. Data tagging to video segments:
   As mentioned before, the situations such as "speech scene" situation can be a good tag for video segments. Currently, we are trying to extract additional information from transcripts. The name of a speaker, attendants in a meeting/conference, a visitor and location of visit, etc. With this data, video segment retrieval can be much more efficient.

Conclusion

We described the idea of the Spotting by Association in news video. By this method, video segments with typical semantics are detected by associating *language clues* and *image clues*.

Our experiments have shown that many correct segments can be detected with our method. Most of the detected segments fit the typical situations we introduced in this paper. We also proposed new applications by using detected news segments.

There are many areas for future work. One of the most important areas is the improvement of *key-image* and *key-sentence* detection. Another is the check of effectiveness with other kinds of videos.

(a) (b)

Figure 5: Example of outdoor scenes

Table 9: Usage of outdoor scenes

video	outdoor scenes
Video1	34
Video2	39

difficult because small faces and human figures are more difficult to detect. The same can be said to outdoor scene detection.

Automatic face and outdoor scene detection is still under development. For the experiments in this paper, we manually pick them. Since the representative image of each cut is automatically detected, it takes only a few minutes for us to pick those images from a 30-minute news video.

Association by DP

The sequence of *key-sentence*s and that of *key-image*s are associated by Dynamic Programming.

Basic Idea

The detected data is the sequence of *key-image*s and that of *key-sentence*s to which starting and ending time is given. If a *key-image* duration and a *key-sentence* duration enough overlap (or close to each other) and the suggested situations are compatible, they should be associated.

In addition to that, we impose a basic assumption that the order of *key-image* sequence and that of *key-sentence* sequence are the same. In other words, there is no reverse order correspondences. Consequently, dynamic programming can be used to find the correspondences.

The basic idea is to minimize the following penalty value P.

$$P = \sum_{j \in Sn} Skip_{\mathrm{s}}(j) + \sum_{k \in In} Skip_{\mathrm{i}}(k) + \sum_{j \in S, k \in I} Match(j,k) \quad (1)$$

where S and$I$ are the *key-sentence*s and *key-image*s which have corresponding *clues* in the other modality, Sn and In are those without corresponding *clues*. $Skip_{\mathrm{s}}$ is the penalty value for a *key-sentence* without inter-modal correspondences, $Skip_{\mathrm{i}}$ is for a *key-image* without inter-modal correspondences, and $Match(j,k)$ is the penalty for the correspondences between the j-th *key-sentence* and the k-th *key-image*.

Table 10: Example of cost definition

key-sentence: speech 1.0, meeting 0.6, crowd 0.6, travel/visit 0.6, location 0.6
key-image: face 1.0, people 0.6, scene 0.6

Table 11: Example of sentence cost definition

1.SPEECH/OPINION
keyword's part-of-speech: verb 1.0, noun 0.6
subject type: a proper noun suggesting a human or a social group 1.0, a common noun suggesting a human or a social group 0.8, other nouns 0.3
2.MEETING
keyword's part-of-speech: verb 1.0, noun 0.6
subject type: a proper noun suggesting a human or a social group 1.0, a common noun suggesting a human or a social group 0.8, other nouns 0.3
verb semantics: verbs suggesting attendance 1.0, the other verbs 0.8

In DP path calculation, we allow any inter-modal correspondence unless the duration of a *key-image* and that of a *key-sentence* are mutually too far to be matched[7]. Any *key-sentence* or *key-image* may be skipped (warped), that is left unmatched.

Cost Evaluation

Cost of Skipping (*Skip*)**:** Basically, the penalty values are determined by the importance of the data, that is the possibility of each data having the inter-modal correspondences. In this research, importance evaluation of each *clues* is calculated by the following formula. The skip penalty *Skip* is considered as $-E$.

$$E = E_{\mathrm{type}} \cdot E_{\mathrm{data}}$$

where the E_{type} is the type evaluation, for example, the evaluation of a type "face close-up". E_{data} is that of each *clues*, for example, the face size evaluation for a face close-up. The importance value used for each type in our experiments is shown in Table 10. The calculation of E_{data} is based on how each *clues* fits the category. In the case of face close-up, the importance evaluation is the weighted sum of the pixels which are occupied by a face close-up. Currently, E_{data} for each people image or outdoor scene image is 1.0, since those images are manually detected.

Similarly, E_{data} for key-sentence is calculated based on a keyword's part-of-speech, lexical meaning of subject, etc. An example of this cofficient is shown in Table 11.

[7] In our experiments, the threshold value is 20 seconds.

Table 5: Conditions for *key-sentence* detection

type	condition
SPEECH OPINION	active voice and affirmative, not future tense, subject as a human or a social group, not "it"
MEETING CONFERENCE	affirmative, not future tense
CROWD	affirmative, not future tense
VISIT TRAVEL	affirmative, not future tense, subject as human, at least one location name in a sentence
LOCATION	preposition (in, at, on, to, etc.) + location name

Table 6: Key-sentence detection result

	speech	meeting	crowd	visit	location
Video1	40/3/1	20/1/0	33/4/0	41/33/0	89/59/5
Video2	28/3/0	22/6/0	24/3/0	39/34/1	65/39/2
Video3	34/5/1	15/2/1	22/2/0	39/33/0	70/50/4

key-sentence: X is the number of sentences which include keywords; Y is the sentences removed by the above keyword screening; Z is the number of sentences incorrectly removed[4].

Image Clue Detection

A dominant portion of a news video is occupied by human activities. Consequently, human images, especially faces and human figures, have important roles. In the case of human visits or, movement outdoor scenes carry important information: who went where, how was the place, etc. We consider this a unit of *image clues*, and we call it a *key-image*.

Key-image

In this research, three types of images, face close-ups, people, and outdoor scenes are considered as *image clues*. Although these *image clues* are not strong enough for classifying a topic, there usage has a strong bias to several typical situations. Therefore, by associating the *key-image*s and *key-sentence*s, the topic of an image can be clarified, and the focus of the news segment can be detected.

The actual usage of the three kinds of images are shown in Table 7, 8 and 9. Among them, the predominant usage of face close-ups is for speech, though a human face close-up has the role of identifying the subject of other acts: a visitor of a ceremony; a criminal for a crime report, etc. Similarly, an image with small faces or small human figures suggests a meeting, conference, crowd, demonstration, etc. Among them, the predominant usage is the expression for a meeting or conference. In such a case, the name of a conference such as "Senate" is mentioned, while the people attending the conference are not always mentioned. Another usage of people images is the description about crowds, such as people in a demonstration.

In the case of outdoor scenes, images describe the place, the degree of a disasters, etc. Since the clear distinction of the roles is difficult, only the number of images with outdoor scenes is shown in Table 9.

[4] In this evaluation, difficult and implicit expressions which do not include words implying the *clues*. Therefore, we assume the keyword spotting results include all of the needed *language clues*.

Table 7: Usage of face close-up

video	speech	others	total
Video1	59	10	69
Video2	80	12	92

Other usages are personal introduction(4), action(2), audience/attendee(3), movie(2), anonymous(2), exercising(2), sports(1), and singing(4).

(a) (b)

Figure 4: Example of people images

Key-image Detection

First, the videos are segmented into cuts by histogram based scene change detection(SH95; HS95); The tenth frame[5] of each cut is regarded as the representative frame for the cut. Next, the following feature extractions are performed for each representative frame.

Face Close-up Detection In this research, human faces are detected by the neural-network based face detection program(RBK96). Most face close-ups are easily detected because they are large and frontal. Therefore, most frontal faces[6], less than half of small faces and profiles are detected.

People Image and Outdoor Scene Detection As for images with many people, the problem becomes

[5] The first few frames are skipped because they often have scene change effects.

[6] As described in (RBK96), the face detection accuracy for frontal face close-up is nearly satisfactory.

Table 8: Usage of people images

video	meeting	crowd	total
Video1	16	16	32
Video2	9	43	52

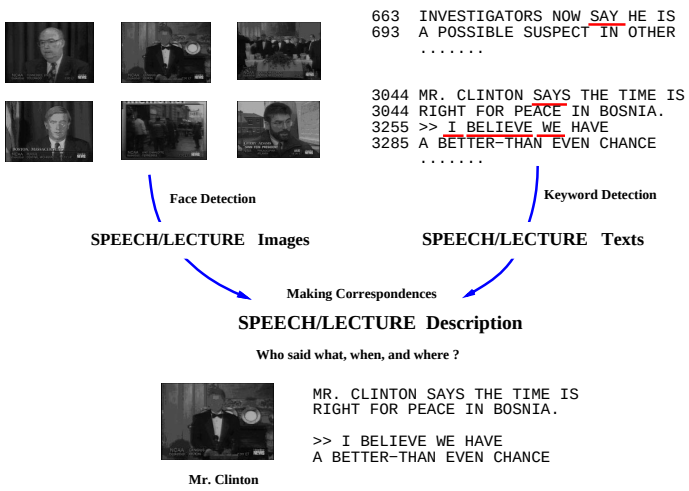

Figure 3: Basic idea of Spotting by Association

Table 2: Example of speech sentences

- MR. CLINTON SAYS THE TIME IS RIGHT FOR PEACE IN BOSNIA.
- TOMORROW, MR. CLINTON TALKS PEACE IN ANOTHER PART OF EUROPE.
- I THINK IT'S FOR PUBLICITY, FOR HIMSELF TO GET THE IRISH VOTE IN THE U.S., TO BE HONEST.
- I WAS ON THE EDGE AND DIDN'T KNOW IT.

lecture scene at a rate of 92%. Some words suggesting meeting/conference, crowd, visit/travel situations are shown in Table 4. Similarly, a location name often appears with outdoor scenes that are the actual scenes of that location.

Screening Keywords

As we can see in Table 3, some words such as "talk" are not sufficient keys. One of the reasons is that "talk" is often used as a noun, such as "peace talk". In such a case, it sometimes mentions only the topic of the speech, not the speech action itself. Moreover, negative sentences and those in future tense are rarely accompanied by the real images which show the mentioned content. Consequently, keyword spotting may cause large amount of false detections which can not be recovered by the association with image data.

To cope with this problem, we parse a sentence in transcripts, check the role of each keyword, and check the semantics of the subject, the verb, and the objects. Also, each word is checked for expression of a location.

1. Part-of-speech of each word can be used for the keyword evaluation. For example, "talk" may be better evaluated when it is used as a verb.
2. If the keyword is used as a verb, the subject or the object can be semantically checked. For example, the subject must be a human(s) or a representative of a social organization in the case of SPEECH/OPINION *clues*. For this semantic check, we use the *Hypernym* relation in the WordNet(Mil90): Word A is a hypernym of word B if word A is a superset or generalization of word B; Therefore, if one of the hypernyms of the subject word is "human" or "person", etc., the subject can be considered as a human(s).
3. Negative sentences or those in future tense can be ignored.
4. A location name which follows several kinds of prepositions such as "in", "to" is considered as a *language clue*.

Table 3: Keyword usage for speech

Indirect Narration

word	speech	not speech	rate
say	118	11	92%
tell	28	3	90%
claim	12	6	67%
talk	15	37	29%

Direct Narration or Live Video

word	speech	not speech	rate
I (my, me)	132	16	89%
we (our, us)	109	37	75%
think	74	15	84%
believe	12	10	55%

Table 4: Keyword usage for meeting and visiting

word	human meet	others	rate
meet	31	9	78%
see	15	59	20%
word	human visit	others	rate
visit	21	1	95%
come	30	62	32%

Process

In *key-sentence* detection, keywords are detected from transcripts. Separately, transcripts are parsed by the Link Parser(ST93). Keywords are syntactically and semantically checked and evaluated by using the parsing results. Since the transcripts of CNN Headline News are rather complicated, less than one third of the sentences are perfectly parsed. However, if we focus only on subjects and verbs, results are more acceptable. In our experiments, subjects and verbs are correctly detected at a rate close to 80%.

By using these results, part of speech of each keyword, and lexical meanings of the subject, verb, and object in a sentence are checked. The words to be checked and the conditions are listed in Table 5. A sentence including one or more words which satisfy these conditions is considered a *key-sentence*.

The results are shown in Table 6. The figure (X/Y/Z) in each table shows the numbers of detected

(a) (b) (c) (d)

Figure 1: Example of images in news videos

SPEECH/OPINION & FACE — Who spoke what? Where?

MEETING/CONFERENCE & PEOPLE — Who met whom? What subject?

OUTDOOR SCENE & LOCATION — Where? What event? Who visit where?

Figure 2: Typical situations

Situation Spotting by Association

From the above discussion, it is clear that the association between language and image is an important key to video content detection. Moreover, we believe that an important video segment must have mutually consistent image and language data. Based on this idea, we propose the "Spotting by Association" method for detecting important *clues* from each modality and associating them across modalities. This method has two advantages: the detection can be reliable by utilizing both images and language; the data explained by both modalities can be clearly understandable to the users.

For the above *clues*, we introduce several categories which are common in news videos. They are, for language, SPEECH/OPINION, MEETING/CONFERENCE, CROWD, VISIT/TRAVEL, and LOCATION; for image, FACE, PEOPLE, and OUTDOOR SCENE. They are shown in Table 1.

Inter-modal coincidence among those *clues* expresses important situations. Examples are shown in Fig.2. A pair of SPEECH/OPINION and FACE shows one of the most typical situation, in which someone talk about his opinion, or reports something. A pair of MEETING/CONFERENCE and PEOPLE show a conventional situation such as the Congress.

A brief overview of the spotting for a speech or lecture situation is shown in Fig.3. The *language clues* can be characterized by typical phrases such as "He says" or "I think", while *image clues* can be characterized by face close-ups. By finding and associating these images and sentences, we can expect to obtain speech or lecture situations.

Language Clue Detection

The transcripts of news videos are automatically taken from a NTSC signal, and stored as text. The simplest way to find *language clues* is keyword spotting from the texts. However, since keyword spotting picks many unnecessary words, we apply additional screening by parsing and lexical meaning check.

Table 1: Clues from language and image

language clues	
SPEECH OPINION	speech, lecture, opinion, etc.
MEETING CONFERENCE	conference, congress, etc.
CROWD PEOPLE	gathering people, demonstration, etc.
VISIT/TRAVEL	VIP's visit, etc.
LOCATION	explanation for location, city, country, or natural phenomena

image clues	
FACE	human face close-up (not too small)
PEOPLE	more than one person, faces or human figures
OUTDOOR-SCENE	outdoor scene regardless of natural or artificial.

Simple Keyword Spotting

In a speech or lecture situation, the following words frequently appear as shown in Table 2[2].

indirect narration: say, talk, tell, claim, acknowledge, agree, express, etc.

direct narration: I, my, me, we, our, us, think, believe, etc.

The first group is a set of words expressing indirect narration in which a reporter or an anchor-person mentions someone's speech. The second group is a set of words expressing direct narration which is often live video portions in news videos. In those portions, people are usually talking about their opinions.

The actual statistics on those words are shown in Table 3. Each row shows the number of word occurrences in speech portions or other portions[3]. This means if we detect "say" from an affirmative sentence in the present or past tense, we can get a speech or

[2]Since they are taken from closed-caption, they are all in upper case.

[3]In this statistics, words in a sentence of future tense or a negative sentence are not counted, since real scenes rarely appear with them.

Spotting by Association in News Video

Yuichi NAKAMURA
Institute of Information Sciences and Electronics,
University of Tsukuba,
Tsukuba City, 305, Ibaraki, JAPAN
(yuichi@image.is.tsukuba.ac.jp)

Takeo KANADE
The Robotics Institute,
Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213
(tk@cmu.cs.edu)

Abstract

This paper introduces the Spotting by Association method for video analysis, which is a novel method to detect video segments with typical semantics. Video data contains various kinds of information by means of continuous images, natural language, and sound. For use in a Digital Library, it is essential to segment the video data into meaningful pieces. To detect meaningful segments, we should associate data from each modality, including video, language, and sound. For this purpose, we propose a new method for segment spotting by making correspondences between *image clues* detected by image analysis and *language clues* created by natural language analysis. As a result, relevant video segments with sufficient information in every modality are obtained. We applied our method to closed-captioned CNN Headline News. Video segments with important situations, that is a speech, meeting, or visit, are detected fairly well.

Introduction

Recently, a large amount of video data has been gathered in Digital Libraries, and prepared for public or commercial use. The Informedia project(WKSS96) is one of the Digital Libraries, in which news and documentary videos are stored. Its experimental system provides news and documentary video retrieval by user queries from text or speech input.

Since the amount of data is enormous, efficient retrieval techniques are becoming more and more important. Data presentation techniques are also required to show large amounts of data to the users. Suppose that we are looking for video portions in which the U.S. president gave a talk about Ireland peace at some location. Then, if we simply ask "Mr. Clinton" and/or "Ireland" from news data in 1995 or 1996, we might get hundreds of video portions. It may take a considerable amount of time to find the right data. In this sense, we need two kinds of data management: One is semantical organization and tagging of the data; The other is data presentation that is structural and clearly understandable.

In this paper, we introduce the novel method to analyze the structure of news video data. This method is aimed to make the retrieval process more efficient and to meet more complicated query requests. First, we define *language clues* and *image clues* which are common in news videos, and introduce the basic idea of situation detection. Then, we describe inter-modal association between images and language. By this method, relevant video segments with sufficient information in every modality are obtained.

We applied our method to closed-captioned CNN Headline News. In the experiment, segments with typical important situations, such as a speech, meeting, or visit, etc. are detected fairly well.

Video Content Spotting by Association

Necessity of Multiple Modalities

When we see a news video, we can partially understand topics even if images or audio is missing. For example, when we see an image as shown in Fig.1(a), we guess that someone's speech is the focused. A face close-up and changes in lip shape is the basis of this assumption. Similarly, Fig.1(b) suggests a car accident and the extent of damage, though the suggestion is not always correct[1].

However, video content extraction from only language or image data may be misleading. Suppose that we are trying to detect a speech or lecture scene. With recent techniques in computer vision, face detection is not intractable. However, this is not enough. For example, Fig.1(c) is a face close-up; it is a criminal's face, and the video portion is devoted to a crime report. The same can be said about the language portion. Suppose that we need to detect someone's opinion from a news video. A human can do this perfectly if he reads the transcript and considers the contexts. However, current natural language processing techniques are far from human ability. Considering a sentence which starts with "They say", it is difficult to determine, without deep knowledge, whether the sentence mentions a rumor or is really spoken as an opinion.

[1] Actually, the car was exploded by a missile attack, not by a car accident.