Video Content Processing for Audiovisual Education

Yuichi NAKAMURA Yuichi OHTA
Institute of Information Sciences and Electronics
University of Tsukuba
Tsukuba 305-8573, JAPAN {yuichi@is.tsukuba.ac.jp}

1 Introduction

Recently, various multimedia tools are becoming available on personal computers. Advances on computer hardwares enabled us to handle complex media. Video is one of the most important sources for that purpose. Digital video archiving is now a hot topic, and not a few research projects are ongoing, and expected to greatly contribute to educations.

However, we have not fully exploited the potential of videos on computers. Finding appropriate videos out of a library and detecting appropriate portions in them costs much time and care. We also have problems in making our original video contents for audiovisual education, since it needs much intensive work in planning, recording, editing, and so on.

Thus we still have difficulties in using videos as educational sources, while humans are accumulating great amount of knowledge in terms of video every day. We need intensive research for analyzing videos, and convert them into effective media for practical use.

In this paper, we will introduce our approaches to these problems.

2 Toward Effective Video Usage

A video is a continuous medium. It enable us to record movements which are difficult to record in other media. A video can contain image, sound, and speech. It is effective to give an explanation of events by using the above modality combination.

These characteristics, however, cause serious drawbacks. Since the data can be large and redundant, to look through a video often requires too much time even when we want only a portion of the content. From the viewpoint of production, videotaping requires intensitive and intelligent work which is usually performed by cameramen and editors.

For the above reasons, it has been strongly desired to investigate video browsing, tagging, and summarization. Video production assistance and semantic tagging through the production process is also necessary.

Thus not a few methods have been developed to cope with these problems. For example, some cut/shot detection algorithms, which detect primitive boundaries of a video, show good performance, and are now the bases for video structure analysis. Video browsing, which provides the cut/shot catalog of a video, is also becoming a common technique.

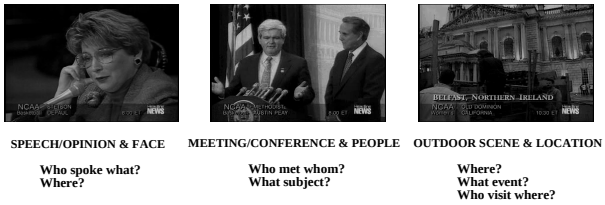


Figure 1: Typical situations

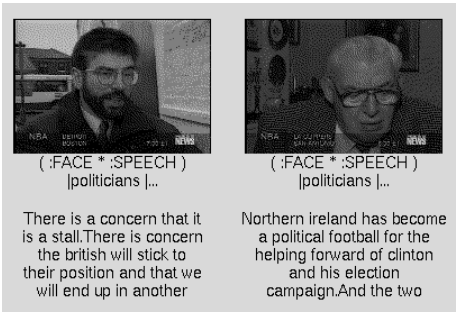


Figure 2: Speech scenes in a news video

However, there are still left open problems: video story segmentation, summarization, semantic tagging, and retrieval are required for the effective video as a knowledge sources.

Our researches aim to cope with the above unsolved problems. One is on story segmentation scheme for detecting important and meaningful pieces from videos[3]. The system is designed to detect typical scenes as shown in Figure 1. The system detects important segments of these categories, whose example is shown in Figure 2. The figure shows two segments, each of which is focused on someone's speech/opinion. Then, the users easily recognize and memorize "who said what".

Another research we will present in this paper is video production. An overview is shown in Figure 3. In this research, we are investigating the basis for realizing *virtual cameramen* and *virtual editors* who capture and emphasize the right place to which attention should be paid. In other words, this research aims at the mechanism for observing human activities in a similar way people would pay attention if they present, and the mechanism for effectively communicating the activities.

In the followings sections, we will describe the video analysis scheme in Section 3, and the production scheme in Section 4.

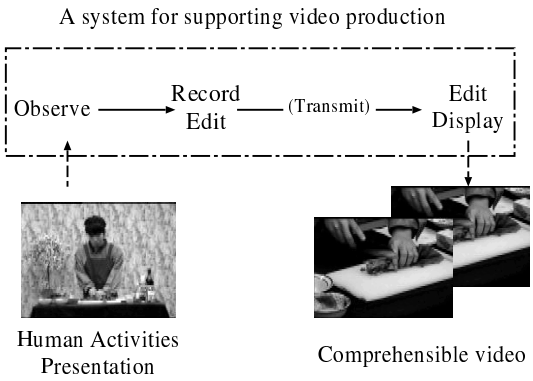


Figure 3: A system for supporting human communication

3 Video Structure Analysis

Digital Libraries gather a large amount of video data for public or commercial use. Since the amount of data stored in the libraries is enormous, in addition to efficient retrieval, data presentation techniques are also required to show large amounts of data to the users. Suppose a user is looking for video portions in which the U.S. president gave a talk about Ireland peace at some location. Then, if the user simply asks video segments related to "Mr. Clinton" and/or "Ireland" from news data in 1995 or 1996, hundreds of video segments may be retrieved. It may take a considerable amount of time to find the right data from that set. In this sense, we need two kinds of data management. One is semantical organization and tagging of the data, and the other is data presentation that is structural and clearly understandable.

For this purpose, it is effective to detect a topic essence in terms of one to several representative pairs of image and language data, for example, three pairs of a picture and a sentence. Image and language data corresponding to the same portion of a story should be chosen in this selection. These segments are the portions which the film/TV producers want to report, and are the portions which are easily understandable even when they are shown separately from others. Therefore, to detect those segments and to organize video archives based on them will be an essential technique for digital video libraries.

We introduce the Spotting by Association method, which detects relevant video segments by associating image data and language data. This method is aimed to make the retrieval process more efficient and to allow for more sophisticated queries. 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 from every modality are obtained.

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

Figure 4: Example of images in news videos

3.1 Video Content Spotting by Association

When we see a news video, we can understand topics at least partially, even if either images or audio is missing. For example, when we see an image as shown in Figure 4(a), we guess that someone's speech is the focus. A facial close-up and changes in lip shape is the basis of this assumption. Similarly, Figure 4(b) suggests the news reports a car accident and the extent of damage[1].

However, video content extraction from only language or image data is not reliable. Suppose that we are trying to detect a speech or lecture scene. Figure 4(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.

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 Figure 1. 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

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

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.

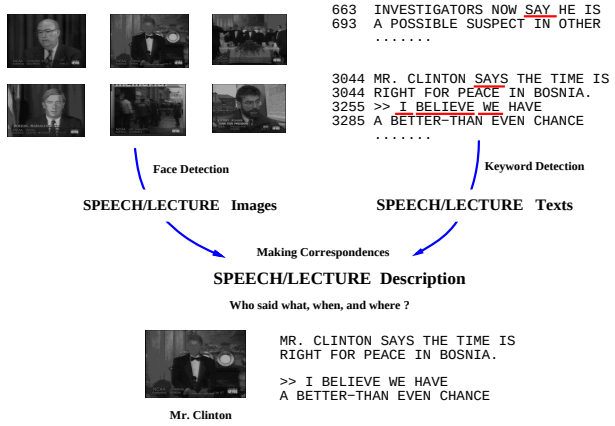


Figure 5: Basic idea of Spotting by Association

situation is shown in Figure 5. 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.

3.2 Language Clue Detection

The transcripts of news videos are automatically taken from a NTSC signal, and stored as text. The simplest way to detect *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.

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.

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

Table 2: Example of speech sentences

- MR. CLINTON SAYS THE TIME IS RIGHT FOR PEACE IN BOSNIA.
- I THINK IT'S FOR PUBLICITY, FOR HIMSELF TO GET THE IRISH VOTE IN THE U.S., TO BE HONEST.

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%

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 lecture scene at a rate of 92%.

We manually pick up those words by consulting a thesaurus, and chose actual keywords according to the statistics. The keywords suggesting MEETING/CONFERENCE, CROWD, VISIT/TRAVEL situations were chosen in the same way.

Screening Keywords

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 a 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 (details are shown in [3]) .

3.3 Image Clue Detection

In this research, three types of images, face close-ups, people, and outdoor scenes are considered as *image clues* and we call them *key-image*s. Although these *image clues* are not strong enough for classifying a topic, there usage has a strong bias to several typical situations.

The actual usage of face close-ups is shown in Table 4. The predominant usage of face close-ups is for speech, though a human face close-up has the role of identifying

[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.

Table 4: 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).

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.

First, the videos are segmented into cuts by histogram based scene change detection [6, 2]; The tenth frame[4] of each cut is regarded as the representative frame for the cut. Next, the following feature extractions are performed for each representative frame.

Human faces are detected by the neural-network based face detection program [5]. Most face close-ups are easily detected because they are large and frontal. Therefore, most frontal faces[5], less than half of the small faces and profiles are detected.

Automatic small face detection 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.

3.4 Association by DP

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 have 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 a *key-image* sequence and that of a *key-sentence* sequence are the same. In other words, there is no reverse order correspondence. Consequently, dynamic programming can be used to find the correspondence.

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 correspondence, $Skip_{\mathrm{i}}$ is for a *key-image* without inter-modal correspondence, and $Match(j,k)$ is the penalty for the correspondence between the j-th *key-sentence* and the k-th *key-image*. The value is basically determined by the durations and the categories of *clue*s.

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[6]. Any *key-sentence* or *key-image* may be skipped (warped), that is left unmatched.

3.5 Experiments

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

Figure 6 shows the association process by DP. The columns show the *key-sentence*s and the rows show *key-image*s. The correspondences are calculated from the paths' cost. In this example, 167 *key-image*s, 122 *key-sentence*s are detected; 69 correspondence cases are successfully obtained.

One of the results is shown in Figure 7. Each pair of a picture and a sentence is an associated pair. The picture is a *key-image*, and the sentence is a *key-sentence*. The position of the pair is determined by the situations defined in Section 3.1: segments for VISIT/TRAVEL or LOCATION are placed in the top row; the MEETING or CROWD segments are in the second row; SPEECH/OPINION segments are in the bottom row. Thus, the first row shows Mr. Clinton's visit to Ireland and the preparation for him in Belfast; the second row explains the politicians and people in that country; the third row shows each speech or opinion about Ireland peace. As we can see in this example, we can grasp the rough structure of the topic by taking a brief look at the explainer.

4 Presentation Video Production

TV programs and videos are produced by intensive work of cameramen and editors. Cameramen move their cameras, and changing the camera angles and zoom factors. By this framing work, important portions are tracked and captured in a frame. Editors select the best view, *i.e.* they determine which image to use out of the images from multiple cameras. Moreover, they usually remove unnecessary or redundant shots or scenes. By these works, an audience's attention is drawn to the right portion.

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

[5]As described in [5], the face detection accuracy for frontal face close-up is nearly satisfactory.

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

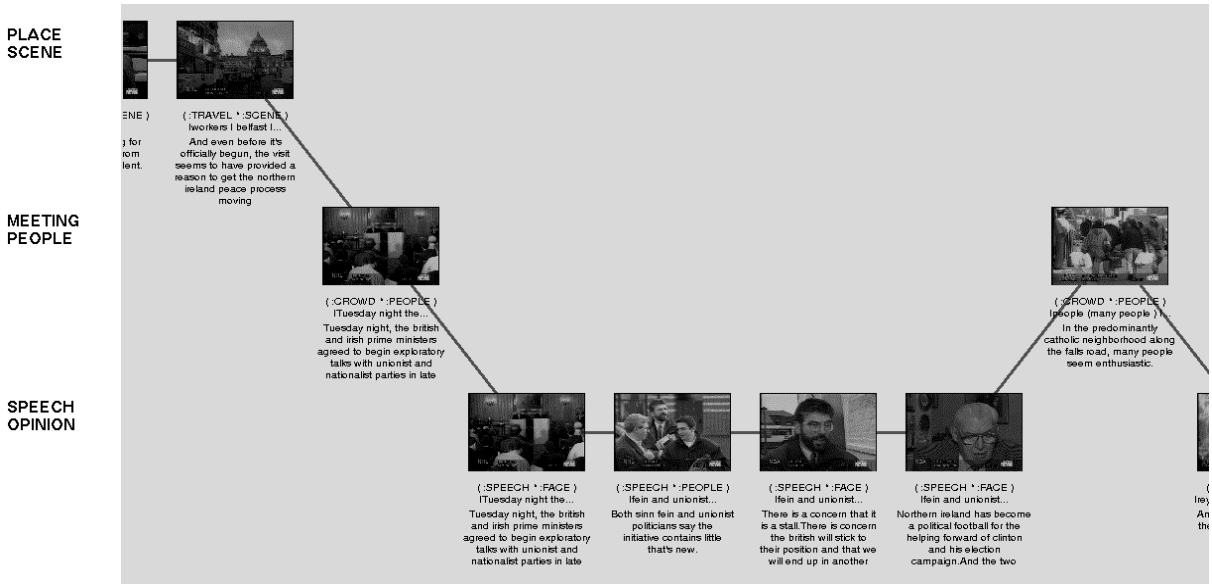


Figure 7: News video TOPIC EXPLAINER

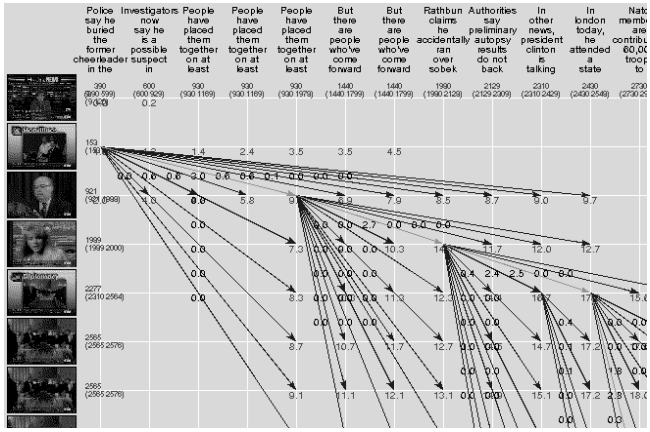


Figure 6: Correspondence between sentences and images

In this way, the intelligent videotaping system needs the function for interpreting a subject's intention, and reporting or recording the right focus[7]. Summarization is also important if we want to keep the audience's attention.

The requirements for the intelligent videotaping support system can be listed as follows.

- a function for interpreting human intentions
- a function for recognizing the focus of attention
- a function for emphasizing the focus of attention

The most useful keys to interpret human intentions are gestures and speech. Recently, many researches are devoted to recognizing gestures which have predefined meaning such as a sign language. In artificial and simple situations, some of them are at the level of practical use.

However, in unconstrained presentation or demonstration, humans use gestures in much more flexible way, often cooperatively with speech or other modalities. In this sense, if we want to understand human gestures as an effective means of communications, we need to research on how gestures are used in such situations. Moreover, since human behaviors are heavily dependent on environments around him/her, we need to investigate the variety of situations.

To cope with this problem, we first built a prototype system for recording human behavior in multimodal way. We gathered the records of human behaviors in the context of presentation, and built a prototype database named Multimodal Multi-view Integrated Database (MMID)[4]. Then, we have been investigating presentation video production by recognizing the speakers intentions.

4.1 Multimodal Data Acquisition

We developed a multimodal data recording system, which records videos, human body motions, motion labels, audio, and speech transcripts are recorded.

Videos are taken from multiple (currently 6) cameras. Each camera aims at a different portion of the presenter scene: the whole scene; the upper half of the body; the right hand; the left hand; the objects on a table; a stage view from the left side. The angle of the cameras for the upper body, hands, and objects are controlled by a host computer by using the position of those parts measured by the magnetic sensors. In this way, the multiple cameras shoot at all important portions which potentially attract viewer's attention. An example of the views is shown in Figure 8. Videos are digitized into MPEG or Motion JPEG. The videos from the multiple cameras for each scenario are completely synchronized.

Rough transcripts were prepared beforehand. They are modified if the speaker change the phrases or add a different phrase. An example of the transcript is shown in Table 5. Each line is separated such that it forms a phrase or a sentence. Each line contains a frame num-

[7]Hereafter, we use the word "focus" to represent the focus of attention which a person giving a presentation wants to show, and the audience should to look.

Figure 8: Example of multi-view video data

Table 5: Example transcript

7020	これは（あの）温サラダにする分です。 (These are for hot salad.)
7101	すでに洗ってありますので、 (There are already washed)
7215	はい、（えーっと）まず葉先と軸とを分けます。 (Then, (well), first split the leaves and the stem)

ber, which is the time code in the corresponding video(s) added afterward, and several words which the presenter spoke at that moment.

Motions are measured by multiple (currently 6) magnetic sensors. Each sensor measures six degrees of freedom: the position (x, y, z) and the orientation $(rotation, role, azimth)$. The sensors are attached to the presenter's head, both hands, both shoulders, and the back. The measurement range is about 5m in radius, and the sampling rate is set to 30Hz. Since we currently attach 6 sensors on a speaker's body, 6 sequences of positional and orientational values are obtained, each of which includes 6 values at every 1/30 second. Each measurement is associated with the frame number of the corresponding video(s).

Motion labels are manually added to the motion records. Currently, we are using 19 categories for motion labels as shown in Table 6. They are chosen so that each of them can be an atomic operation which cannot be divided into other operations.

4.2 MMID

With the above system, we build a prototype database MMID, which contains presentation activities in terms of video, audio, motion captured data, and speech transcripts, all of which are related by their occurrence time.

The contents can be retrieved by specifying an example or a template in one of the modalities. From the retrieved data, we can easily overlook how multiple modalities are cooperatively used in human communications.

MMID can potentially give the following information.

1. Frequency of a specific gesture or a speech
2. Variations of gestures or speeches used for a specific purpose
3. Cooccurrence of gestures and speeches in a specific situation
4. Individuality or differences of gestures and speeches among persons

Table 6: Motion label

put, put-in, take-out, pick-up, cut, stab, push, hit, twist, fold, pull, rub, shake, knead, stir, scoop-up, turn-back, fix-to, turn-up

Thus if enough variety of data are stored in MMID, it can be a good tool for designing user interface or multimedia contents handling system.

Contents

Currently, MMID has two kinds of data. One is a collection of original presentations, and the other is a collection of cooking shows from TV broadcast programs.

In our original presentations, the presenter demonstrates many kinds of gestures which we usually see in actual presentations; for example, deictic movement (pointing gestures), spatial movements, and pictographs. Each scenario has from 30sec to 2min length, and was played by 6 different people. The total length is about 50min (8.5min/person $\times$ 6 person).

In addition, the cooking shows are recorded from TV broadcast, and transcripts are manually added. Each of them is 25min in length (total 200min). Motions contained in those data are mainly operations by hands such as cutting some materials. The cooking show data lack the motion data and simultaneous multiple views compared with the original presentation data. However, they are good sources because they are easy to record, and speeches and acts inside them seem to be natural. Moreover, a demonstrator usually describe his movements by his speech. For example, a demonstrator puts an egg into a bowl saying "then, put this into water" at the same time. In this sense, a cooking show is one of the most useful data in which speeches and motions are mostly synchronous.

Content Retrieval and Display

For this retrieval, we can think query and retrieval schema as follows.

Motion: Gesture or posture detection and retrieval by searching similar motion sequence.

Transcript: Retrieval for word, morphological form, and case. Semantic retrieval for specific situations, such as cutting or assembling.

Video: Similar scene retrieval, face detection, object detection, etc.

Currently, motion, posture, and transcript retrieval have been implemented. Others and query by a combination of multiple modalities are under development.

4.3 Focus detection

One of the most important portion in the presentation is an object or place which is pointed by the speaker. A movement itself is sometimes the focus, when a speaker is performing an important operation. We are now investigating both of the above situations.

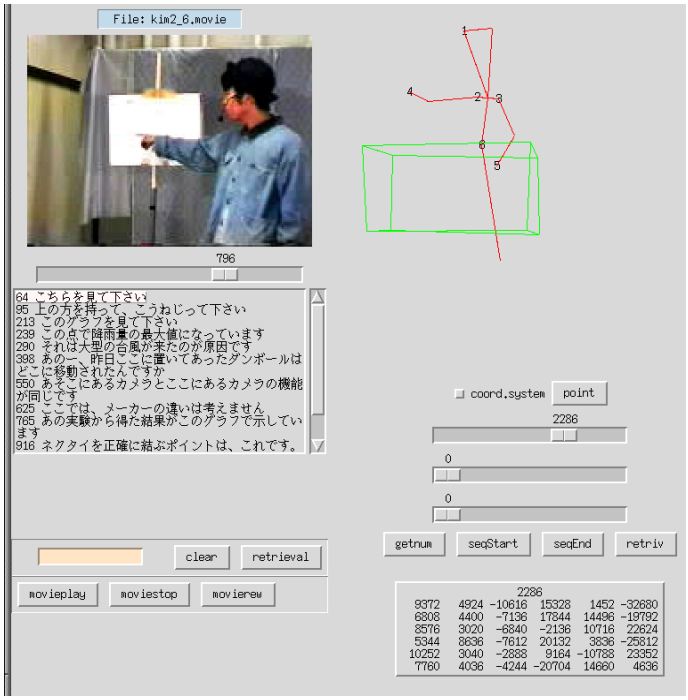


Figure 9: Example view of comprehensive display

Pointing Gesture Detection

The target is the recognition of gestures for pointing an object, a direction, or a location. The most typical case is the situation in which a person is saying "this" or "that" with his hand stretched. The problem is, however, not so simple. We have to check a variety of situations.

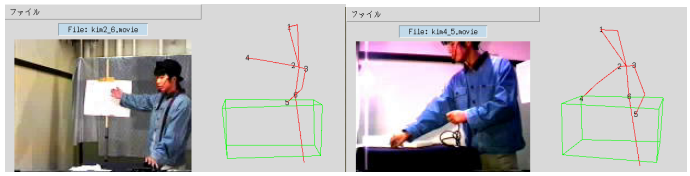
From the language point of view, we can check how demonstrative pronouns are used. "KORE" in Japanese usually has the same meaning as "this", and it is often used in a deictic way. Our statistics show 100 out of 162 cases[8] with "KORE" are deictic situations. In contrast to the above case, " それ (SORE)" has different characteristics, though "SORE" is often considered as "that" or "it". Only 3 out of 102 cases are deictic, and the predominant usage is anaphoric use[9]. This means that we need deeper analysis for deictic situation detection with "SORE" if we do not use gesture information.

On the other hand, situations with stretched hand are shown in Figure 10. 26 out of 78 cases taken from the original presentation are for pointing.

By combining features from speech and motion, the detection rate and error rate is improved. In our method, if features from motions and speech are close enough to each other, their scores, which supports the existence of a deictic motion, are summed up. The decision are made based on that score. This method uses the statistics gathered from MMID: the cooccurence rate of a demonstrative pronoun and a deictic movement; the cooccurence rate of each typical motion feature and a deictic movement; time difference between each typical motion feature and each demonstrative pronoun, etc.

[8]They are taken from 5 cooking show videos.

[9]referring a word or a sentence previously given.



(a) deictic case (b) simple operation

Figure 10: Example of typical gestures

これを入れます
(Put/pour this in)

野菜を加えます
(add some vegetables)

Figure 11: Example of put-in situations.

Important operation detection

There are typical important operations in presentations. In cooking show, for example, they are operations of cutting, putting, boiling, and so on.

Basically, an important situation, for example cutting operation, is well explained by speech. It is natural that a human describes his movements by his speech if he wanted to draw attention.

By combining motion detection and natural language analysis, we will obtain a small set of relevant segments from a large sequential medium. We are currently investigating the use of speech. Motion analysis is left for future work.

4.4 Editing/Summarizing Presentation

Our editing/summarizing scheme is based on the following idea:

- Choose the best view in which the focused object or place is captured at the best resolution.
- Choose the essential frames in which the focus is clear and important. We call these essential frames as "key-frame".

In this way, by choosing the view and the frame for reporting or recording, we eventually produce better videos than those taken with one fixed camera without edit.

view selection

In cooking shows, a person giving presentation usually want to show one of the following views: operations such as cooking, food, or overview. If operations are the focus, we can choose a camera which capture the motions of the person with the appropriate resolution. If food or materials are the focus, the person tends to specify them by deictic movement or typical words. On the contrary, when the person explains the overview of the cooking, he is not intending to show details on the table or his motions. In this case, wide-angled view of the studio is suited.

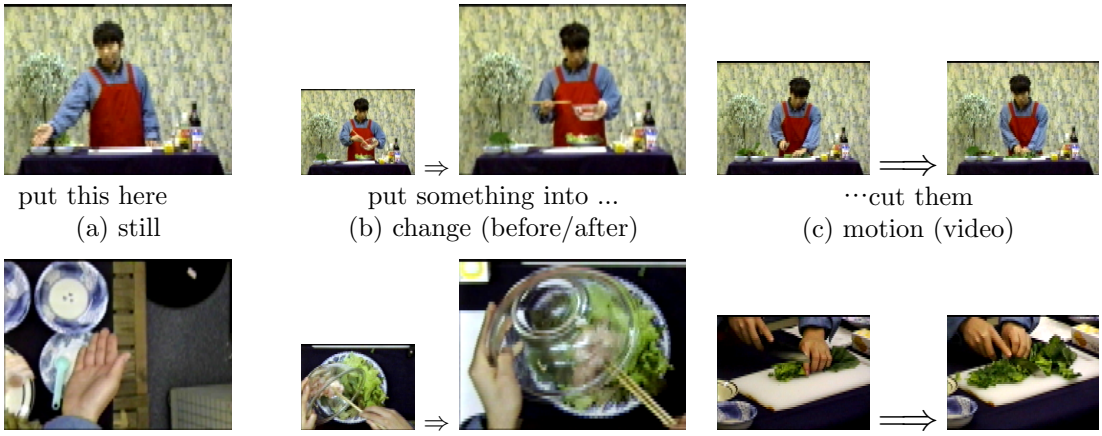


Figure 12: Focus and key-frame detection and display

Therefore, straightforward use of the above deictic movements detection and important operation detection can be a good basis for view selection.

key-frame selection

Similarly to the view selection, deictic movements and important operations are good clues for selecting key-frames.

At the current stage, the following frames are labeled as key-frames, and they are shown to the audience according to some criteria.

1. deictic movements
2. illustrator movements
3. movements to call the audience's attention
4. important operations

By collecting these frames and reducing redundant portions, we obtain a summarization of a presentation. Generally speaking, story structure analysis of the presentation is necessary for better summarization, this is left for future works.

examples

Here we show a simple example to demonstrate our framework. The above method works for a relatively simple input. The upper row in Figure 12 shows the result of key-frame selection. The images are taken from the wide-angle view of the presentation.

They are different in the sense how people should see. For the left case, a deictic movement requested the user to see a still object. The middle column gives the case that the state change is important. The still image of that object is usually enough. We need to look at the change between "before" and "after". In the case in right column, an operation which we need to see from its beginning to its end is detected. A moving image should be given for that portion.

The lower row shows the example in which view selection is added. As you can see in this example, the focus becomes much clearer than the upper column. We can easily notice where we should look.

The above example is a simple case, and we need to exploit more advanced scheme which works for more general cases.

5 Summary

In this paper, we introduced our approach to video analysis and production, which can potentially contribute to audiovisual education.

The Spotting by Association technique is the method for structure analysis of news videos. Internal stories in news videos are parsed and a comprehensible explanation is given to the users. MMID stores presentation activities in terms of audio, video, human body motion, and transcripts. MMID can serve as a basis for systematic and statistical analysis of those modalities. Our video editing/summarizing scheme supports video production. The system detects the focus of a presentation, selects the best view, and summarizes a video.

Since videos contain enormous information, we have many open problems on video analysis and production. Intensive researches are necessary for categorization, structure analysis, tagging, retrieval, and so on.

References

[1] A. Bobick. Movement, activity, and action. *MIT Media Lab Preceptual Computing Section*, TR-413, 1997.

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

[3] Y. Nakamura and T. Kanade. Semantic analysis for video contents extraction — spotting by association in news video. *ACM Multimedia*, pages 393–401, 1997.

[4] Y. Nakamura, Y. Kimura, Y. Ye, and Y. Ohta. Mmid: Multimodal multi-view integrated database for human behavior understanding. *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages pp.540–545, 1998.

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

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