Semantic Analysis for Video Contents Extraction — Spotting by Association in News Video

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

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

Abstract

Spotting by Association method for video analysis is a novel metliod to detect video segments with typical semantics. Video data contains various kinds of information through continuous images, natural language, and sound. For videos to be stored and retrieved in a Digital Library, it is essential to segment the video data into meaningful pieces. To detect meaningful segments, we need to identify the segment in each modality (video, language, and sound) that corresponds to the same story. For this purpose, we propose a new method for making correspondences between ***image clues*** detected by image analysis and *language clues* detected by natural language analysis. As a result, relevant video segments with sufficient information from every modality are obtained. We applied our method to closed-captioned CNN Headline News. Video segments with important events, such as a public speech, meeting, or visit. are detected fairly well.

1 Introduction

Digital Libraries gather a large amount of video data 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 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 wliicli 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 tlic 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 wliicli 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.

So far few researches have dealt with this problem. It is common to give a topic explanation by using the first. frame/image of the first cut/shot and the first sentence in a transcript. This representative pair is often a poor topic explanation, for example, an anchor person's close-up with a too much general description. To cope with image selection problem, Zhang, et. al, proposed a method for key-frame selection by using several image features such as colors, textures, and temporal features including caniera operations [ZLSW95]. Smith and Kanade proposed video skimming by selecting video segments based on TFIDF, camera motion, human face, captions on video, and so on [SK97]. By joining the selected segments, a new video which gives a rough idea about the topic is obtained. They are good techniques which are broadly applicable, since they do not require deep content analysis.

There are, however, still open problems to tackle. One is the semantic classification of each segment. For effective topic indexing or explanation, we need to know what a segment describes. Another is the correspondence problem between image and language. As mentioned above, we need to detect image and language data corresponding to the same portion of a story. If they are taken from different portions, the pair may become misleading to the users[1].

To handle these problems, we introduce the Spotting by Association method, which detests 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.

[1] For example, a close-up of a policeman's face and the name of the criminal.

By this method, relevant video segments with sufficient information from every modality are obtained.

We applied our method to closed-captioned CNN Headline News, from which segments with typical important situations, such as public speeches, meetings, or visits are detected fairly well.

2 Video Content Spotting by Association

2.1 Necessity of Multiple Modalities

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 Fig.1(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, Fig.1(b) suggests the news reports **a** car accident and the extent of damage'.

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. 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.

2.2 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 eacli 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. Esaniples 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.

Table 1: Clues from language **and** image

language clues	
SPEECH OPINION	speech, lecture, opinion, etc.
CONFERENCE	
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.

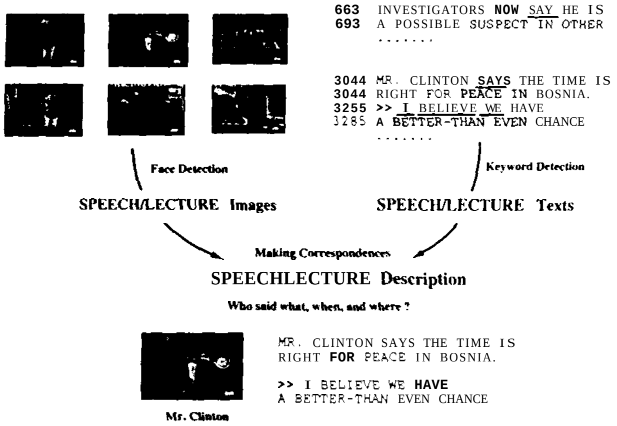
663 INVESTIGATORS NOW SAY HE IS
693 A POSSIBLE SUSPECT IN OTHER
....
3044 MR. CLINTON SAYS THE TIME IS
3044 RIGHT FOR PEACE IN BOSNIA.
3255 >> I BELIEVE WE HAVE
3285 A BETTER-THAN EVEN CHANCE
....
Face Detection
Keyword Detection
SPEECH/LECTURE Images
SPEECH/LECTURE Texts
Making Correspondences
SPEECHLECTURE Description
Who said what, when, and where ?
MR. CLINTON SAYS THE TIME IS
RIGHT FOR PEACE IN BOSNIA.
>> I BELIEVE WE HAVE
A BETTER-THAN EVEN CHANCE
Mr. Clinton

Figure 3: Basic idea of Spotting by Association

3 Language Clue Detection

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

3.1 Simple Keyword Spotting

In a speech or lecture situation, the following words frequently appear **as** shown in Table 2[3].

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 espressiiig 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 wliicli is often live vidco portions in news videos. In

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

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

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

Figure 1: Example of images in news videos

SPEECWOPINION & 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

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.

Table 3: Keyword usage for speech

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

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%
scc	15	59	20%

word	human visit	others	rate
visit	21	1	95%
come	30	62	32%

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[4]. 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%. 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.

3.2 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. hloreovcr, 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 wliicli can not be recovered by tlie association with image data.

To cope with this problem, we parse a sentence in tran-

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

scripts, check tlie role of each keyword, and check the semantics of tlie subject, tlie 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 tlie 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 wliicli follows several kinds of prepositions such **as** "in", "to" is considered **as** a ***language clue.***

3.3 Process

In ***key-sentence*** detection, keywords are detected from transcripts. Separately, transcripts are parsed by tlie Link Parser [ST93]. Keywords are syntactically and semantically checked and evaluated by using tlie parsing results. Since the transcripts of CNN Headline News are rather complicated, less than one tliird of tlie 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 usiug these results, part of speech of each keyword, arid lexical meanings of tlie subject, verb, and object in **a** sentence are checked. Tlic words to be cliecked and tlie conditions are listed in Table 5. A sentence including one or more words wliicli satisfy these coiiditions 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 ***key-sentence:*** X is the number of sentences which include keywords; Y is the sentences removed by the above keyword screening; Z is tlie number of sentences incorrectly removed[5].

4 Image Clue Detection

A dominant portion of **a** news video is occupied by human activities. Consequently, human images, especially faces and human figures, liave important roles. In tlie 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.***

Table 5: Conditions for ***key-sentence*** detection

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

	speech	**meeting**	**crowd**	**visit**	**location**
Videol	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

4.1 Key-image

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

The actual usage of tlie three kinds of images are shown in Table 7, 8 and 9. Among them, tlie predominant usage of face close-ups is for speech, though **a** liuman face close-up lias tlie role of identifying tlie subject of other acts: a visitor of **a** ceremony; **a** criminal for **a** crime report, etc. Similarly, an image witli small faces or small human figures suggests **a** meeting, conference, crowd, demonstration. etc. Among tliem, 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 tlie people attending tlie conference are not always mentioned. Anotlier 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, tlie 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.

4.2 Key-image Detection

First, the videos are segmented into cuts by liistogram based scene change detection [SH95, HS95]; The tenth frame[6] of each cut is regarded **as** the representative frame for tlie cut. Next, tlie following feature extractions are performed for each representative frame.

[5] **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.***

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

video	speech	others	total
Videol	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

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', less tlian half of the small faces and profiles are detected.

People Image and Outdoor Scene Detection

As for images with many people, the problem becomes difficult because small faces and human figures are more difficult to detect. The same can be said of 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.

5 Association by DP

The sequence of ***key-sentences*** and that of ***key-images*** are associated by Dynamic Programming.

5.1 Basic ldea

The detected data is the sequence of ***key-images*** and that of ***key-sentences*** to which starting and ending time is given. If **a** ***key-image*** duration and ***a bey-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-rmage*** 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.

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

(a) (b)

Figure 5: Example of outdoor scenes

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

$$P = \sum_{j\in Sn} Skip_s(j) + \sum_{k\in In} Skip_i(k) + \sum_{j\in S, k\in I} Match(j,k) \quad (1)$$

where S and **I** are the ***key-sentences*** and ***bey-images*** wliicli have corresponding ***clues*** in the other modality, Sn and In are those without corresponding *clues*. **Skip,** is the penalty value for **a** ***key-sentence*** without inter-modal correspondence, $Skip_i$ is for **a** ***key-image*** without inter-modal correspondence, and ***Match***(j, k) is tlie penalty for tlie correspondence between the j-th *key-sentence* and the k-th *key-image*.

In DP path calculation, we allow any inter-modal correspondence unless tlie duration of **a** ***key-image*** and that of a ***key-sentence*** are mutually ***too*** far to be matched[8]. Any ***key-sentence*** or ***key-image*** may be skipped (warped), that is left unmatched.

5.2 Cost Evaluation

Skipping Cost (*Skip*):

Basically, the penalty values are determined by the importance of the data, that is the possibility of each data having the inter-modal correspondence. In this research, importance evaluation of each *clue* is calculated by the following formula. The skip penalty ***Skip*** is considered **as** $-E$.

$$E = E_{type} \,.\, E_{data} \quad (2)$$

where the ***E,,,,*** is the type of evaluation, for example, the evaluation of **a** type "face close-up". E_{data} is that of each ***clue,*** 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 ***clue*** 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-sentences*** is calculated based on **a** keyword's part-of-speech, lesical meaning of subject, etc. An example of this coefficient is shown in Table 11.

'As **described** in [RBK96], **tlie face** detection **accuracy for frontal face** close-up *is* nearly **satisfactory.**

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

Table 9: Usage of outdoor scenes

video	outdoor scenes
Videol	34
Video2	39

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

Matching Cost (Match):

The evaluation of correspondence is calculated by the following formula.

$$Match(i,j) = M_{time}(i,j) . M_{type}(i,j) \qquad (3)$$

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

A *key-image's* duration $(d,)$ is the duration of the cut from wliicli tlie *key-image* is taken; the starting and ending time of a sentence in the speech is used for *key-sentence* duration (d_s). In tlie case where tlie esact speech time is difficult to obtain, it is substituted by tlie 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.

6 Experiments

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

They are segmented into cuts by scene change detection, then eacli poster frame, *i.e.* representative image for each cut is detected. Nest, the face detection, people detection, and outdoor scene detection are applied to each poster frame. Currently, only the face close-up detection is automated, the rest are created manually. Each data is registered as a *key-image,* then the importance is evaluated.

Transcripts are automatically obtained by closed-caption. They are segmented into sentences, and parsed by Link Parser. Then, through keyword detection and screening by checking semantics, *key-sentences* are detected. All transcript processing is done without human assistance, since the *key-sentence* detection results are satisfactory. For each *key-sentence,* importance is calculated similarly to the *key-image* evaluation. Finally, inter-modal correspondences between obtained *key-images* and *key-sentences* are calculated by DP.

6.1 Results

Fig.6 shows the association results by DP. The columns show the *key-sentences* and the rows show *key-images.* The correspondences arc calculated from tlie paths' cost. In this example, 167 *key-images*, 122 *key-sentences* are detected; 69 correspondence cases are successfully obtained.

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

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

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

- *key-image* or *key-sentence* detection errors
- 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 speecli or talk situation.

6.2 Usage of the Results

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

1. Summarization and presentation tool:
   Around 70 segments are spotted for eacli 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. Tliis 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.9. Each pair of a picture and a sentence is an associated pair. The picture is a *key-image*, and the

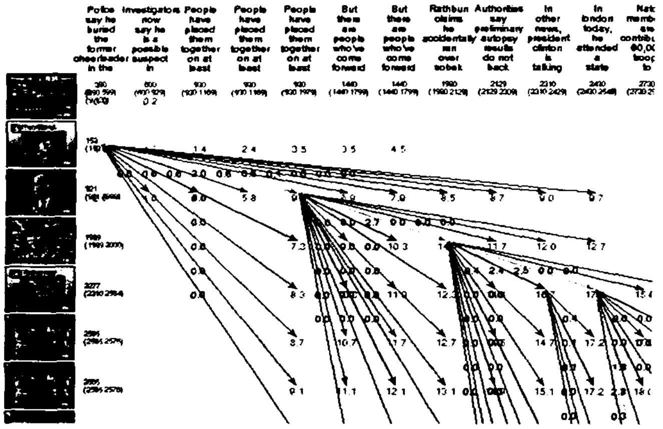

Figure 6: Correspondence between sentences and images

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	**452**	215	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 the number of *clues* for which association is failed in spite of having real correspondences,** E is the **number of wrong association**, *i.e.* **mismatching.**

sentence is a *key-sentence*. The position of the pair is determined by tlie situations defined in Section 2: 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, tlie 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.

In this view, the time order of segments is kept only inside each row. This is mainly for saving the space. If we keep the order across the row, *i.e.* if all tlie segments are placed in the order of their presented time, we get the view shown in Fig.8. This view enables us to overlook how the topic is organized. Visit and place information is given first., meeting information is given second, then **a** few public speeches and opinions are given. As we can see in this example, we can grasp the rough structure of the topic by taking **a** brief look at the explainer.

2. Data tagging **to** video segments:
   As mentioned before, tlie 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 these data, video segment retrieval can be much more efficient.

Table **14:** Spotting result 2

	face	people	scene
speech	**199 165**	**24 12**	
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, tlie number of **found correspondences** (X) **and the number** of **correct** correspondences ***(Y).***

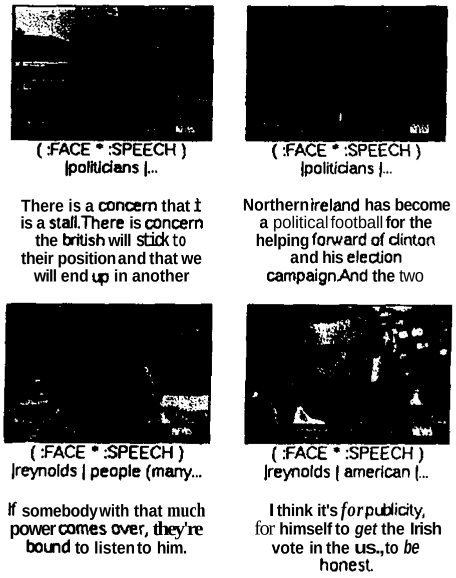
(:FACE * :SPEECH)
|politicians |...
There is a concern that i is a stall.There is concern the british will stick to their position and that we will end up in another
(:FACE * :SPEECH)
|politicians |...
Northern ireland has become a political football for the helping forward of clinton and his election campaign.And the two
(:FACE * :SPEECH)
|reynolds | people (many...
If somebody with that much power comes over, they're bound to listen to him.
(:FACE * :SPEECH)
|reynolds | american |...
I think it's *for* publicity, for himself to *get* the Irish vote in the us.,to *be* honest.

Figure 9: Details in TOPIC EXPLAINER

7 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 tlic most important areas is the improvement of *key-image* and *key-*sentence detection. Another is to check the effectiveness of this method with other kinds of videos.

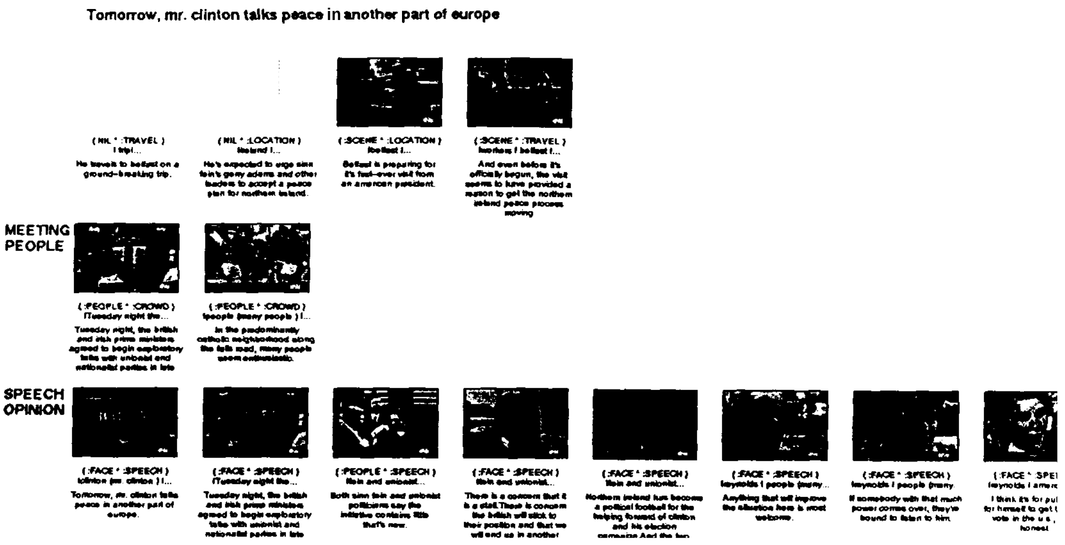

Figure 7: News video TOPIC EXPLAINER (Category)

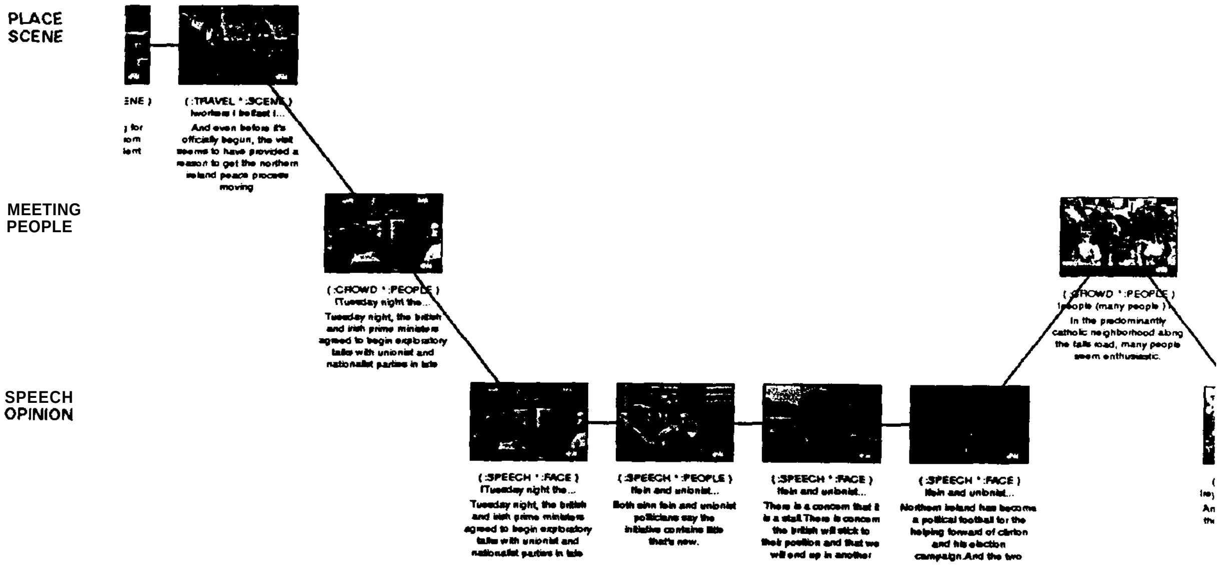

Figure 8: News video TOPIC EXPLAINER (Category + Time Order)

References

[HS95] Hauptmann, A. and Smith, M. Video Segmentation in the Informedia Project. In *IJCAI-95, Workshop on Intelligent Multimedia Information Retrieval*, 1995.

[Mil90] Miller, G. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, Vol. 3. No. 4, 1990.

[RBK96] Rowley, H., Baluja, A. and Kanade, T. Neural Network-Based Face Detection. *Image Understanding Workshop*, 1996.

[SH95] 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.

[SK97] Smith, M. and Kanade, T. Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques. *IEEE CVPR*, 1997.

[ST93] Sleator, D. and Temperley, D. Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies,* 1993.

[WKSS96] Wactlar, H., Kanade, T., Smith, M. and Stevens, S. Intelligent Access to Digital Video: The Informedia Project. *IEEE Computer,* Vol. 29, No. 5. 1996.

[ZLSW95] Zhang, H., Low, C., Smoliar, S. and Wu, J. Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution. *Proc. ACM Multimedia,* 1995.