Proceedings of ISDL'97, pp. 143–150, Tsukuba, Japan, 1997.

# Accessing Video Contents: Cooperative Approach between Image and Natural Language Processing

Takeo Kanade
Carnegie Mellon University
Pittsburgh, PA 15213, USA
tk@cs.cmu.edu

Shin'ichi Satoh
NACSIS
Tokyo 112, Japan.
satoh@rd.nacsis.ac.jp

Yuichi Nakamura
University of Tsukuba
Tsukuba, 305, Japan.
yuichi@is.tsukuba.ac.jp

## Abstract

Digital video libraries become much more important. In achieving them, access and extraction methods of semantic contents of videos are essential technologies. The paper demonstrates the benefits of multi-modal video analysis to extract semantic contents of videos. Two systems, Name-It and Spot-It, are introduced as example systems taking this approach. Name-It detects faces in news videos and associates with their names. Spot-It classifies video segments into several meaningful categories. Their results can enhance performance of both retrieval and presentation for digital video libraries. The successful results demonstrate importance of our approach.

Keywords: digital library, video analysis, image processing, natural language processing

## 1 Introduction

The Informedia project [1] is one of the digital library projects whose goal is to develop technologies for structuring the storage of a large amount of video data consisting of news and documentary videos; the data would then be available for convenient retrieval by public or commercial users. The project's experimental system allows users to retrieve news and documentary video by means of text or speech queries.

Video has several distinctive properties when compared with text documents: (1) Video is more effective, comprehensive and appealing to humans than text documents because of their spatio-temporal visual properties; (2) Information in video is more "raw" and dispersed, which makes it difficult to extract specific "keywords"; and (3) Video presents information linearly in time, which makes it more diffucult to scan for reviewing or browsing within a shorter period of time than the realtime length of video.

Accessing and extracting video contents is crucial for video digital libraries with effective and efficient retrieval and presentation abilities. Figure 1 shows several sample images taken from news videos. When we see the video segment shown in Figure 1(a), we usually guess that its content is someone's speech (and recognize the speaker as Bill Clinton). Likewise, we associate the segment of Figure 1(b) with an event that takes place in Congress, and Figure 1(c) with a car accident. Such associations of an events with video segments help retrieval and presentation for video digital libraries, since these associations can lead to "real" content-based retrieval, as opposed to "image" content-based retrieval.

(a) (b) (c)

Figure 1: Example segments of news videos

Automated extraction of content from videos is extremely difficult, if not impossible, at this point. Video, however, contains multi-modal information, including image sequences, audio (including speech), closed captions, and transcripts. Some important information in videos appears only in one modality or in multiple modality in different forms. Therefore, cooperative integration of multiple modalities, while each result may be imperfect, can be used to increase the capability and reliability in extracting video contents.

In this paper, we describe examples of multi-modal video understanding which integrates image and natural language processing techniques. The first example is a system called "Name-It," which extracts facial images from image sequences and name them from transcripts, then integrates this information to obtain face-name association. The Name-It system also attempts to support retrieval of image sequences by giving the names of people whose faces may appear in the video. The "Spot-It" system is the second system we studied. The system divides video into segments, each of which is classified into one of several types, such as someone's speech and Congress/meeting, by using several image and language clues. The system allows reorganization of a video into a form much more compact than the original, making it effective for video presentation.

These systems demonstrate how the content infor-

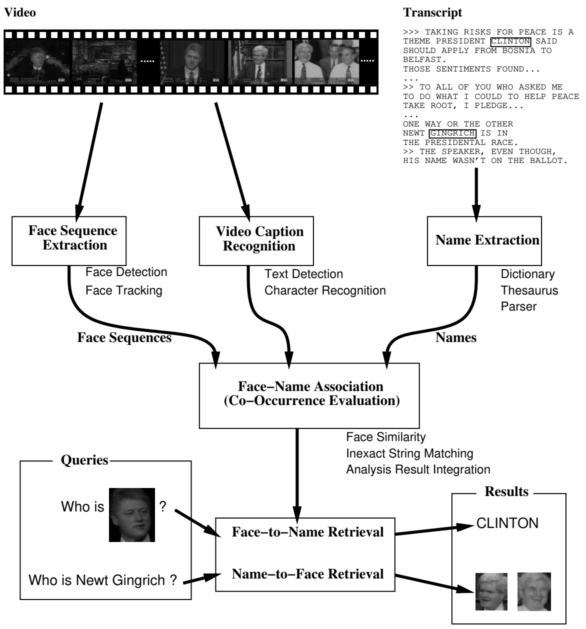



Figure 2: Architecture of Name-It

mation of video, such as face-name association or video segment classification, can be obtained by combining multi-modal information extraction.

# 2 Name-It: Detecting and Naming Faces in Video

## 2.1 Overview of Name-It

The purpose of Name-It is to associate names and faces in news videos [2, 3]. Several potential applications might include: (1) News video viewer which can interactively provide text description of the displayed face, (2) News text browser which can provide facial information of names, (3) Automated video annotation generation by naming faces.

To achieve Name-It system, we employ the architecture shown in Figure 2. Since we use closed-captioned CNN Headline News for our target, given news are composed of a video portion and a transcript portion. From video images, the system extracts faces of persons who might be mentioned in transcripts. Meanwhile, from transcripts, the system extracts words corresponding to persons who might appear in videos. Then, the system evaluates the association of the extracted names and faces. Both names and faces are extracted from videos, therefore, they furnish additional timing information, i.e., at what time in videos they appear. The association of names and faces is evaluated with a "co-occurrence" factor using their timing information. Co-occurrence of a name and a face expresses how often, and how well the name coincides with the face. In addition, the system also extracts video captions from video images. Extracted video captions are recognized to obtain text information, then used to enhance face-name association quality.

## 2.2 Image Processing

The image processing portion of Name-It is necessary for extracting faces of persons who might be mentioned in transcripts. Those faces are typically shown under the following conditions: (a) frontal, (b) close-up, (c) centered, (d) long duration, (e) frequently. Given a video as input, the system outputs a two-tuple list: timing information (start $\sim$ end frame), and face identification information. Some of the conditions above will be used to generate the list; others will be evaluated later using information provided by that list. The image processing portion also contributes for video caption recognition, which provides rich information for face-name association.

### 2.2.1 Face Tracking

To extract face sequences from image sequences, Name-It applies face tracking to videos. Face tracking consists of 3 components; face detection, skin color model extraction, and skin color region tracking.

First, Name-It applies face detection to every frame within a certain interval of frames, e.g., 10 frames. The system uses the neural network-based face detector [4] which detects mostly frontal faces at various sizes and locations. The face detector can also detect eyes; we use only faces in which eyes are successfully detected to ensure that the faces are frontal and close-up.

Once a face is detected, the system extracts a skin color model [3]. Once a face region is detected in a frame, the skin color model of the face region is captured as the Gaussian model in $(R, G, B)$ space. The model is applied to the subsequent frames to detect skin candidate regions. Face region tracking is continued until a scene change is encountered or until no succeeding face region is found.

### 2.2.2 Face Identification

To infer the "frequent" occurrence of a face, face identification is necessary. Namely, we need to determine whether one face sequence is identical to another.

To make face identification work effectively, we need to use frontal faces. The best frontal view of a face will be chosen from each face sequence [3]. We first apply the face skin region clustering method to all detected faces. Then, the center of gravity of the face skin region is calculated and compared with the eye locations to evaluate a frontal factor. The system then chooses the face having the largest frontal factor as the most frontal face in the face sequence.

We choose the eigenface-based method to evaluate face identification [5]. Each of the most frontal faces is converted into a point in the 16-dimensional eigenface space. Face identification can be evaluated as the face distance, i.e., the Euclidean distance between two corresponding points in the eigenface space.

#### 2.2.3 Video Caption Recognition

Video captions are directly attached to image sequences, and give text information. In many cases, they are attached to faces, and usually represent a person's name. Thus video caption recognition provides rich information for face-name association, though, they do not necessarily appear for all faces of persons of interest.

To achieve video caption recognition, the system first detects text regions from video frames. Several filters including differential filters and smoothing filters are employed to achieve this task. Clusters with bounding regions that satisfy several size constraints are selected as text regions. The detected text regions are preprocessed to enhance video caption image quality. First, the filter that minimize intensities among frames is applied. This filter suppresses complicated and moving background, yet enhances characters because they are placed at the exact position for a sequence of frames. Next, the linear interpolation filter is applied to quadruple the resolution. Then template-based character recognition is applied. Current system can recognize only upper-case letters, but it achieved 76% character recognition rate.

Since character recognition results are not perfect, inexact matching between the results and character strings is essential to utilize imperfect results for face-name association. We extended the edit distance method [6] to cope with this problem. Assume that $C$ is the character recognition result, and $N$ is a word. The similarity $S_c(C, N)$ is defined to represent that $C$ *might be* $N$.

## 2.3 Natural Language Processing

The system extracts name candidates from transcripts using natural language processing technologies. The system is expected not only to extract name candidates, but also to associate them with scores. The score represents the likelihood that the associated name candidate might appear in the video. To achieve this task, combination of lexical and grammatical analysis and the knowledge of the structure of news is employed.

First, the dictionary and parser are used to extract proper nouns as name candidates. The agent of an act such as speech or attending meeting obtains a higher score. In doing this, the parser and thesaurus are essential. In a typical news video, an anchor person appears first, talks about an overview of the news, and mentions the name of the person of interest. The system also uses news structure knowledge like this. Several such conditions are employed for score evaluation

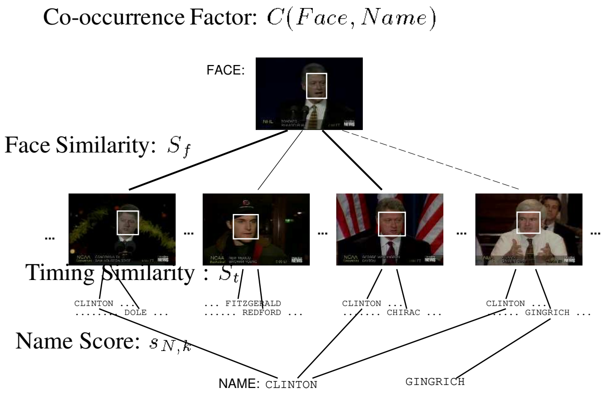



Figure 3: Co-occurrence Factor Calculation

[3]. The system evaluates these conditions for each word in the transcripts by using a dictionary (the Oxford Advanced Learner's Dictionary [7]), thesaurus (WordNet [8]), and parser (Link Parser [9]). Then, the system outputs a three-tuple list: a word, timing information (frame), and a normalized score.

The execution time for a 30-minute news video is approximately 1.5 hours on an SGI workstation (MIPS R4400 200MHz). Most of that time is consumed by parsing.

## 2.4 Integration

#### 2.4.1 Face-Name Association Algorithm

In this section, the algorithm for retrieving face candidates by a given name is described. We use the co-occurrence factor to integrate image and natural language analysis. Let $N$ and $F$ be a name and face, respectively. The co-occurrence factor $C(N, F)$ is expected to have a degree which represents the fact that the face $F$ is likely to have the name $N$. Think of the faces $F_a, F_b, \cdots$ and the names $N_p, N_q, \cdots$, and $F_a$ corresponds to $N_p$. Then $C(F_a, N_p)$ should have the largest value among co-occurrence factors of any combinations of $F_a$ and the other names (e.g., $C(F_a, N_q)$, etc.), or of the other faces and $N_p$ (e.g., $C(F_b, N_p)$, etc.). Retrieval of face candidates by a given name is realized as follows using the co-occurrence factor:

1. Calculate co-occurrences of combinations of all face candidates and the given name.

2. Sort co-occurrences.

3. Output faces that correspond to the N largest co-occurrences.

Retrieval of name candidates by a face is realized as well.

#### 2.4.2 Integration by Co-occurrence Calculation

In this section, the co-occurrence factor $C(N, F)$ of a face $F$ and a name $N$ is defined. Assume that we have the two-tuple list of face sequences (timing, face identification): $\{(t_{F_i}, F_i)\} = \{(t_{F_1}, F_1), (t_{F_2}, F_2), ...\}$, the three-tuple list of name candidates (word, timing, score): $\{(N_j, t_{N_j,k}, s_{N_j,k})\} = \{(N_1, t_{N_1,1}, s_{N_1,1}), (N_1, t_{N_1,2}, s_{N_1,2}), ...$ $(N_2, t_{N_2,1}, s_{N_2,1}), ...\}$, and the two-tuple list of video captions (timing, recognition result): $\{(t_{C_i}, C_i)\} = \{(t_{C_1}, C_1), (t_{C_2}, C_2), ...\}$. Note that $t_{F_i}$ and $t_{C_i}$ have duration, e.g., ($t_{start,F_i} \sim t_{end,F_i}$); so we can then define the duration function as $dur(t_{F_i}) = t_{end,F_i} - t_{start,F_i}$. Also note that a name $N_j$ may occur several times in a video, so each occurrence is indexed by $k$. We define the face similarity between faces $F_i$ and $F_j$ as $S_f(F_i, F_j)$ using the distance in the eigenface space. The caption similarity between a video caption recognition result $C$ and a word $N$, $S_c(C, N)$, and the timing similarity between times $t_i$ and $t_j$, $S_t(t_i, t_j)$, are also defined. The caption similarity is defined using the edit distance, and the timing similarity represents coincidence of events. Then the co-occurrence factor $C(N, F)$ of the face $F$ and the name candidate $N$ is defined as follows;

$$
\begin{aligned}
C(N, F) & \\
&= \frac{\sum_i S_f(F_i, F)(\sum_k s_{N,k} S_t(t_{F_i}, t_{N,k}) + *)}{\sqrt{\sum_i S_f^2(F_i, F) dur(t_{F_i}) \sum_k s_{N,k}^2}} \\
* &= w_c \sum_j S_t(t_{C_j}, t_{F_i}) S_c(C_j, N).
\end{aligned}
$$

Intuitively, the numerator of $C(N, F)$ becomes larger if $F$ is identical to $F_i$ AND $F_i$ coincides with $N$ having the larger score. To prevent "anchor person problem," (An anchor person coincides with almost any name. A face/name coincides with any name/face should correspond to NO name/face.) $C(N, F)$ is normalized with the denominator. $w_c$ is the weight for caption recognition results. Roughly speaking, when a name and a caption match and the caption and a face match at the same time, the face equivalently coincides with $w_c$ occurrences of that name. We use 1 for the value of $w_c$.

### 2.5 Experiments and Discussion

The Name-It system was implemented on an SGI workstation. We processed 10 CNN Headline News videos (30 minutes each), i.e., a total of 5 hours of video. The system extracted 556 face sequences from videos. Name-It performs name candidate retrieval from a given face, and face candidate retrieval from a given name. In face-to-name retrieval, the system is given a face, then outputs name candidates with co-occurrence factors in descending order. Likewise, in name-to-face retrieval, the system outputs face candidates of a given name with co-occurrence factors in descending order.

Figure 4(a) through (d) show the results of face-to-name retrieval. In each result, an image of a given face and ranked name candidates associated with co-occurrence factors are shown. A correct answer is shown with a circled ranking number. Figure 4(e) through (h) show the results of name-to-face retrieval. The top-4 face candidates are shown in order from left to right with corresponding co-occurrence factors. These results demonstrate that Name-It achieves effective face-to-name and name-to-face retrieval with actual news videos.

We have to note that there are some faces not being mentioned in the transcripts, but described in video captions. These faces can be named only by incorporating video caption recognition (e.g., Figure 4(d) and Figure 4(h)). Although these faces are not always the most important in terms of news topics, namely, "the next to the most important," video caption recognition surely enhance performance of Name-It. The overall accuracy that the correct answer is involved in top-5 candidates is 33% in face-to-name retrieval, and 46% in name-to-face retrieval.

## 3 Spot-It

### 3.1 Spot-It Objectives

Data summarization and presentation techniques, in addition to efficient retrieval, are required to navigate the users, since the amount of data stored in the libraries is enormous. 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.

The *topic explainer* view obtained by our method is shown in Figure 5. Each pair of a picture and a sentence is an associated pair for a typical situation. The vertical position of the pair is determined by the situations: 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

| | | |
|---|---|---|
| 1 | MILLER | 0.145916 |
| 2 | VISIONARY | 0.114433 |
| 3 | WISCONSIN | 0.1039 |
| 4 | RESERVATION | 0.103132 |

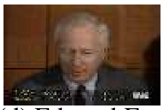
(a) Bill Miller, singer

| | | |
|---|---|---|
| 1 | WARREN | 0.177633 |
| 2 | CHRISTOPHER | 0.032785 |
| 3 | BEGINNING | 0.0232368 |
| 4 | CONGRESS | 0.0220912 |

(b) Warren Christopher, the former U.S. Secretary of State

| | | |
|---|---|---|
| 1 | FITZGERALD | 0.164901 |
| 2 | INDIE | 0.0528382 |
| 3 | CHAMPION | 0.0457184 |
| 4 | KID | 0.0351232 |

(c) Jon Fitzgerald, Actor



| | | |
|---|---|---|
| 1 | EDWARD | 0.0687685 |
| 2 | THEAGE | 0.0550148 |
| 3 | ATHLETES | 0.0522885 |
| 4 | BOWL | 0.0508147 |

(d) Edward Foote, University of Miami President


(e) given "CLINTON" (Bill Clinton)


(f) given "GINGRICH" (Newt Gingrich, 1st and 2nd candidates)


(g) given "NOMO" (Hideo Nomo, pitcher of L.A. Dodgers, 2nd candidate)


(h) given "LEWIS" (Lewis Schiliro, FBI, 2nd candidate)

Figure 4: Face-Name Association Results

politicians and people in that country; the third row shows each speech or opinion about Ireland peace. The horizontal position of the pair is determined by the order of its presented time. Thus, this view also 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. In addition to the above example, the situations such as "speech scene" situation can be good tags for video segments. With these data, video segment retrieval can be much more efficient.

## 3.2 Spotting by Association

### 3.2.1 Key Idea

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.

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. |

Inter-modal coincidence among those *clues* expresses important situations. Examples are shown in Figure 6. 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 Figure 7. 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.

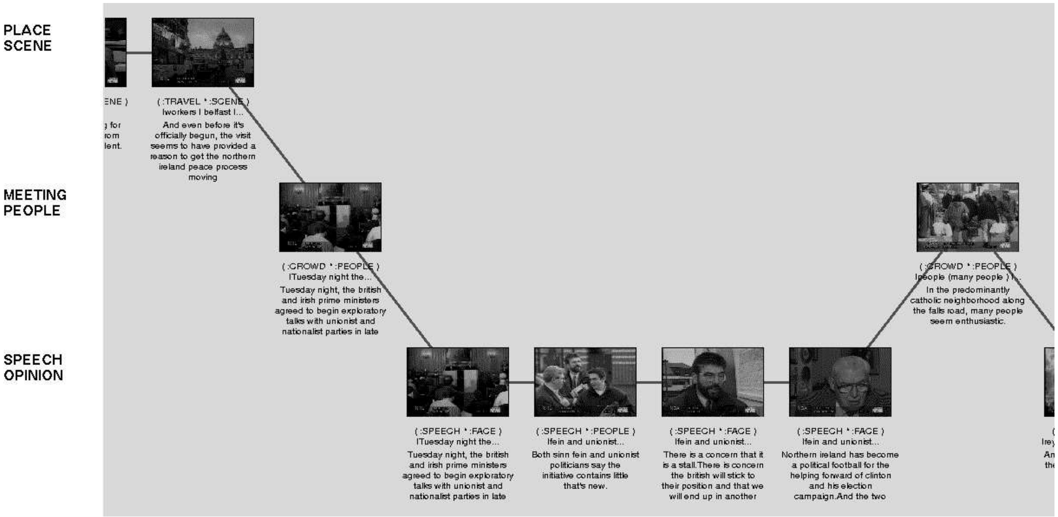


Figure 5: News video TOPIC EXPLAINER (Category + Time Order)

Figure 6: Typical situations

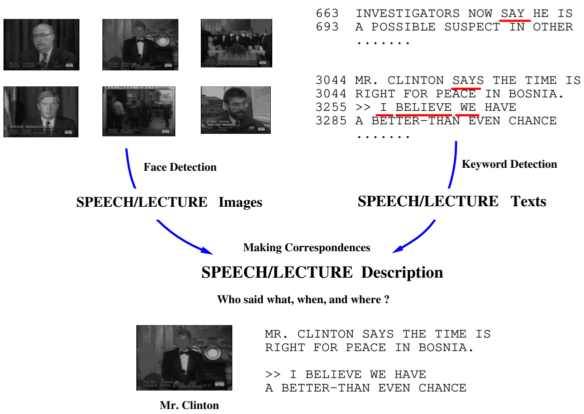


Figure 7: Basic idea of Spotting by Association

### 3.2.2 Language Clues

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.

In a speech or lecture situation, for example, the following words frequently appear, *e.g.* "Mr. President says ..." or "I think it's ...".

**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 2. Each row shows the number of word occurrences in speech portions or other portions[1]. 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%. Similarly, sets of words for other situations were manually determined based on news video transcripts.

However, 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

[1] 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 2: 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% |

subject, the verb, and the objects. Also, each word is checked for expression of a location. The details are skipped because of the lack of space (see [10]).

### 3.2.3 Image Clues

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*.

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 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.

### 3.2.4 Association by DP

The detected data are the sequence of *key-image*s and that of *key-sentence*s to which starting and ending

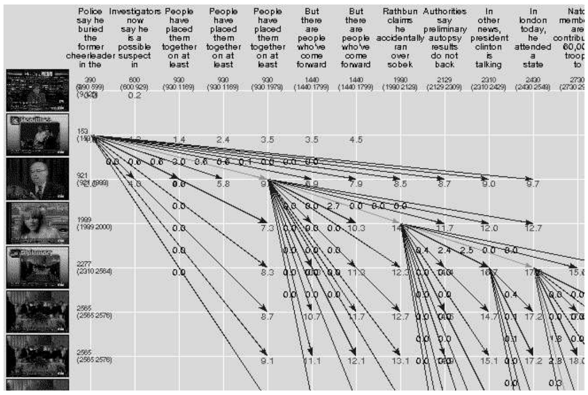


Figure 8: Correspondence between sentences and images

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 for this association is to minimize the following penalty value $P$.

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

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*.

## 3.3 Spot-It Experiments

We chose 6 CNN Headline News videos from the Informedia testbed. Each video is 30 minutes in length. They are segmented into cuts by scene change detection, then each poster frame, *i.e.* representative image for each cut is detected. Next, 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

Table 3: Spotting result (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 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.

parsed by Link Parser. Then, through keyword detection and screening by checking semantics, *key-sentence*s are detected. All transcript processing is done without human assistance, since the *key-sentence* detection results are satisfactory. Finally, inter-modal correspondences between obtained *key-image*s and *key-sentence*s are calculated by DP.

Figure 8 shows an example of 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. In this example, 167 *key-image*s, 122 *key-sentence*s are detected; 69 correspondence cases are successfully obtained. Total numbers of matched and unmatched *key-data* in 6 news videos are shown in Table 3.

Around 70 segments are spotted for each 30-minute news video. This means an average of 3 segments, *i.e.* associated pairs of *key-image*s and *key-sentence*s in a minute. If a topic is not too long, we can place all the segments in one topic into one window. This view is a good presentation and good summarization as already shown in Figure 5.

# 4 Conclusions

We reviewed importance of video information handling in digital libraries, and its difficulties especially in retrieval and presentation. To overcome these problems, access and extraction of semantic contents of video are the key technology. Due to the video property, a multi-modal approach is effective to achieve this task, i.e., integration of image and natural language processing. We introduced two systems, Name-It and Spot-It, taking a multi-modal video analysis approach. Successful results of the systems revealed effectiveness of the approach, as well as importance of semantic contents of videos.

# Acknowledgement

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9411299. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

# References

[1] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: The informedia project," *IEEE Computer*, vol. 29, no. 5, pp. 46–52, 1996.

[2] S. Satoh and T. Kanade, "Name-It: Association of face and name in video," in *Proc. of CVPR*, 1997.

[3] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in video by the integration of image and natural language processing," in *Proc. of IJCAI*, 1997.

[4] H. Rowley, S. Baluja, and T. Kanade, "Human face detection in visual scenes," Tech. Rep. CMU-CS-95-158, School of Computer Science, Carnegie Mellon University, 1995.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[6] P. A. V. Hall and G. R. Dowling, "Approximate string matching," *ACM Computing Surveys*, vol. 12, no. 4, pp. 381–402, 1980.

[7] The Oxford Text Archive. http://ota.ox.ac.uk/.

[8] G. Miller, "WordNet: An on-line lexical database," *Int. J. of Lexicography*, vol. 3, no. 4, 1990.

[9] D. Sleator, "Parsing english with a link grammar," in *Third Int. Workshop on Parsing Technologies*, 1993.

[10] Y. Nakamura and T. Kanade, "Semantic analysis for video contents extraction – spotting by association in news video," in *Proc. of ACM Multimedia*, 1997.