Assembling Personal Speech Collections by Monologue Scene Detection from a News Video Archive

Ichiro IDE*
ide@is.nagoya-u.ac.jp, ide@nii.ac.jp

Naoki SEKIOKA†
nsekioka@murase.m.is.nagoya-u.ac.jp

Graduate School of Information Science, Nagoya University; Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Tomokazu TAKAHASHI
Japan Society for the Promotion of Science / Nagoya University
ttakahashi@murase.m.is.nagoya-u.ac.jp

Hiroshi MURASE
Graduate School of Information Science, Nagoya University
murase@is.nagoya-u.ac.jp

ABSTRACT

Monologue scenes in news shows are important since they contain non-verbal information that could not be expressed through text media. In this paper, we propose a method that detects monologue scenes by individuals in news shows (news subjects) without external or prior knowledge on the show. The method first detects monologue scene candidates by face detection in the frame images, and then excludes scenes overlapped with speech by anchor-persons or reporters (news persons) by dynamically modeling them according to clues obtained from the closed-caption text and from the audio stream. As an application of monologue scene detection, we also propose a method which assembles personal speech collections per individual that appear in the news. Although the methods still need further improvement for realistic use, we confirmed the effectiveness of employing multimodal information for the tasks, and also saw interesting outputs from the automatically assembled speech collections.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Experimentation

Keywords

Face detection, dynamic speech modeling, closed-caption text, personal name annotation

1. INTRODUCTION

Recent advance in data storage technologies has provided us with the ability to archive many hours of video streams accessible as online digital data. Among various genres, we are focusing on television news shows in order to obtain useful knowledge concerning the real-world. Since one of the main focus of news shows is to report social activities in the human society, they are rich in human-related information.

Among the human-related information, monologues by individuals (news subjects) is the most informative when considered as multimedia data since they contain non-verbal information such as expressions, moods, tensions, and even health conditions of the speaker that cannot be observed from text-based news sources such as newspapers. Considering such advantages, we propose an automatic monologue scene detection method from broadcast news video streams exploiting image, audio, and text information in the input stream.

A monologue scene is defined as a "*video segment that contains an event in which a single person, a news subject not a news person, speaks for a long time without interruption by another speaker*", according to the high-level feature extraction task definition in the TRECVid evaluation workshop [16, 15]. The major approach of the works submitted to the workshop tries to detect monologue scenes by removing scenes with news people, that is, anchor-persons or reporters.

For example, in Hauptmann et al.'s work [3], news people and other people are distinguished by looking up names obtained from overlayed captions recognized by Video-OCR in a list of news persons' names collected from broadcasters' web pages. In another work by Amir et al. [1], monologue scenes are detected by high-level feature-based mod-

*Also affiliated to National Institute of Informatics. 2–1–2, Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan.
†Currently at Kyocera Corp.

MIR'06, October 26–27, 2006, Santa Barbara, California, USA.

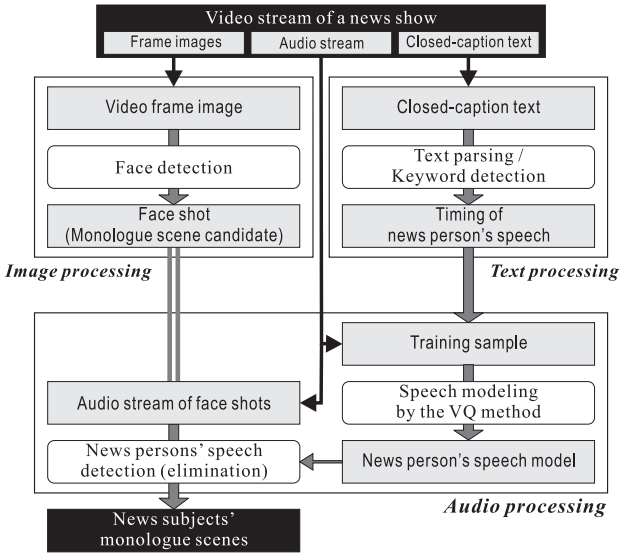

Figure 1: Flow of the proposed monologue scene detection method.

els, which are composed of low-level feature-based models created from image, audio, and text features obtained from manually annotated training data.

These works refer to external information other than the source input video, and also require prior knowledge which makes them difficult to be applied when such information or knowledge are not available, leaving aside the cost for manually annotating training data. In this paper, we propose a monologue detection method that does not require external information or prior training. It detects monologue scenes solely from the input video stream by dynamically creating news persons' speech models. In addition, we also propose a preliminary attempt to assemble personal speech collections by clustering monologues associated with particular individuals.

Monologues have been focused as part of other human-related scenes in some early works. Nakamura and Kanade proposed a method to detect and annotate several human-relates scenes including a 'speech scene' and at the same time showed the effectiveness of focusing on such scenes for summarization [13]. Ide et al. also proposed a selective indexing method focusing on human-related scenes including a 'speech/ report shot' [7].

These works, however, did not consider whether the audio stream accompanying a detected shot was overlapped with a news person's speech or not. We consider that it is important to detect true monologue scenes where a news subject speaks in his/ her own voice. It is, however, difficult to judge whether the voice is actually spoken by the person without prior external knowledge. Thus, in this paper, we propose a method that at least eliminates false monologue scenes with a news person's speech overlapped.

The paper is organized as follows: Section 2 describes the proposed monologue scene detection method with an evaluation experiment. Section 3 describes the method of assembling personal speech collections by annotating monologue scenes with personal names, together with the result of an evaluation of the entire process. Section 4 concludes the paper.

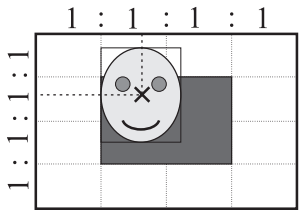

Figure 2: Location of a face region in a 'face shot'.

2. MONOLOGUE SCENE DETECTION

This section describes a method that detects monologue scenes by news subjects; people other than news persons, by eliminating scenes overlapped with a news person's speech. Since news persons' speech models are created dynamically exploiting image, audio, and text information in the input video stream, no external or prior knowledge on them are required for the proposed method.

The process flow of the monologue scene detection method is shown in Figure 1. First, shots including a visually significant face are detected as monologue scene candidates (hereafter, *face shots*) by image processing. Meanwhile, timings of news persons' speech are estimated from certain clues in the closed-caption text by text processing. Speech models are then created from the corresponding audio stream. Next, the speech models are compared against the entire audio stream to detect all scenes with speeches by news persons. Finally, monologues by news subjects are detected by eliminating speech scenes by news persons from all the *face shots*. In this way, we should be able to detect true monologue scenes better; at least most of those with news person's speech overlapped are eliminated.

2.1 Detecting monologue scene candidates

First, monologue scene candidates are extracted by image processing; face detection.

2.1.1 'Face shot' detection

Since the most important feature of a monologue scene is the existence of a face, we start from detecting shots with a visually significant face. After shot segmentation by Chi-square examination between RGB histograms of adjacent frames as a pre-process, faces are detected from the frames. When a monologue scene is shot, a camera-person would usually try to capture the expression of the subject's face in the video frame. Therefore, a face in a monologue scene tends to be relatively large, and is usually in the center of the frame. Accordingly, the following two conditions are applied to frames where a face is detected in order to detect *face shots*; monologue scene candidates.

- **Size**:
  Larger than 8% of the frame size [1].

- **Location**:
  The centroid is located within the blocks in the center of a frame, as illustrated in Figure 2.

[1]The ratio was determined so that faces with approximately 80 pixels square and larger should be detected in the video with a size as specified in Table 1.

Table 1: Specification of the video data used in the experiment.

News show	*NHK News 7* (in Japanese)
Length	890 [minutes]
	(20 to 30 [minutes/day])
Period	Jan. 1, 2004 ∼ Jan. 31, 2004
	(31 [days])
Format	MPEG-1, NTSC
Frame size	352 × 240 [pixels]
Frame rate	30 [frames/second]

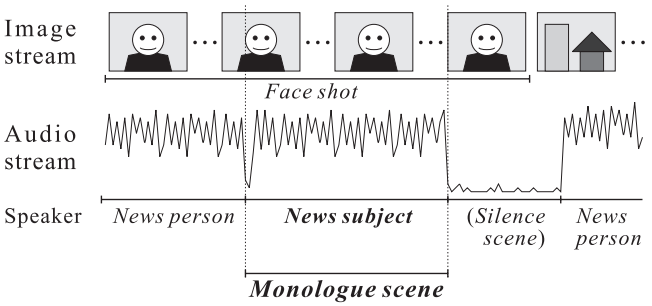

Figure 3: Detection of monologue scenes in a 'face shot'. A monologue scene is generally equal to or part of a 'face shot', which makes it a sub-shot structure. The final output may, however be multiple scenes concatenated across shots.

2.1.2 Experiment on 'face shot' detection

The *face shot* detection process was applied to actual news video streams obtained from a Japanese broadcaster. For face detection, the program in the OpenCV library [8] was used. This program implements a rapid face detection algorithm using Haar-like features [17], which is considered as one of the best performing face detectors currently available. Table 1 shows the specifications of the news video data used in the experiment. As a result, an average recall of 78.5% and a precision of 30.4% were achieved for detecting monologue (candidate) shots by news subjects just from image features; *face shot* detection. Notice that the result is evaluated by whether a detected *face shot* is a news subject's monologue or not, but not the performance of the face detection.

There were a few false negatives caused by the direction of the face that the face detector could not handle, and also by occlusions by hats and so on. On the other hand, most of the false positives were caused by monologue scenes by news persons; the anchor-person reading a news in the studio, reporters covering from news sites and so on. Other false positives were scenes with a news subject in the image, but overlapped with a news person's speech.

2.2 Detecting monologue scenes by news subjects —eliminating news persons' speech

Since the common feature of the causes of false positives is that the audio stream contains speech by a news person, next, the proposed method tries to eliminate these. In this section, we describe a method that dynamically creates models for news persons' speech and in the end eliminate them from the input video. An experiment shows how much of the false positives are eliminated and the precision increases by applying the method.

As shown in Figure 3, there are scenes with news person's speech or with no speech at all (*Silence scenes*) in a *face shot*. In order to eliminate such scenes, the text and audio processing shown in Figure 1 are applied to the *face shots*. As a pre-process, scenes with low sound level are detected and eliminated from the *face shots*.

Next, as the first step, speech models are created for each news person in a video stream. The timing of a speech by a news person is detected by certain clues in the closed-caption text. Once the models for news persons' speech are created, as the second step, they are compared against the audio stream of all the *face shots* to detect and eliminate scenes with a news person's speech. Details of each process follows.

2.2.1 Pre-process: Low audio level detection

When the FFT power spectrum in the audio stream is lower than a given threshold and continues as such for a period of time S_{length}, it is considered as a *silence scene*, and is eliminated from a *face shot*.

2.2.2 Estimating scenes with news persons' speech

In order to create speech models of news persons, samples are collected by estimating the timing [2] of news person's speech according to certain clues in the closed-caption text. This approach enables the proposed method to detect news persons in the audio stream without any external or prior knowledge on them.

News persons consist of an anchor-person who reads a news in the studio and reporters who cover from news sites.

In order to estimate the timing of an anchor-person's speech, we assumed that the first person who speaks in a news show is the anchor-person. Therefore, the first sentence in the closed-caption text and its timing is considered as the beginning of an anchor-person's speech.

On the other hand, a reporter's speech is estimated according to the contents of the preceding speech by the anchor-person. After carefully studying the closed-caption text, the following two conditions were set to detect a reporter's speech. If a sentence satisfies either of the conditions, the following sentences are considered as a reporter's speech.

1. **Addressing a reporter**
   The end of the sentence matches the pattern:
   "[proper noun] + *san* (Mr./ Ms.)."

2. **Real-time conversation with a reporter**
   The sentence is in the present tense and includes any of the three keywords:

   - '*kisha* (reporter)'
   - '*shuzai* (report)'
   - '*chukei* (live report)'

A Japanese morphological analysis system JUMAN [10] and a parsing system KNP [11] were used to analyze the parts of speech and the tense.

[2] Although the appearance of closed-caption text usually lags behind the actual speech in the audio stream, the provided closed-caption text in the archive was already synchronized to the audio stream by word-spotting technologies.

Table 2: Specifications of the audio stream.

Sampling rate	16 [kHz]
Bit rate	16 [bit]
Pre-emphasis	$1 - 0.97z^{-1}$
Frame length	256 [points]
Frame shift length	128 [points]
Window type	Hamming window
Audio feature	18 LPC cepstrum coefficients
Codebook size	128
Distance measure	Euclidian distance

In either of the cases, the samples are extracted from A_{length} succeeding seconds starting from the beginning point estimated from the closed-caption text, excluding the *silence scenes*.

2.2.3 Creating speech models of news persons

Speech models are created from the audio stream of each scene detected in 2.2.2. The model is based on the VQ (Vector Quantization) method generally used for speaker identification.

The VQ method models a speech by composing a codebook per individual that consists of centroids of short-term spectra clusters obtained from sample speech data. In order to identify the person of an input speech, each codebook is applied to quantize the speech, and the distortion of the quantization is measured, where the person corresponding to the least distorted model is identified as the speaker.

As for the short-term spectra feature, LPC (Linear Predictive Coding) cepstrum is used, which is generally considered to represent personal features of speech well. The short-term analysis is composed of the following processes:

1. Pre-emphasis
2. Zero-level normalization
3. Low audio level (*silence scene*) detection
4. Feature extraction (LPC Cepstrum analysis)

This process is applied to both the training and the test data in order to extract short-term speech features.

2.2.4 Eliminating scenes with speeches by news persons

Since the news persons' speech estimation from the closed-caption text does not cover all scenes with news persons' speech, all the speech models are compared with the audio stream of all the *face shots* by VQ distortion, where the matched scenes (speech scenes by a news person) are eliminated from the *face shots* together with the *silence scenes*. As a result, monologue scenes by news subjects remain.

2.3 Experiment on monologue scene detection

The proposed method was evaluated by applying it to the same news video data used in 2.1.2 (Table 1). The parameters were set as shown in Table 2 and also as follows: $S_{length} = 0.5$ [seconds], and $A_{length} = 10$ [seconds].

As a result, an average recall of 76.6% and precision of 55.0% were obtained. Compared to the experiment in 2.1.2, the average precision improved by approximately 25% while the average recall remained almost equivalent. This shows the effect of eliminating speech scenes by news persons. Note that strictly speaking, the results cannot be compared directly, since the experiment in 2.1.2 was evaluated per shot, while this experiment was evaluated per scene which is a sub/super-shot structure independent of the shot structure.

There were very few false positives due to the oversights of anchor-person's speech. It occurred when there was a special news at the beginning of a show covered by a reporter, or an important speech by a news subject. On the other hand, most of the false positives were due to recorded reports where the reporter's speech could not be detected according to the text conditions set in 2.2.2. In such cases, no conversation takes place between the anchor-person and the reporter, thus the reporter's speech often starts suddenly after an anchor-person's speech.

3. ASSEMBLING PERSONAL SPEECH COLLECTIONS

As a usage of the detected monologue scenes, we propose to assemble personal speech collections composed of monologue scenes by individuals that appear in the news. In order to assemble such collections, it is first, necessary to annotate personal names to the detected monologue scenes. Personal name candidates are extracted from the closed-caption text within a story which the monologue scene belongs to.

Since there are usually more than one name candidates, the monologue scenes are next clustered according to the name candidate vectors. Thus, the personal speech collections are assembled. Details of the process follows.

3.1 Annotating personal name candidates to monologue scenes

Personal name candidates are annotated to the monologue scenes by names obtained from within news stories that include them.

3.1.1 Annotation of personal name candidates

As a related work, the Name-It system by Satoh et al. is a pioneering work in face–name association [14]. Referring to their work, we extract and count personal name candidates in the closed-caption text that satisfy the following two conditions:

1. A personal name that appears in the news story that the monologue scene belongs to.
2. A personal name that appears just one sentence before or after the monologue scene.

When a personal name takes a nominative case in a sentence, it is treated as a relatively reliable candidate by counting it as w_c (> 1) counts. In the following experiment, w_c was empirically set to 3.5.

Personal name detection. A personal name is detected according to the dictionary and the method proposed by Ide et al. [4]. This method is based on the nature that in Japanese language, the suffix generally determines the semantic attribute of a noun compound. A brief description of the method is as follows:

1. Each sentence of a closed-caption text is analyzed by a Japanese morphological analysis system JUMAN [10].

2. Noun compounds are extracted according to the morphemes, followed by semantic attribute analysis based on a suffix dictionary.

The suffix dictionary is a semi-automatically collected list of suffices that represent personal attributes.

News story segmentation. As for the news stories, they are segmented by another method proposed by Ide et al. [5]. A brief description of the method is as follows:

1. Create keyword vectors for each sentence. Keyword vectors for four semantic attributes; general, personal, locational/ organizational, and temporal, are formed by noun compounds. The latter two are analyzed in the same way as the personal names by referring to a different suffix dictionary, and all the others are classified as general nouns.

2. For each sentence boundary, concatenate w adjacent vectors on both sides of the boundary. Measure the similarity of the two concatenated vectors by calculating the cosine of the angle between them. Choose the maximum similarity among all the window sizes: w. The maximum of w was set to 10.

3. Combine the similarities in each semantic attribute and detect a topic boundary when it does not exceed a threshold. According to a training with manually given topic boundaries, an optimal weight of 0.23 for general, 0.21 for personal, 0.48 for locational/ organizational, and 0.08 for temporal nouns, and a threshold of 0.17 were obtained.

4. Concatenate over-segmented stories by measuring the similarity of the keyword vectors between adjacent stories.

3.1.2 *Experiment on the annotation of personal name candidates*

The results obtained from the experiment in 2.3 were annotated with personal name candidates. For the evaluation, a monologue scene is considered as successfully annotated when there are personal names that match the manually given ground-truth within the candidates with top-three high counts. As a result, 16.6% of the monologue scenes were correctly annotated.

This result is far from satisfactory, but since the proposed method relies on names in the closed-caption text, it is impossible to obtain correct name candidates if a person is not mentioned in the closed-caption text. If such cases are excluded, 47.2% of the monologue scenes were correctly annotated. Furthermore, if the false positives (mis-detected reporter shots and so on) from the monologue scene detection may also be discarded, the rate increases to 61.9%, which shows the individual ability of the annotation method itself. Incorrect annotations were caused mostly in the following cases:

- The story was discussing mostly about someone else, usually a very important politician than the person actually in the monologue scene.

- Several monologue scenes appeared in a sequence.

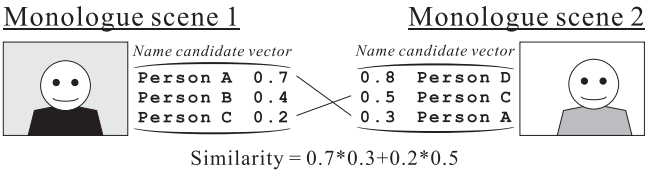

Figure 4: Example of similarity evaluation between personal name candidates of monologue scenes.

3.2 Clustering monologue scenes per individual

As the final step, the monologue scenes annotated by personal name candidates are assembled as speech collections. Feature vectors composed of the top-three name candidates and their counts are clustered by the nearest-neighbor method. To update the centroid of each cluster, the name candidate vector of a newly input monologue scene is compared with the vector of the previous centroid. The similarity of the vectors are evaluated by the cosine measure, as exemplified in Figure 4.

3.3 Experiment on personal speech collection

Applying the process described in the paper from the beginning, personal speech collections were automatically assembled. As for the data set, video data obtained from news shows one month in addition to the data shown in Table 1 were used (Jan. 1, 2004 to Feb. 29, 2004; 60 [days] or 1,700 [minutes]). No manual corrections were made during the processes —from face detection to monologue scene clustering— for this experiment. As a result, an average recall of 37% and precision of 52% were obtained as the classification accuracy. For example, the results of the top-three large monologue collections are shown in Table 3, and an example of a collection is shown in Figure 5.

False positives were caused mostly by oversights of reporters' speeches that appeared as recorded reports together with the mis-annotation of personal names to the monologue scenes. The most common mis-annotations were annotated as 'Prime Minister Koizumi' of Japan at the time of the broadcast. Such mis-annotations tend to be caused by popular people that appear as news subjects frequently, regardless to the main focus of the story. Since these problems are difficult to be solved solely by the proposed method, we should refer to audio-visual features in the clustering process in the future.

Although there are many non-monologue scenes with or without the annotated person, we found it quite interesting to watch monologues after monologues of a person.

4. CONCLUSION

In this paper, we proposed a monologue scene detection method that does not require external information or prior training, together with a report on an attempt to assemble personal speech collections from the detected monologue scenes.

The proposed monologue scene detection method made use of the existence of a visually significant face region in the video frame, and then eliminated scenes overlapped with a news person's speech by dynamically modeling the news persons' speech from the audio stream with clues obtained from the closed-caption text. We experimented the effect

Table 3: Results of the top-three large speech collections.

Collection name	Prime Minister Koizumi	President Bush	Chief Cabinet Secretary Fukuda	Average
Manually extracted ground-truth monologue scenes	49	9	11	
Correctly detected as monologue scenes	38	13	11	
and also correctly annotated and clustered	**10**	**5**	**4**	
but no names appear in the text or incorrectly annotated	*18*	*8*	*7*	
and correctly annotated but incorrectly clustered	*10*	*0*	*0*	
Recall	20%	56%	36%	37%
Precision	36%	62%	57%	52%

Jan 15, 2004 [05'38"-05'58"] **Jan 15, 2004** [06'56"-07'20"] **Jan 21, 2004** [13'09"-14'07"]

Jan 15, 2004 [14'08"-14'31"] **Feb 07, 2004** [07'37"-08'02"]

(a) Correctly clustered monologue scenes

Jan 20, 2004 [13'08"-13'15"] **Feb 07, 2004** [07'07"-07'20"]

(1) *Monologue scenes by news persons*

Jan 16, 2004 [20'05"-20'20"] **Jan 20, 2004** [13'28"-13'30"] **Jan 21, 2004** [15'53"-16'46"]

Jan 21, 2004 [16'58"-17'05"] **Jan 21, 2004** [17'06"-17'41"] **Feb 18, 2004** [16'25"-16'30"]

(2) *Monologue scenes overlapped with news person's speech*

(b) Incorrectly clustered monologue scenes

Figure 5: Example of an automatically assembled personal speech collection: President Bush.

of this multimodal approach, which showed approximately 25% improvement in the average precision while maintaining the average recall.

One potential drawback of the method depending on the application is that it cannot be run in real-time since it needs to scan through a stream twice, besides the offline closed-caption and audio stream synchronization process. Other drawbacks are that it cannot handle news shows with two anchor-persons, and that it does not ensure that the speech is actually spoken by the person in the image. These issues should be considered in the future, and be solved by incremental accumulation of speech collections together with the usage of visual features such as synchronism of lip movements to the speech.

Speech collections assembled from the detected monologue scenes were not necessarily satisfactory at this point, but in the future, we will try to obtain better clustering results by employing audio-visual features of the person in the monologue scene; combination of the proposed method with speaker and face clustering. This should also be effective in distinguishing individuals with same names. Using video-OCR technologies to obtain better name candidates should also improve the results. Once these attempts should succeed, speech collections for various individuals will be created automatically just from a large news video archive.

We will also see the effectiveness of handling monologue scenes somewhat special than other scenes when generating news summaries in another work [6], and also works such as [2] should also merit from the proposed method.

5. ACKNOWLEDGMENTS

Parts of this work were supported by the Grants-in-Aid for Scientific Research (#15700116, #16016289, #18049035, and #18700080) and the 21st century COE program from the Ministry of Education, Culture, Sports, Science and Technology, and the Japan Society for the Promotion of Science, and also by the Research Grant from Kayamori Foundation of Information Science Advancement (#K17ReX-202).

The video data used in the experiments were provided from the National Institute of Informatics Broadcast Video Archive [9], through a joint research project. Parts of the implementation was done using the Speech Signal Processing Toolkit: SPTK [12].

We would like to thank Dr. Chiyomi Miyajima at Nagoya University for her support and professional advice especially in the speech modeling part of the proposed method.

6. REFERENCES

[1] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. R. Naphade, A. P. Nastev, C. Neti, H. Nock, J. R. Smith, B. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID-2003 video retrieval system. In *Online Proc. TRECVID 2003*, November 2003.

[2] S. Bocconi, F. Nack, and L. Hardman. Using theoretical annotations for generating video documentaries s. In *Proc. IEEE 2005 Intl. Conf. on Multimedia and Expo*, July 2005.

[3] A. G. Hauptmann, D. Ng, R. Baron, M. G. Christel, P. Duygulu, C. Huang, W.-H. Lin, H. D. Wactlar, N. Moraveji, C. G. Snoek, G. Tzanetakis, J. Yang, and R. Jin. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Online Proc. TRECVID 2003*, November 2003.

[4] I. Ide, R. Hamada, S. Sakai, and H. Tanaka. Semantic analysis of television news captions referring to suffixes. In *Proc. Fourth Intl. Workshop on Information Retrieval with Asian Languages*, pages 37–42, November 1999.

[5] I. Ide, H. Mo, N. Katayama, and S. Satoh. *Image and Video Retrieval —Third Intl. Conf., CIVR2004, Dublin, Ireland, July 2004, Procs.—*, volume 3115 of *Lecture Notes in Computer Science*, chapter Topic threading for structuring a large-scale news video archive, pages 123–131. Springer-Verlag, July 2004.

[6] I. Ide, H. Mo, N. Katayama, and S. Satoh. Exploiting topic thread structures in a news video archive for the semi-automatic generation of video summaries. In *Proc. 2006 IEEE Intl. Conf. on Multimedia and Expo*, pages 1473–1476, July 2006.

[7] I. Ide, K. Yamamoto, and H. Tanaka. *Advanced Multimedia Content Processing —First Intl. Conf. AMCP'98, Osaka, Japan—*, volume 1554 of *Lecture Notes in Computer Science*, chapter Automatic video indexing based on shot classification, pages 87–102. Springer-Verlag, January 1999.

[8] Intel Corp. Open source computer vision library. http://www.intel.com/technology/computing/opencv/.

[9] N. Katayama, H. Mo, I. Ide, and S. Satoh. *Advances in Multimedia Information Processing —PCM2004 Fifth Pacific Rim Conf. on Multimedia, Tokyo, Japan, November/December 2004, Procs. Part II—*, volume 3332 of *Lecture Notes in Computer Science*, chapter Mining large-scale broadcast video archives towards inter-video structuring, pages 489–496. Springer-Verlag, December 2004.

[10] Kyoto University, Kurohashi Lab. Japanese morphological analysis system, JUMAN. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html.

[11] Kyoto University, Kurohashi Lab. Japanese parsing system, KNP ver. 2.0. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html.

[12] Nagoya Institute of Technology, Tokuda Lab. Speech signal processing toolkit: SPTK. http://kt-lab.ics.nitech.ac.jp/˜tokuda/SPTK/.

[13] Y. Nakamura and T. Kanade. Semantic analysis for video contents extraction —spotting by association in news video. In *Proc. Fifth ACM Intl. Conf. on Multimedia*, pages 393–401, November 1997.

[14] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, January–March 1999.

[15] A. F. Smeaton. *Image and Video Retrieval —Fourth Intl. Conf., CIVR2005, Singapore, July 2005, Procs.—*, volume 3568 of *Lecture Notes in Computer Science*, chapter Large scale evaluations of multimedia information retrieval: The TRECVid experience, pages 11–17. Springer-Verlag, July 2005.

[16] United States, National Institute of Standards and Technology. TRECVid evaluation. http://www-nlpir.nist.gov/projects/trecvid/.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, December 2001.