Reducing the Cost of Metadata Generation by Using Video/ Audio Indexing and Natural Language Processing Techniques

Hidetaka Kuwano[†], *Yoshihiro Matsuo, and Katsuhiko Kawazoe*

Abstract

Reducing the cost of generating metadata will allow more broadcast contents to be transmitted with advanced viewing options. In this article, we describe SceneCabinet, a system that automatically extracts scene-based semantic metadata from video content. It extracts meaningful video slices and their associated textual information such as the title, synopsis, and keywords by using natural language processing based on the results of speech and on-screen text recognition. Moreover, it can import video program scripts and use them for automatic keyword extraction. SceneCabinet provides an intuitive user operation interface including a video browser with key images detected automatically based on scene changes, on-screen text, camerawork, speech, and music information. Experiments showed that SceneCabinet can significantly reduce metadata generation costs.

1. Introduction

In recent years, technology for distributing broadcast content and metadata over broadband IP (Internet protocol) networks has been studied with the objective of achieving a new style of TV viewing that will create a new market for broadcasting and telecommunication services. Metadata is a key to implementing scene-based advanced TV viewing services such as scene navigation, digest viewing, and highlight viewing.

To achieve scene-based TV viewing services, information such as what scenes are contained in the program and the times at which the scenes appear must be described as metadata before the program is broadcast. For example, for a news program, the metadata should describe the content of each individual news topic; for a sports program, it should describe information such as the approximate times when exciting events like goal shots or home runs occur. On the other hand, the task of generating such scene-based semantic metadata requires a great number of operations when it is done entirely manually. Thus, reducing the cost of generating metadata seems to be the key to distributing more contents that allow advanced viewing services.

After reviewing segmentation metadata and conventional methods of metadata generation, we describe SceneCabinet, a system that automatically extracts semantic metadata which previously could only be created manually and at high cost. It automatically extracts meaningful video slices and their titles, synopses, and keywords using natural language processing based on the results of speech and on-screen text recognition and program scripts. SceneCabinet also provides an intuitive user operation interface. We also present experimental results that show the reduction in metadata generation cost.

2. Segmentation metadata

An international standard format for video content metadata for advanced TV viewing has been established by the TV-Anytime Forum [1]. With TV-Anytime, each individual slice in a video content is defined as a “segment”, and information describing the segment is referred to as “segmentation metadata”. This segmentation metadata includes the starting

† NTT Cyber Solutions Laboratories
Yokosuka-shi, 239-0947 Japan
E-mail: kuwano.hidetaka@lab.ntt.co.jp

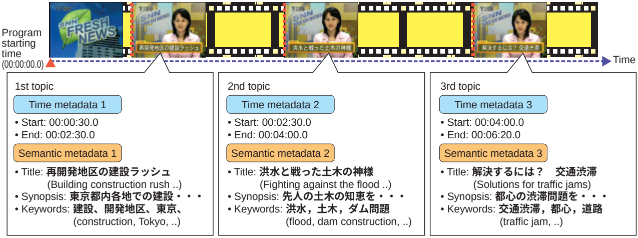

Program starting time (00:00:00.0)
Time
SNN FRESH NEWS
再開発地区の建設ラッシュ
洪水と戦った土木の神様
解決するには？ 交通渋滞
1st topic
Time metadata 1
• Start: 00:00:30.0
• End: 00:02:30.0
Semantic metadata 1
• Title: 再開発地区の建設ラッシュ (Building construction rush ..)
• Synopsis: 東京都内各地での建設・・・
• Keywords: 建設、開発地区、東京、 (construction, Tokyo, ..)
2nd topic
Time metadata 2
• Start: 00:02:30.0
• End: 00:04:00.0
Semantic metadata 2
• Title: 洪水と戦った土木の神様 (Fighting against the flood ..)
• Synopsis: 先人の土木の知恵を・・・
• Keywords: 洪水，土木，ダム問題 (flood, dam construction, ..)
3rd topic
Time metadata 3
• Start: 00:04:00.0
• End: 00:06:20.0
Semantic metadata 3
• Title: 解決するには？　交通渋滞 (Solutions for traffic jams)
• Synopsis: 都心の渋滞問題を・・・
• Keywords: 交通渋滞，都心，道路 (traffic jam, ..)

Fig. 1. Example of segmentation metadata for a Japanese news program.

and ending times of the segment in the program and textual information that includes the segment title, synopsis, and keywords. For convenience in the following discussion, we refer to these as “time metadata” and “semantic metadata”, respectively.

Specific examples of time and semantic metadata for a Japanese news program are shown in **Fig. 1**. In ordinary broadcasting, a news program presents a number of news topics in order. In advanced TV viewing services, there will be useful menus for searching for interesting news topics and selecting topics from a list by genre (politics, the economy, or sports). To implement a TV viewing service for news programs in which news topics can be viewed individually, information about the starting and ending times of each topic within the entire news program are specified as time metadata and the topic title, synopsis, and keywords needed for topic searches are specified as semantic metadata.

3. Conventional methods of metadata generation

When time metadata is generated manually, the video content must be viewed from beginning to end to check the whole contents by using a video replay controller for a VCR (video cassette recorder). The time spent playing and fast-forwarding to find the scenes to be checked accounts for a large portion of the total task time. Furthermore, the concentration required so as not to overlook a scene while fast-forwarding makes this task very tiring for the person doing it. For semantic metadata, it is necessary to analyze the program contents, assign an appropriate title, synopsis, and set of keywords, and write down the information by hand while viewing the playback of the video content. When this is done manually, the segmentation metadata generation task may take as much as ten times the program duration, although this depends on the program genre and service content. Therefore, there is a strong demand for technology for reducing the cost of generating segmentation metadata.

In recent years, many systems [2] and [3] for automatically generating segmentation metadata that employ a video and audio indexing function, such scene change detection and speech recognition, have been proposed. Nevertheless, segmentation metadata created by these systems alone cannot always represent the above-mentioned time and semantic metadata in a way that facilitates advanced TV viewing. It turns out that segmentation metadata generation using the conventional systems is very costly because most of the results of automatic metadata generation obtained by those systems are inadequate for advanced TV viewing and need to be revised. Reducing the cost of generating segmentation metadata is thus a major requirement for distributing greater amounts of video contents for advanced TV viewing.

4. Proposed workflow

An effective way to reduce the cost is to automate time-consuming manual tasks. **Figure 2** shows the overall proposed workflow on the assumption that digitized video content exists. Once the provisional time and semantic metadata has been obtained by combining the automatic video, audio indexing, and natural language processing techniques of [4]-[8] in step 1, the operator can simply check that data and revise it as necessary in step 2. Thus, it is possible to

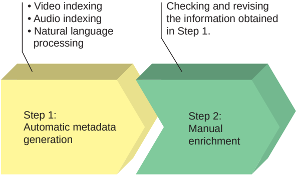

• Video indexing
• Audio indexing
• Natural language processing
Checking and revising the information obtained in Step 1.
Step 1:
Automatic metadata generation
Step 2:
Manual enrichment

Fig. 2. Workflow.

reduce the overall task time for segmentation metadata generation.

4.1 Time metadata generation task

Time metadata could be generated in a short time if it were possible to browse the overall flow of the video content and immediately start the playback from any scene in the program. In [4], [5], and [6], technology is proposed for

(a) automatically detecting time periods of each shot, on-screen text, camerawork, speech, and music, which are useful for scene-based checking of video content and

(b) a graphical user interface (GUI) that displays the detected scene information as an easy-to-view list of key images, allowing immediate playback of the scene beginning at that image when it is selected from the list.

News topics usually begin with a shot containing a person (news anchor) and on-screen text showing the topic (title) so process (a) can be used for checking the approximate start and end times of individual news topics. The GUI allows the operator to immediately jump to the desired scene and generate time metadata through fine adjustments of the time based on the detected scenes.

4.2 Semantic metadata generation task

If textual information such as the title, synopsis, and keywords of a segment can be obtained automatically without checking the video content, then the task's duration can be shortened. In [6], [7], and [8], technology is proposed for

(c) automatically recognizing on-screen text and speech and converting this information into text form and

(d) automatically extracting words and sentences that correspond to video titles, synopses, and keywords from the textual data of the recognition results by natural language processing with a language model.

For TV news programs, the accuracy of on-screen text recognition is over 80% and that of speech recognition is over 90% (in Japanese), especially for the anchor's speech [6], [7]. For example, headline text superimposed at the beginning of each news topic is suitable for use as topic titles, so the recognition results for the on-screen text can be used for that purpose. As the synopsis, it is possible to use the recognition results for the anchor's speech in the news studio, and for keywords, suitable words can be selected from all of the on-screen text and speech recognition results for each topic by natural language processing. The accuracy of the extracted semantic metadata for news programs is about 80% [8]. This enables the operator to simply check and revise the extraction results for generating the final semantic metadata. **Figure 3** shows an outline of semantic metadata generation by natural language processing.

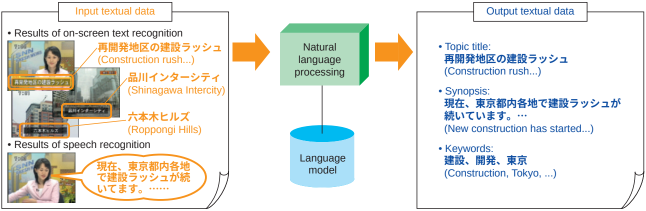


Fig. 3. Outline of semantic metadata extraction process.

5. SceneCabinet system

We developed SceneCabinet on the basis of the technology described in [4]-[8]. The functional configuration of the system is illustrated in **Fig. 4**. The main components of the system are the metadata generation engine and the metadata authoring GUI.

5.1 Metadata generation engine

The metadata generation engine detects scene changes, camerawork, and music in the video/audio content and recognizes the on-screen text and the speech. It outputs the time information and image data for each detected scene and the various recognition results as text data. In addition, natural language processing is used to automatically extract the title, synopsis, and keywords from the recognition results for the on-screen text and speech. There is also an error correction function for recognizing on-screen text using a language model.

5.2 Metadata authoring GUI

The metadata authoring GUI module provides the user with a GUI function for checking and revising the output of the metadata generation engine to produce the final segmentation metadata. The GUI design for the checking and revising of automatically generated information is also an important factor in reducing the cost of the metadata generation task. **Figure 5** shows an example screen shot of the metadata authoring GUI. The main components are discussed below.

(1) Video browser and playback monitor

The video browser displays a list of key images showing the results output by the metadata generation engine. The user can easily browse the content of the entire video. Each image is color-coded to indicate the scene type (scene change, on-screen text, etc.), so the operator can easily select among only the type of images he or she wants to check. When an image is selected in the video browser, the video is immediately played on the playback monitor, beginning at the selected image. In addition, it is possible to play the scene frame by frame or jump at various time intervals (e.g., 10 s). Using the video browser and playback monitor, the operator can check and revise time metadata easily. Another feature is that the playback monitor has a sound level waveform display function, which makes it possible to check for the presence or absence of audio in a fixed interval without playing the video.

(2) Metadata editor

The metadata editor has a button that lets the user automatically extract semantic metadata based on on-screen text and speech recognition results for the video segment being edited. The user can check, edit, and revise the results of automatic semantic metadata extraction.

(3) Scenario import

The metadata authoring GUI is also equipped with a function for importing existing textual data such as the program scenario and the anchor’s manuscript and automatically extracting semantic metadata from them. Although it depends on the program genre, the program scenarios and manuscripts usually include

(a) the estimated time schedule for the beginning and end of each topic in the program and

(b) the content of the narration and anchor’s speech for each topic.

Therefore, (a) can be used as key information for generating time metadata and (b) can be used for generating semantic metadata. Script data, if available,

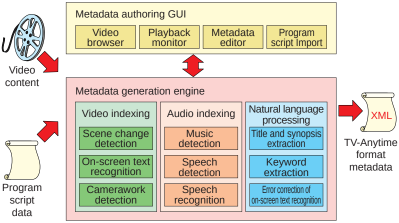
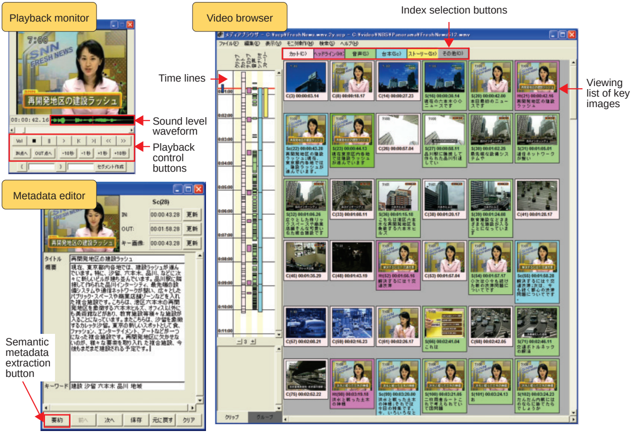


Fig. 4. Overview of SceneCabinet system.



Sc(28)

IN: 00:00:43.28 更新

OUT: 00:01:58.28 更新

キー画像: 00:00:43.28 更新

タイトル 再開発地区の建設ラッシュ

概要 現在、東京都内各地では、建設ラッシュが進んでいます。特に、汐留、六本木、品川、などに次々に新しいビルが建ち並んでいます。品川駅に隣接して作られた品川インターシティ、最先端の設備システムや通信ネットワークが整い、広々としたパブリック・スペースや商業店舗ゾーンなどを入れた複合施設です。こちらは、港区六本木の再開発地区を象徴する六本木ヒルズ、オフィス以外にも美術館などがあり、教育施設等様々な施設が入ることになっています。またこちらは、汐留を象徴するカレッタ汐留。東京の新しいスポットとして食、ファッション、エンターテイメント、アートなどが一つになった複合施設です。再開発地区に欠かせないのが、様々な要素を取り入れた複合施設、今後もまだまだ建設される予定です。

キーワード 建設 汐留 六本木 品川 地域

Fig. 5. Example screen shot of metadata authoring GUI.

can be automatically combined with the video and audio recognition results to obtain semantic metadata that is more pertinent, thus shortening the task time even more. The segmentation metadata created with this system is finally output as an XML (extensible markup language) file in the TV-Anytime format.

6. Experimental results and discussion

To investigate the task of generating time and semantic metadata for a real broadcast news program (as shown in Fig. 2), we conducted experiments to compare the task times when using SceneCabinet and when performing the task entirely by hand.

6.1 Task time measurement

The task was performed by four operators for each task model. The task times were measured and averaged. Furthermore, multiple news program videos were used for each task model, so data was collected for four times the number of programs for each task model, and average values were obtained.

6.2 Target task

We wanted to determine the effect of our system on user operation cost, so the task in step 1 in Fig. 2, “automatic metadata generation”, which can be executed by running batch programs at night, was out of the scope of our measurements. The task of step 2, “manual enrichment”, was the main focus of these experiments.

6.3 Detailed rules for generating segmentation metadata

A news program generally presents a number of news topics in order. In each individual news topic, the anchor first presents a summary of the topic and then video contents that explain the topic in detail are broadcast. Then, the next news topic follows. Headline text showing the topic title is superimposed at the beginning of each topic. Based on these characteristics, we established the following rules for the operator’s task of generating time and semantic metadata for each individual news topic.

1) Time metadata: “Start time” is defined to be within the time period with no voice just before

the anchor's first utterance about the topic. "End time" is defined to be within the time period with no voice just after the content of the topic has finished.

2) Semantic metadata: "News title" is defined as when headline text is superimposed at the beginning of the topic. "Synopsis" is defined from the anchor's speech. "Keywords" are defined by selecting one suitable genre and five suitable words from all of the on-screen text and speech contents.

6.4 Results

The bar graph in **Fig. 6** shows the total task times in step 2 in Fig. 2 for both time and semantic metadata generation, normalized with respect to video length. When done entirely manually, the task required about 3.5 times the length of the video; when SceneCabinet was used, it took about 1.6 times the length of the video. These results show that using the SceneCabinet could reduce the task time to about 46% compared with the time required to do it manually. A breakdown of the task times shows that the time could be reduced to about 45% for both the segment metadata and the semantic metadata.

(1) Time metadata generation task

According to the results of questionnaires given to the operators, using the images in which the anchor appears and displaying the on-screen text as a list on the video browser as cues made it particularly easy to find the starting and ending times of the news topics. Furthermore, the various playback control buttons and the sound level waveform viewing function of the playback monitor also effectively reduced the time spent playing and fast-forwarding during the task.

(2) Semantic metadata generation task

The video summary function of the metadata editor could extract a pertinent title, synopsis, and set of keywords for each news topic. Thus the task could be completed by simply revising parts of the output without checking the whole content of the news topic. This reduced both the time required and the amount of user fatigue compared with doing the work entirely manually.

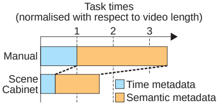


Fig. 6. Results of task time measurement.

6.5 Discussion

For news programs, the ability to automatically obtain information that is close to the final desired metadata by video and audio recognition and natural language processing contributed greatly to operation cost reduction. However, the metadata generation cost depends on a number of things, including the task rules, the program genre, and the skill of the operator. For example, the task rules for these experiments specified that five keywords were to be generated. If that rule were relaxed to 'one or more', we expect that the task time for SceneCabinet's model would be shortened markedly because the operators could just select a suitable word in the results of automatic semantic metadata extraction. That is to say, the efficiency of the operation cost reduction varies according to the assumed TV viewing service and task rules. We believe that these results can be used as basic data for future work assuming actual service conditions.

Concerning operator skill, because comparatively explicit rules for guidance in performing the task content can be specified for news programs, the operators were able to perform the task smoothly and without confusion, even when they did not have specialized skills. On the other hand, in tasks such as selecting important plays in sports programs such as digest scenes in soccer matches, the operator must have specialized knowledge such as expertise in the sport or creative skills such as generating metadata that includes presentation effects in the digest. Even in such cases, however, if the task can be divided into two stages: the first involves generating "basic" segmentation metadata that includes every scene expected to be used in the digest and the second involves creative personnel using those basic metadata to create the final metadata. Thus, work rules can easily be defined for the former task, and video recognition technology can also be applied effectively. Thus, we can also expect that a system such as SceneCabinet will lighten the burden on creative personnel and reduce the overall cost of the task.

From the above discussion, the important points in reducing the cost of generating segmentation metadata are:

- Being able to obtain information that is close to the final desired information by automatic metadata generation technology such as video recognition
- Separating out the parts of the overall task for

which clear rules for performing the task can be established.

7. Conclusion

Our segmentation metadata extraction system "SceneCabinet" is effective at reducing the time required, and hence the cost, for generating metadata for news programs. In the future, building on what we have achieved in this work, we will continue research and development on automatic metadata generation engines and user interface technology to reduce the cost of metadata generation tasks that need more creative skill than those involved in news programs.

References

[1] http://www.tv-anytime.org

[2] http://www.virage.com

[3] http://www.mediasite.com

[4] Y. Taniguchi, A. Akutsu, and Y. Tonomura, "PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing," Proceedings of ACM Multimedia 97, pp. 427-436, 1997.

[5] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, "Video Handling with Music and Speech Detection," Proceedings of IEEE Multimedia, Vol. 5, No. 5, pp. 17-25, 1998.

[6] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima, "Telop on Demand: Video Structuring and Retrieval based on Text Recognition," Proceedings of International Conference on Multimedia and Expo 2000, pp. 759-762, 2000.

[7] K. Ohtsuki, T. Matsuoka, S. Matsunaga, and S. Furui, "Topic Extraction based on Continuous Speech Recognition in Broadcast-news Speech," Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 527-534, 1997.

[8] Y. Hayashi, K. Ohtsuki, K. Bessho, O. Mizuno, Y. Matsuo, S. Matsunaga, M. Hayashi, T. Hasegawa, and N. Ikeda, "Speech-based and Video-supported Indexing of Multimedia Broadcast News," Proceedings of SIGIR, pp. 441-442, 2003.

Hidetaka Kuwano

Research Engineer, NTT Cyber Solutions Laboratories.

He received the B.S. and M.S. degrees in information engineering from Niigata University, Niigata in 1993 and 1995, respectively. In 1995, he joined NTT Human Interface Laboratories, Yokosuka. Since then he has been engaged in R&D of video analysis, video structuring, and video OCR systems. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Information Processing Society of Japan and received the Young Engineer Award from IEICE in 2000.

Yoshihiro Matsuo

Senior Research Engineer, NTT Cyber Solutions Laboratories.

He received the B.S. and M.S. degrees from Osaka University, Osaka in 1988 and 1990, respectively. In 1990, he joined NTT Communications and Information Processing Laboratories. His current interests include machine translation and cross-language information access.

Katsuhiko Kawazoe

Senior Research Engineer, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in engineering from Waseda University, Tokyo in 1985 and 1987, respectively. Since joining NTT in 1987, he has mainly been engaged in R&D of radio communication systems, satellite communication systems, and the personal handy-phone system (PHS). His specialty is forward error correction systems. He is currently a co-chairman of the Association of Radio Industries and Businesses Working Group for Broadcasting Systems based on a Home Server. He is a member of IEICE and received the Young Engineer Award from IEICE in 1995.