Annotation-Based Multimedia Summarization and Translation

Katashi Nagao
Dept. of Information Engineering
Nagoya University
and CREST, JST
Furo-cho, Chikusa-ku,
Nagoya 464-8603, Japan
nagao@nuie.nagoya-u.ac.jp

Shigeki Ohira
School of Science and
Engineering
Waseda University
3-4-1 Okubo, Shinjuku-ku,
Tokyo 169-8555, Japan
ohira@shirai.info.waseda.ac.jp

Mitsuhiro Yoneoka
Dept. of Computer Science
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku,
Tokyo 152-8552, Japan
yoneoka@img.cs.titech.ac.jp

Abstract

This paper presents techniques for multimedia annotation and their application to video summarization and translation. Our tool for annotation allows users to easily create annotation including voice transcripts, video scene descriptions, and visual/auditory object descriptions. The module for voice transcription is capable of multilingual spoken language identification and recognition. A video scene description consists of semi-automatically detected keyframes of each scene in a video clip and time codes of scenes. A visual object description is created by tracking and interactive naming of people and objects in video scenes. The text data in the multimedia annotation are syntactically and semantically structured using linguistic annotation. The proposed multimedia summarization works upon a multimodal document that consists of a video, keyframes of scenes, and transcripts of the scenes. The multimedia translation automatically generates several versions of multimedia content in different languages.

1 Introduction

Multimedia content such as digital video is becoming a prevalent information source. Since the volume of such content is growing to huge numbers of hours, summarization is required to effectively browse video segments in a short time without missing significant content. Annotating multimedia content with semantic information such as scene/segment structures and metadata about visual/auditory objects is necessary for advanced multimedia content services. Since natural language text such as a voice transcript is highly manageable, speech and natural language processing techniques have an essential role in our multimedia annotation.

We have developed techniques for semi-automatic video annotation integrating a multilingual voice transcription method, some video analysis methods, and an interactive visual/auditory annotation method. The video analysis methods include automatic color change detection, characterization of frames, and scene recognition using similarity between frame attributes.

There are related approaches to video annotation. For example, MPEG-7 is an effort within the Moving Picture Experts Group (MPEG) of ISO/IEC that is dealing with multimedia content description (MPEG, 2002). MPEG-7 can describe indeces, notes, and so on, to retrieve necessary parts of content speedily. However, it takes a high cost to add these descriptions by hands. The method of extracting them automatically through the video/audio analysis is vitally important. Our method can be integrated into tools for authoring MPEG-7 data. The linguistic description scheme, which will be a part of the amendment to MPEG-7, should play a major role in this integration.

Using such annotation data, we have also developed a system for advanced multimedia processing such as video summarization and translation. Our video summary is not just a shorter version of the original video clip, but an interactive multimedia presentation that shows keyframes of important scenes and their transcripts in Web pages and allow users to interactively modify summary. The video summarization is customizable according to users' favorite size and keywords. When a user's client device is not capable of video playing, our system tranforms video to a document that is the same as a Web document in HTML format.

The multimedia annotation can make delivery of multimedia content to different devices very effective. Dissemination of multimedia

content will be facilitated by annotation on the usage of the content in different purposes, client devices, and so forth. Also, it provides object-level description of multimedia content which allows a higher granularity of retrieval and presentation in which individual regions, segments, objects and events in image, audio and video data can be differentially accessed depending on publisher and user preferences, network bandwidth and client capabilities.

2 Multimedia Annotation

Multimedia annotation is an extension of document annotation such as GDA (Global Document Annotation) (Hasida, 2002). Since natural language text is more tractable and meaningful than binary data of visual (image and moving picture) and auditory (sound and voice) content, we associate text with multimedia content in several ways. Since most video clips contain spoken narrations, our system converts them into text and integrates them into video annotation data. The text in the multimedia annotation is linguistically annotated based on GDA.

2.1 Multimedia Annotation Editor

We developed an authoring tool called multimedia annotation editor capable of video scene change detection, multilingual voice transcription, syntactic and semantic analysis of transcripts, and correlation of visual/auditory segments and text.

Figure 1: Multimedia Annotation Editor

An example screen of the editor is shown in Figure 1. The editor screen consists of three windows. One window (top) shows a video content, automatically detected keyframes in the video, and an automatically generated voice transcript. The second window (left bottom) enables the user to edit the transcript and modify an automatically analyzed linguistic markup structure. The third window (right bottom) shows graphically a linguistic structure of the selected sentence in the second window.

The editor is capable of basic natural language processing and interactive disambiguation. The user can modify the results of the automatically analyzed multimedia and linguistic (syntactic and semantic) structures.

2.2 Linguistic Annotation

Linguistic annotation has been used to make digital documents machine-understandable, and to develop content-based presentation, retrieval, question-answering, summarization, and translation systems with much higher quality than is currently available. We have employed the GDA tagset as a basic framework to describe linguistic and semantic features of documents. The GDA tagset is based on XML (Extensible Markup Language) (W3C, 2002), and designed to be as compatible as possible with TEI (TEI, 2002), CES (CES, 2002), and EAGLES (EAGLES, 2002).

An example of a GDA-tagged sentence follows:

```
<su><np opr="agt" sem="time0">Time</np>
<v sem="fly1">flies</v>
<adp opr="eg"><ad sem="like0">like</ad>
<np>an <n sem="arrow0">arrow</n></np>
</adp>.</su>
```

The `<su>` element is a sentential unit. The other tags above, `<n>`, `<np>`, `<v>`, `<ad>` and `<adp>` mean noun, noun phrase, verb, adnoun or adverb (including preposition and postposition), and adnominal or adverbial phrase, respectively.

The `opr` attribute encodes a relationship in which the current element stands with respect to the element that it semantically depends on. Its value denotes a binary relation, which may be a thematic role such as agent, patient, recipient, etc., or a rhetorical relation such as cause, concession, etc. For instance, in the above sentence,

<np opr="agt" sem="time0">Time</np> depends on the second element <v sem="fly1">flies</v>. opr="agt" means that *Time* has the agent role with respect to the event denoted by *flies*. The sem attribute encodes a word sense.

Linguistic annotation is generated by automatic morphological analysis, interactive sentence parsing, and word sense disambiguation by selecting the most appropriate item in the domain ontology. Some research issues on linguistic annotation are related to how the annotation cost can be reduced within some feasible levels. We have been developing some machine-guided annotation interfaces to simplify the annotation work. Machine learning mechanisms also contribute to reducing the cost because they can gradually increase the accuracy of automatic annotation.

In principle, the tag set does not depend on language, but as a first step we implemented a semi-automatic tagging system for English and Japanese.

2.3 Video Annotation

The linguistic annotation technique has an important role in multimedia annotation. Our video annotation consists of creation of text data related to video content, linguistic annotation of the text data, automatic segmentation of video, semi-automatic linking of video segments with corresponding text data, and interactive naming of people and objects in video scenes.

To be more precise, video annotation is performed through the following three steps.

First, for each video clip, the annotation system creates the text corresponding to its content. We developed a method for creation of voice transcripts using speech recognition engines. It is called multilingual voice transcription and described later.

Second, some video analysis techniques are applied to characterization of visual segments (i.e., scenes) and individual video frames. For example, by detecting significant changes in the color histogram of successive frames, frame sequences can be separated into scenes.

Also, by matching prepared templates to individual regions in the frame, the annotation system identifies objects. The user can specify significant objects in some scene in order to reduce the time to identify target objects and to obtain a higher recognition accuracy. The user can name objects in a frame simply by selecting words in the corresponding text.

Third, the user relates video segments to text segments such as paragraphs, sentences, and phrases, based on scene structures and object-name correspondences. The system helps the user select appropriate segments by prioritizing them based on the number of the detected objects, camera motion, and the representative frames.

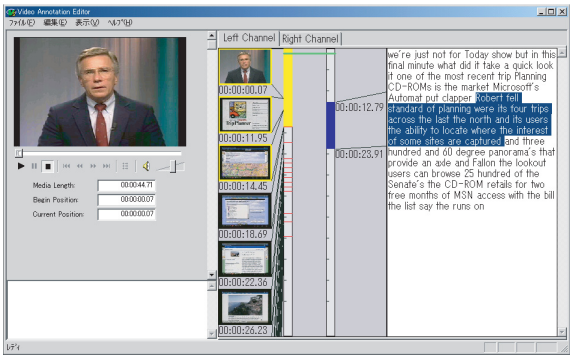
2.4 Multilingual Voice Transcription

The multimedia annotation editor first extracts the audio data from a target video clip. Then, the extracted audio data is divided into left and right channels. If the average for the difference of the audio signals of the two channels exceeds a certain threshold, they are considerd different and transfered to the multilingual speech identification and recognition module. The output of the module is a structured transcript containing time codes, word sequences, and language information. It is described in XML format as shown in Figure 2.

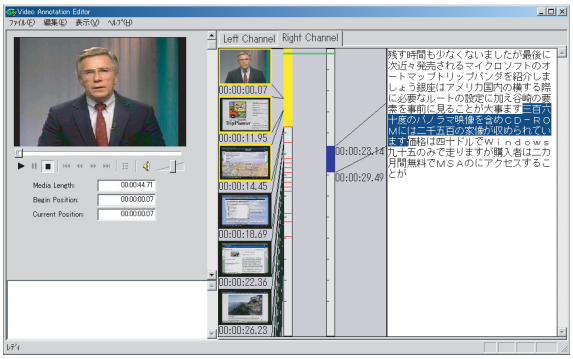
```
<transcript lang="en" channel="l">
<w in="20.264000" out="20.663000">Web grabber </w>
<w in="20.663000" out="21.072000">is a </w>
<w in="21.072000" out="21.611000">very simple </w>
<w in="21.611000" out="22.180000">utility </w>
<w in="22.180000" out="22.778000">that is </w>
<w in="22.778000" out="23.856000">attached to </w>
<w in="23.856000" out="24.215000">Netscape </w>
<w in="24.215000" out="24.934000">as a pull down menu </w>
<w in="24.934000" out="25.153000">and </w>
<w in="25.153000" out="25.462000">allows you </w>
<w in="25.462000" out="25.802000">to take </w>
<w in="25.802000" out="26.191000">your Web content </w>
<w in="26.191000" out="27.039000">whether it's a </w>
<w in="27.039000" out="27.538000">MPEG file </w>
...
</transcript>
```

Figure 2: Transcript Data

Our multilingual video transcriptor automatically generates transcripts with time codes and provides their reusable data structure which allows easy manual corretion. An example screen of the mulitilingual voice transcriptor is shown in Figure 3.



Left Channel



Right Channel

Figure 3: Multilingual Voice Transcriptor

2.4.1 Multilingual Speech Identification and Recognition

The progress of speech recognition technology makes it comparatively easy to transform speech into text, but spoken language identification is needed for processing multilingual speech, because speech recognition technology assumes that the language used is known. While researchers have been working on the multilingual speech identification, few applications based on this technology has been actually used except a telephony speech translation system. In the case of the telephone translation system, the information of the language used is self-evident; at least, the speaker knows; so there are little needs and advantages of developing a multilingual speech identification system.

On the other hand, speech data in video do not always have the information about the language used. Due to the recent progress of digital broadcasting and the signal compression technology, the information about the language is expected to accompany the content in the future. But most of the data available now do not have it, so a large amount of labor is needed to identify the language. Therefore, the multilingual speech identification has a large part to play with unknown-language speech input.

A process of multilingual speech identification is shown in Figure 4. Our method determines the language of input speech using a simple discriminant function based on relative scores obtained from multiple speech recognizers working in parallel (Ohira et al., 2001).

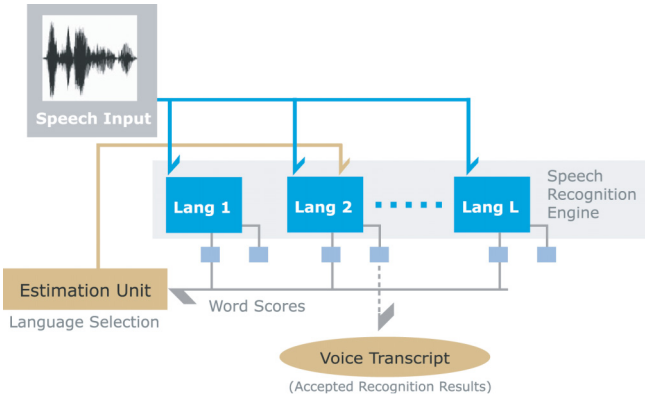


Figure 4: Configuration of Spoken Language Identification Unit

Multiple speech recognition engines work simultaneously on the input speech. It is assumed that each speech recognition engine has the speaker independent model, and each recognition output word has a score within a constant range dependent on each engine.

When a speech comes, each recognition engine outputs a word sequence with scores. The discriminant unit calculates a value of a discriminant function using the scores for every language. The engine with the highest average discriminant value is selected and the language is determined by the engine, whose recognition result is accepted as the transcript. If there is no distinct difference between discriminant values, that is not higher than a certain threshold, a judgment is entrusted to the user.

Our technique is simple, it uses the existing speech recognition engines tuned in each language without a special model for language identification and acoustic features.

Combining the voice transcription and the video image analysis, our tool enables users to create and edit video annotation data semi-automatically. The entire process is as shown

in Figure 5.

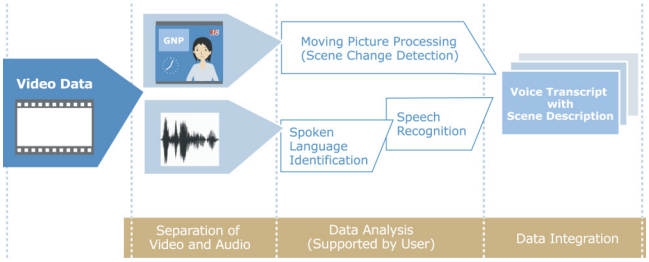


Figure 5: Multilingual Video Data Analysis

Our system drastically reduces the overhead on the user who analyzes and manages a large collection of video content. Furthermore, it makes conventional natural language processing techniques applicable to multimedia processing.

2.5 Scene Detection and Visual Object Tracking

As mentioned earlier, visual scene changes are detected by searching for significant changes in the color histogram of successive frames. Then, frame sequences can be divided into scenes. The scene description consists of time codes of the start and end frames, a keyframe (image data in JPEG format) filename, a scene title, and some text representing topics. Additionally, when the user specifies a particular object in a frame by mouse-dragging a rectangular region, an automatic object tracking is executed and time codes and motion trails in the frame (series of coordinates for interpolation of object movement) are checked out. The user can name the detected visual objects interactively. The visual object description includes the object name, the related URL, time codes and motion trails in the frame.

Our multimedia annotation also contains descriptions on auditory objects in video. The auditory objects can be detected by acoustic analysis on the user specified sound sequence visualized in waveform. An example scene description in XML format is shown in Figure 6, and an example object description in Figure 7.

3 Multimedia Summarization and Translation

Based on multimedia annotation, we have developed a system for multimedia (especially, video) summarization and translation. One of

```
<scene>
<seg in="0.066733" out="11.945279"
     keyframe="0.187643"/>
<seg in="11.945279" out="14.447781"
     keyframe="12.004385"/>
<seg in="14.447781" out="18.685352"
     keyframe="14.447781"/>
...
</scene>
```

Figure 6: Scene Description

```
<object>
<vobj begin="1.668335" end="4.671338" name="David"
      description="anchor" img="o0000.jpg"
      link="http://...">
  <area time="1.668335" top="82" left="34"
        width="156" height="145"/>
  <area ... />
</vobj>
...
</object>
```

Figure 7: Object Description

the main functions of the system is to generate an interactive HTML (HyperText Markup Language) document from multimedia content with annotation data for interactive multimedia presentation, which consists of an embedded video player, hyperlinked keyframe images, and linguistically-annotated transcripts. Our summarization and translation techniques are applied to the generated document called a multimodal document.

There are some previous work on multimedia summarization such as Informedia (Smith and Kanade, 1995) and CueVideo (Amir et al., 1999). They create a video summary based on automatically extracted features in video such as scene changes, speech, text and human faces in frames, and closed captions. They can process video data without annotations. However, currently, the accuracy of their summarization is not for practical use because of the failure of automatic video analysis. Our approach to multimedia summarization attains sufficient quality for use if the data has enough semantic information. As mentioned earlier, we have developed a tool to help annotators to create multimedia annotation data. Since our annotation data is declarative, hence task-independent and versatile, the annotations are worth creating if the multimedia content will be frequently used in different applications such as automatic editing

and information extraction.

3.1 Multimodal Document

Video transformation is an initial process of multimedia summarization and translation. The transformation module retrieves the annotation data accumulated in an annotation repository (XML database) and extracts necessary information to generate a multimodal document. The multimodal document consists of an embedded video window, keyframes of scenes, and transcipts aligned with the scenes as shown in Figure 8. The resulting document can be summarized and translated by the modules explained later.

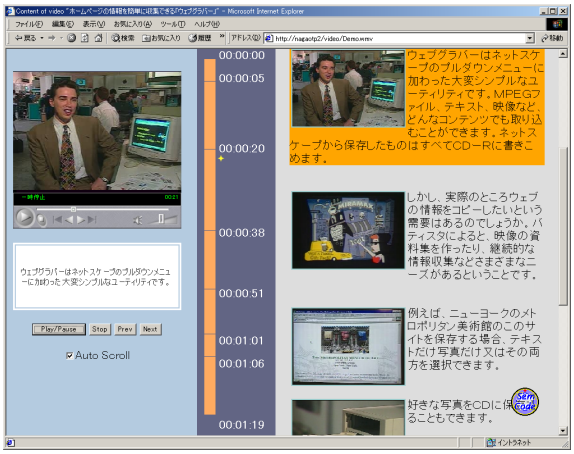


Figure 8: Multimodal Document

This operation is also beneficial for people with devices without video playing capability. In this case, the system creates a simplified version of multimodal document containing only keyframe images of important scenes and summarized transcripts related to the selected scenes.

3.2 Video Summarization

The proposed video summarization is performed as a by-product of text summarization. The text summarization is an application of linguistic annotation. The method is cohesion-based and employs spreading activation to calculate the importance values of words and phrases in the document (Nagao and Hasida, 1998).

Thus, the video summarization works in terms of summarization of a transcript from multimedia annotation data and extraction of the video scene related to the summary. Since a summarized transcript contains important words and phrases, corresponding video sequences will produce a collection of significant scenes in the video. The summarization results in a revised version of multimodal documemt that contains keyframe images and summarized transcripts of selected important scenes. Keyframes of less important scenes are shown in a smaller size. An example screen of a summarized multimodal document is shown in Figure 9.

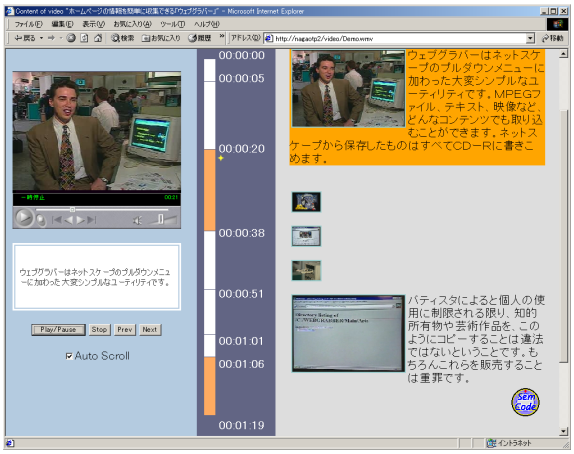


Figure 9: Summarized Multimodal Document

The vertical time bar in the middle of the screen of multimodal document represents scene segments whose color indicates if the segment is included in the summary or not. The keyframe images are linked with their corresponding scenes so that the user can see the scene by just clicking its related image. The user can also access information about objects such as people in the keyframe by dragging a rectangular region enclosing them. The information appears in external windows. In the case of auditory objects, the user can select them by clicking any point in the time bar.

3.3 Video Translation

One type of our video translation is achieved through the following procedure. First, transcripts in the annotation data are translated into different languages for the user choice, and then, the results are shown as subtitles synchronized with the video. The video translation module invokes an annotation-based text translation mechanism. Text translation is also

greatly improved by using linguistic annotation (Watanabe et al., 2002).

The other type of translation is performed in terms of synchronization of video playing and speech synthesis of the translation results. This translation makes another-language version of the original video clip. If comments, notes, or keywords are included in the annotation data on visual/auditory objects, then they are also translated and shown on a popup window.

In the case of bilingual broadcasting, since our annotation system generates transcripts in every audio channel, multimodal documents can be coming from both channels. The user can easily select a favorite multimodal document created from one of the channels. We have also developed a mechanism to change the language to play depending on the user profile that describes the user's native language.

4 Concluding Remarks

We have developed a tool to create multimedia annotation data and a mechanism to apply such data to multimedia summarization and translation. The main component of the annotation tool is a multilingual voice transcriptor to generate transcripts from multilingual speech in video clips. The tool also extracts scene and object information semi-automatically, describes the data in XML format, and associates the data with content.

We also presented some advanced applications on multimedia content based on annotation. We have implemented video-to-document transformation that generates interactive multimodal documents, video summarization using a text summarization technique, and video translation.

Linguistic processing is an essential task in those applications so that natural language technologies are still very important in processing multimedia content.

Our future work includes a more efficient and flexible retrieval of multimedia content for requests in spoken and written natural language. The retrieval of spoken documents has also been evaluated in a subtask "SDR (Spoken Document Retrieval) track" at TREC (Text REtrieval Conference) (TREC, 2002). Johnson (Johnson, 2001) suggested from his group's experience on TREC-9 that new challenges such as use of non-lexical information derived directly from the audio and integration with video data are significant works for the improvement of retrieval performance and usefulness. We, therefore, believe that our research has significant impacts and potetials on the content technology.

References

A. Amir, S. Srinivasan, D. Ponceleon, and D. Petkovic. 1999. CueVideo: Automated indexing of video for searching and browsing. In *Proceedings of SIGIR'99*.

CES. 2002. Corpus Encoding Standard. http://www.cs.vassar.edu/CES/.

EAGLES. 2002. EAGLES online. http://www.ilc.pi.cnr.it/EAGLES/home.html.

Koiti Hasida. 2002. Global Document Annotation. http://i-content.org/GDA/.

S. E. Johnson. 2001. Spoken document retrieval for TREC-9 at Cambridge University. In *Proceedings of Text REtrieval Conference (TREC-9)*.

MPEG. 2002. MPEG-7 context and objectives. http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm.

Katashi Nagao and Kôiti Hasida. 1998. Automatic text summarization based on the Global Document Annotation. In *Proceedings of the Seventeenth International Conference on Computational Linguistics (COLING-98)*, pages 917–921.

Shigeki Ohira, Mitsuhiro Yoneoka, and Katashi Nagao. 2001. A multilingual video transcriptor and annotation-based video transcoding. In *Proceedings of the Second International Workshop on Content-Based Multimedia Indexing (CBMI-01)*.

Michael A. Smith and Takeo Kanade. 1995. Video skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-95-186, School of Computer Science, Carnegie Mellon University.

TEI. 2002. Text Encoding Initiative. http://www.uic.edu/orgs/tei/.

TREC. 2002. Text REtrieval Conference home page. http://trec.nist.gov/.

W3C. 2002. Extensible Markup Language (XML). http://www.w3.org/XML/.

Hideo Watanabe, Katashi Nagao, Michael C. McCord, and Arendse Bernth. 2002. An annotation system for enhancing quality of natural language processing. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-2002)*.