An Overview of Content-utilization Technologies

There is a need for broadcasters to provide viewers with easy access to the large quantities of video content that they have produced through mechanisms that link broadcasting with communications technology. A basic requirement for this is the ability to search through video content on the basis of the meaning of the content. In this article, we provide an overview of concepts and researches related to content recognition, search, and recommendation technologies of new video content services.

1. Introduction

According to the Japanese act, named "Act to Promote the Creation, Protection, and Utilization of Content" (commonly referred to as the "Content Protection Act"), content is defined as "Includes movie, music, theater, literature, photography, comic (manga), animation, computer game and other text, graphics, color, sound, performance and video, or creations combining these elements, or programs for presenting information related to such creations using digital computation devices (meaning the combination of commands given to a computing device to produce a particular result), and is the product of human creative activities that falls into a cultural or entertainment category." According to this definition, the content conceptually encompasses all results of human creativity in the categories of culture and entertainment and includes computer-programs used for presenting such creations. Such products of human creativity have conventionally been called "works" or "productions", but when the term "content" is used, as above, the nuance is more often of creations presented using a computer. Because information must be digitized in order to be processed by computers, we can say that through digitization, the "works" of the analog era come to be referred to as "content".

Thanks to digitization, programs that were first broadcast in the analog era have also joined the set of "content" and can be called broadcast content. Digital broadcasting can also present programs containing a variety of other data types besides video and audio. Moreover, a variety of information related to a program being broadcast can be presented using media such as the World Wide Web. This information can also be called broadcast content since it is provided together with the main broadcast.

In this article, we review the research and development on technologies for presenting broadcast content to more viewers in ways that are easier for them to use. Because broadcast content is mainly composed of non-verbal information (video and audio), metadata, which is data that describes data, is very important for enabling easy access to and facilitating operations on such content. For broadcast content to be used in new ways, we need to be able to extract metadata from the content with as little human intervention as possible and use that metadata to search for and recommend content suited to the viewing situation.

2. Insights into Content-utilization Technologies

2.1 The Role of Metadata

The most basic technology for utilizing content is a search engine that makes it easy for a user to access and view the desired content. The Web, which is the mechanism for distributing and viewing content over the Internet, is now an everyday fixture in many people's lives, and search engines operating on the Web play an enormous role in promoting the spread of the Web.

Figure 1 shows the conceptual organization of a search engine. The search engine gathers content, analyzes the documents to decompose them into words while referring to the HTML tags, and selects index words in consideration of factors like word significance.
They then build an index of these index words and the URLs of the Web pages containing them, which are stored in the index DB. A user's query is in the form of keywords, so the search engine finds search results by matching keywords to the index words in the index to output the conforming Web pages, together with information such as the URLs and titles of the pages. Documents are analyzed and multiple results are ranked, and this process has been developed on the basic technologies of natural language processing and information retrieval.

With the development of these and network technologies, content can now be efficiently gathered from the whole of the Web by search engines, and document content can be analyzed to generate and manage large-scale index DBs automatically. These factors have contributed to the spread of Web search engines.

Most of the broadcast-related content provided on the Web is text, and it can be provided to users through search engines with the same mechanism as described above. In contrast, video and audio content, which is mainly non-textual (hereafter referred to as "video content"), is most often provided as a digital broadcast over radio or cable television (CATV), and it runs in a particular time sequence. On the other hand, provision of such content in an on-demand fashion through dedicated networks or the Web has also become possible recently. Furthermore, many video archives have been created and opened to the public. With the on-demand video content expanding, simple content selection methods such as hierarchical classifications are becoming inadequate, and there is increasing demand for higher-level content search functions, like those for Web content, but suited to the characteristics of video content.

In order to build an efficient content-based search engine for video, an index must be created for collating user queries and video content. However, video and audio consist of continuous time-sequence data, and they have no well-defined basic unit of meaning that is equivalent to words in a text document. Motion video data can be considered as a sequence of still images, so it seems that individual still images could be taken as the basic physical unit. However, such image units do not necessarily carry meaning in the way that words do. To search Web content, an index is built using words as units that carry meaning, and users also express their requests using words (keywords). Thus, index words and keywords can be directly collated. If an index is built using still images instead of words, it will not be a simple matter for users to express queries in terms of still images, and it will be difficult to implement a usable search function for video content using this approach. To make a content-based video search, we need data describing the content at a level that can be collated with the user's queries. This content-descriptive data, or metadata, is of a higher level than the video content, and must be created by attaching it manually to the content or by using...
content analysis to extract it automatically from the content.

2.2 Research on Content-utilization Technologies

As described above, an important step to utilizing video content is to create metadata. The most basic role of metadata is to provide a content-based index for video that can be collated with user’s queries. Metadata describing video content in natural-language-like symbols is needed to handle queries expressed using keywords. Video content also develops over a sequence of scenes that each has meaning; hence, metadata must also contain information regarding the boundaries between individual scenes. Besides this, we need a way to obtain information regarding the image characteristics of scenes so that searches for similar scenes can be made. This sort of information is difficult to express in natural language. The ultimate objective of content analysis research for content utilization is to develop technology for creating this sort of information efficiently.

In content-analysis research domain, researchers are developing technologies that will enable a computer to analyze video and audio content and extract the content-descriptive information required for making this metadata. This requires pattern-recognition technology. As the processing and memory capacity of computers has increased, it has become possible to use machine learning on large quantities of training data and various other statistical techniques for pattern recognition. Research is progressing on using these techniques to perform integrated analyses of the video, audio and language data that comprise video content.

It is very important for a search engine to be able to analyze large quantities of content efficiently. For Web content, it is possible to create an index directly from the text without having to refer to any other information that may be present. In this case, the words that make up the index are extracted from the content itself and contain information at the same level as the content. The metadata for Web content is basically attached by adding special tags when the content is created. It includes information, such as author and copyright, that cannot be obtained by analyzing the content itself but can be used to conduct searches with more-detailed retrieval conditions.

On the other hand, since video content contains a large quantity of data, it is difficult for search engines to collect such content from a broad scope and analyze all of them. Accordingly, for implementation of a large and exhaustive search service, we need a way for the content provider to create descriptive metadata and allow the search engine to collect this data efficiently. To enable this, it is important that standardized, interoperable metadata be created, shared and used. The research on content-utilization technology is intended to promote the standardization and to develop an infrastructure for creating metadata and methods for searching, recommending and presenting video content.

3. Trends in Content-use Technologies

3.1 Content-Analysis Technologies

As described above, information that expresses the content of video, or metadata, would be helpful in searching for programs or scenes within the huge quantities of broadcast content. Here, we give an overview of video analysis technology for automatically extracting metadata from content by computer. Besides techniques to analyze the images themselves, we will also look at techniques that process text in subtitles and scripts and relate them to images.

(1) Shot boundary detection

It would be desirable to divide up video into scenes having different meanings because it would enable searches to be based on the meaning of the content. However, this is a very difficult to do since video content has no well-defined basic unit of meaning. On the other hand, video content, and particularly broadcast programs and movies, is composed of many shots. By shot, we mean a segment of video captured continuously by a single camera. By necessity, techniques for detecting shot boundaries which are not units of meaning were studied during the early stages of research in the content analysis field. Dividing video into shot units is now relatively easy.

The TREC Video Retrieval Evaluation (TRECVID) workshop to evaluate video search technology began in 2001, and it conducted shot-boundary tasks until 2008. TRECVID is backed by the National Institute of Science and Technology (NIST) in the USA and other organizations, and each year it sets out about four tasks for teams from academia and industry to perform. Participants are given a fixed amount of time to report on their computational results for a common set of test videos. The shot-detection method presented by NHK STRL in 2008 had excellent performance in terms of speed and accuracy\(^1\).

(2) Generic object recognition\(^2\)-\(^6\)

Object names are a basic type of metadata, and object

Face-image processing

People are undoubtedly the most frequently appearing objects in broadcast programs and movies. Because of this, there is much research on generating metadata related to people appearing in video. Face detection technology has advanced to the point where even consumer digital cameras embody algorithms for detecting faces in images have been. Image processing is computationally expensive, and for a long time, it was difficult to develop a technology that could detect accurately and quickly. However, a few years ago, Viola et al. devised just such a method, and since then there have been rapid advances in facial-image detection technology. In particular, STRL has devised a stable facial image detecting technology based on Viola et al.’s method.

On the other hand, information regarding who appears in a video cannot be consistently obtained from only facial-image detection results. In particular, the task of facial-image recognition is to determine the name of a person from his or her facial image, and many methods have been proposed for this. In the task of facial-image recognition, however, each new person must be registered in a database, and this process is generally quite complex. STRL is developing a method for registering facial images of unregistered persons relatively easily. Kumar et al. proposed a method for deciding identities of persons from the results of 65 discriminators including gender, race, age, hair color, and expression. This method can handle various changes in images of a particular person.

In general, it is hard to register many faces of unregistered individuals manually for facial image recognition. Thus, multi-modal processing methods have been proposed to gather facial images and related textual information, which includes person’s names, being added manually for other purposes and use them as training data for the person being registered. Satoh et al. developed a method that assigns name metadata to images on the basis of co-occurrence rates of names appearing in open captions and faces appearing in the video. Ramanan et al. semi-automatically created a data set of approximately 600,000 people from television video covering 11 years. They also reported on how the dataset allows records to include changes such as aging or changing physical appearance in individuals. For movie content, the script can also be used. Text from the script and subtitles can be linked to various facial images obtained through face detection processing, and character names can be automatically attached to the scenes where the actor appears. Everingham et al. developed a method in which (1) the script for each actor is compared with the subtitle timing to link the actors' names and the subtitles, (2) people are identified using this information and facial images detected while the actor is talking, and (3) faces are matched to the clothing the actors wear so that names can be attached to the overall scenes in which human body appear. Pham et al. devised a method using facial image detection and captions for attaching names to people appearing in online news videos on the Web. This method works for

---

3 A local feature value for images. The images are filtered using normal distributions of several different standard deviations, points that are stable with respect to shot changes (keypoints) are located, and a histogram of the block locations and changes in brightness for areas surrounding these points is created to form the feature value.

4 A feature value with reduced computational cost relative to that of SIFT can be created by simplifying and increasing the speed of normal distribution computations.
scenes with a number of people in them.

- **Human body detection**
  Face-image detection performs well when persons in the image directly face the camera, but detection rates decrease when the orientation deviates from face on. As well, detection rates also drop when the subject is wearing sunglasses or a hat. Detecting human bodies rather than faces is seen as a way of overcoming such difficulties. One such method uses features called Histograms of Oriented Gradients*5 (HOG)*22. HOG feature values are created using histograms of local changes in the brightness of the image, so they are thought to be good for identifying general shapes, and in turn for identifying human bodies which have relatively large shape variations. Some methods can recognize people by identifying clothing colors that have already been matched to names*16, 19. However, these methods assume that the actor's clothing does not change, so care must be taken when using them.

(4) Event detection
The results of person detection and recognition can provide metadata for responses to search requests such as "scenes with people" or "Mr. ABC", but they cannot provide responses to requests involving a person's state of being or what they are doing. For instance, for sports video to be searched, important events have to be specified. Such events include home runs in baseball or shots taken in soccer. The technology for attaching metadata to scenes containing such occurrences is called event detection.

- **Human posture, action, and behavior detection**
  Posture provides basic information about the state of the body, then technologies have been proposed for posture searching in the form of a query. Ferrari et al.*23 devised a method for differentiating body postures by detecting the head, torso, upper arms and forearms separately. They attempted to detect postures such as placing hands on the hips or folding one's arms. There are also many reports on searching for actions. As with names, scripts and subtitles can be used in movies for this purpose, and labels are attached to actions by drawing relationships between text and image features*26, 29. Another sort of methods relies on captions in images found on the Web. Images depicting certain actions can be automatically gathered from the Web and used to detect particular actions in videos in YouTube*28. As with generic object recognition, it adapts text processing technology in an attempt to recognize specific actions. For example, recognition methods using Probabilistic Latent Semantic Analysis*6 (pLSA)*27 and tensor analysis*7 have been proposed*29. The events recognized include a time component and various postures in various orientations; hence, a wide variety of 2D-image time sequences must be compared in order to discriminate between actions. The problem can be made more tractable by using context to increase recognition performance*30-32. Context here refers to other objects and background information in areas surrounding the phenomenon being recognized. As an example of the benefits of using context, it would be difficult to differentiate scenes of playing a piano and typing on a computer keyboard by observing only finger motions, but if the piano keyboard or computer were recognized in the background, the actions could be distinguished relatively easily by using the contextual (background) information as constraints.

Since 2008, TRECVID has set out event-detection tasks for detecting people performing specific actions, such as running or talking on a mobile phone, in video recorded by cameras installed in an airport. In 2009 NHK STRL came in first place among thirteen teams in the "Running" event task. The method detects people using HOG features and Support Vector Machine (SVM) recognition*7. Then it analyzes the locus obtained by tracking the person's motion*26.

- **Detection of events in sports coverage video**
  Sports fans would naturally like to see missed events in a short time after they occur in a game and find particular occurrences within sports video. To make this possible, metadata can be attached to each event in the video. The research focuses on detecting sports-related

---

*5 A feature value that uses histograms of changes in brightness similar to SIFT, but rather than using keypoints, it describes regions around arbitrary fixed points.

*6 A model in which text or images are generated through a probabilistic process by way of hidden (latent) parameters. For example, a model for which text or images are probabilistically generated by the latent parameters of the topic of the document or the category of the image.

*7 A method that assigns different characteristics to each array of a tensor (multi-dimensional array), which is a mathematical extension of the vector concept. The most significant feature values are then selected by performing an analysis similar to singular value decomposition for arrays.

*8 A machine learning technique in which the recognition is acquired automatically using a large amount of correct data, basically by creating a two-value discriminator function (e.g.: whether it is a face or not, etc.). Its recognition performance is good for unknown data.
events and generating digests of game coverage. Gupta et al. presented a method that trains models for each event, learns automatically from subtitles and transitions in baseball events, and expresses actions and cause-effect relationships in a tree structure called a Storyline. STRL researchers devised a method that exploits the fact that switches from one camera to another in baseball coverage follow typical patterns, and a sequence of such shots can be used to detect events such as home runs or hits. They also proposed a method that attaches metadata identifying pitch types to baseball video. For soccer, however, play on the field is more intense than in baseball, and the sequence of shots is less predictable. Thus, attempts have been made at classifying soccer video content on the basis of the players’ positions on the field and their movements.

3.2 Video Summarizing Technology

Summary videos are very useful for checking many search results quickly to see if they contain the desired content. There are two types of summary video: program highlights and trailers designed to entice viewers to watch the video. There has been very little research on the second type, since it is a difficult problem. NHK STRL’s method automatically generates an introductory (trailer) video for a program by processing a textual summary of the program’s content and correlating it with subtitles.

3.3 Video Search Technology utilizing Physical Characteristics

The atmosphere of an image or impression from the music can be used as a search indicator, even if the image or music in the video cannot be identified by the viewer. There are search methods that use colors and patterns or even the overall composition of images as well as ones based on similarities in the melody or rhythm of music. There have also been proposals for using sketches as queries. STRL researchers have devised a way to make search queries by gathering image fragments that appear frequently in collections of search images, presenting a palette these fragments to the user, and allowing the user to select combinations of fragments. Another method entails the user drawing objects and names attaching the names on the drawn objects on the query screen. The method returns images containing each object, and a natural composite image can be easily created by automatic selecting combinations of the retrieved images.

Research is also advancing on technologies that analyze images in order to find other images of the same scene but taken from different viewpoints or at different times. One such method focuses on motion in the video. It attempts to find video of the same scene by tracking temporal changes in the locus of characteristic points.

3.4 Metadata Production Systems

Metadata used for actual services must be reliable, so manual editing is necessary even when functions that automatically extract metadata from content can be used. Editing is also necessary for bibliographic metadata; content analysis cannot extract author names, copyright notices, or usage permissions.

To improve the efficiency of such editorial work, metadata production systems that combine multiple automatic feature analysis functions, editors and DBs are being developed in cooperative research projects in Europe. These include 4M of the MUSCLE project, M-OntoMat-Anotizer of the aceMedia project, MAD of the PrestoSpace project, and iFinder of Fraunhofer Laboratories. Software from several of these projects is publicly available. Structures for linking subsystems developed by other organizations are also being built. STRL has created the Metadata Production Framework and has released a software package including a metadata editor that integrates various automatic analysis functions and is able to edit metadata.

3.5 Video Content Recommendation and Presentation Technologies

At STRL, we are studying search mechanisms to recommend content automatically to viewers while they are watching television. These systems compare

---


metadata attached to the program or scene being viewed with metadata attached to content being searched and recommend related content to the viewer. We are drawing up a standard set of specifications for metadata interfaces, search and display systems, that will enable the systems providing services and terminals displaying information to be designed separately as long as the systems are developed based on the specifications. We are also looking into automatic methods of building multimedia encyclopedias and generating quiz content using image and natural language analysis.

4. Conclusion

There is broad range of research on image analysis technology for video retrieval, and we have only addressed a small part of it in this article.

Despite TRECVID having set high-order feature-extraction tasks for detecting object and event names, the level of the results achieved by even the best participants is still far from practical. This failure illustrates the so-called semantic gap, i.e., the difficulty of extracting content on a conceptual level from physical values. Many promising image analysis ideas have been proposed for automatically attach metadata, but so far none has been able to bridge the semantic gap.

There are many types of search requests, so a wide variety of technologies for extracting meaning from contents are needed to handle them. What is needed is an environment in which many researchers and technologists can cooperate in developing ways to create metadata automatically.

(Mahito Fujii and Masahiro Shibata)

2) TRECVID Web page:
   http://www-nlpir.nist.gov/projects/trecvid/


49) MPF Web page: http://www.nhk.or.jp/strl/mpf/index.htm