Digital Video Library Network: System and Approach

Master of Philosophy
Research Project First-Term Report

Supervisor

Professor Michael Lyu

Prepared by

Ma Chak Kei (00315340)

Department of Computer Science and Engineering
The Chinese University of Hong Kong

Abstract

Tremendous growth of the Internet population creates a large demand on new applications, and advances in Internet technologies make it feasible to develop new exciting application base on video and broadband network. One of the most hottest topic nowadays is the Digital Video Library Systems. It has a promising application scope in entertainment, information, education or business. However, due to the temporal nature of video, indexing and retrieval of video content is not trivial. Therefore, we will introduce the digital video library network in this paper with a number of techniques concerning the indexing and retrieval of video contents.

Contents

Introduction

Tremendous growth of the Internet population creates a large demand on new applications, and advances in Internet technologies make it feasible to develop new exciting application base on video and broadband network. One of the most hottest topic nowadays is the Digital Video Library Systems, which may becomes the most popular Internet services in many areas including entertainment, information, education or business.

Application of Digital Video Library ranges from education, entertainment, tourist information, cultural services, shopping, interior design, multimedia portals and etc. With the Digital Video Library System, home users may watch movies or sports program on Internet; students may watch recorded lectures/seminars for learning or search for historical event videos for project; people may also search for tutorial video about assembling a computer; or search for news recorded to collect useful information for business decision.

Depending on the content provider and the targeted customer group, sources of video may different a lot in the narrative, formats, and other properties. They may be TV Programs, News, Movies, MTV, Records of Seminars, or even some Synthesized Animations. Furthermore, they may be classified as real-time and stored contents, which may lead to differences on the processing of video.

In this paper, we will describe the system architecture of Digital Video Library in Section 2. Followed by the details of the 4 major components: Video Server, Indexing Server, Query Server and the User Application. Followed by some related works, we will have a brief conclusion at the end.

System Architecture

To provide services on the Internet, the system should attained high availability, high performance with high extensibility. And to make advantage on the Internet environment, we can develop the system in a sense that different components can reside in different computers. Many concepts in Web server can be applied here.

The DVLS is a complex system composed of 4 primary components: Video Server, Indexing Server, Query Server and Client Applications. Base on this architecture, the system may be modeled as a single DVL-Workstation or as distributed systems over the Internet.

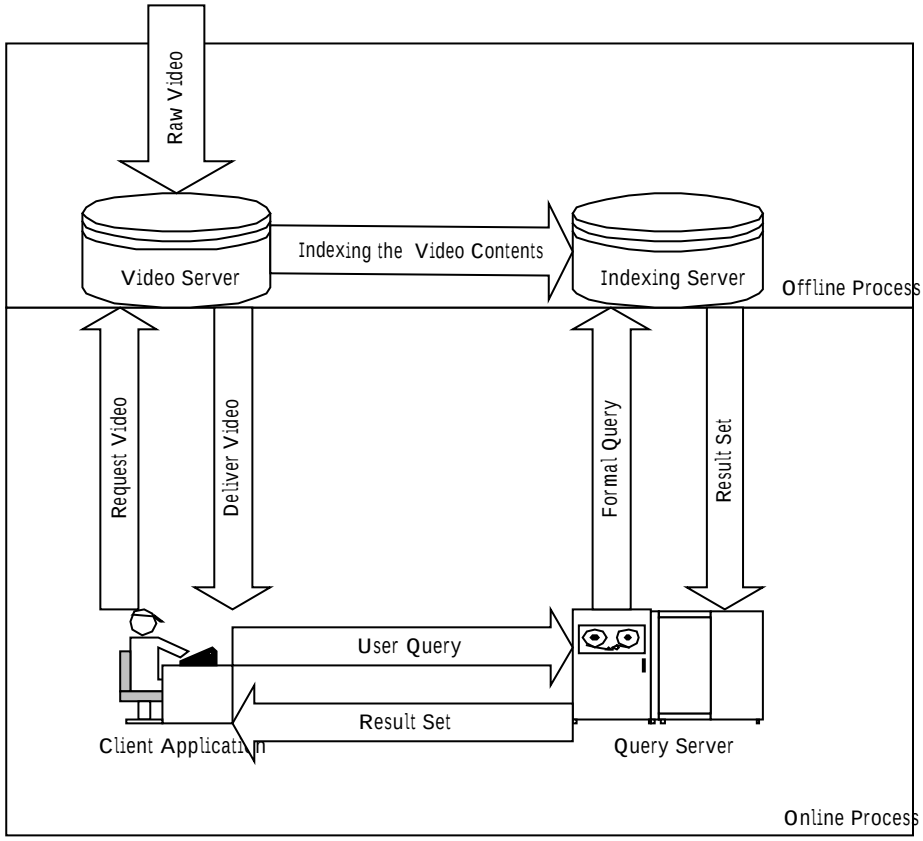

Figure 1: System Overview

Figure 1 shows the relation between these four components as an independent system. At first, the raw video will be segmented by the Video Server into size suitable for storage and stored together with the associated information. The Indexing Server will process the videos, extract various features and keep the index in the database

management system for future query. These offline stages may be carried out as batch jobs or as background jobs as far as they would not affect the online part which will actually affect how the user feel about the system performance. For the online part, beginning with the queries submitted by user at Client Application, then the Query Server will construct formal queries to get the result set from the Indexing Server, and in turns return the results to the Client Application. Finally, the user may request the video from the Video Server base on the returned results and the required video will be delivered to the Client Application.

By separating the tasks to different servers, it is possible to construct a Digital Video Library Network, where each server may gather information from multiple sources, and serve multiple clients as well. In such a way, the sharing of resource will lead to a system that is much more powerful. Figure 2 shows how the components maybe connected to form the Digital Video Library Network.

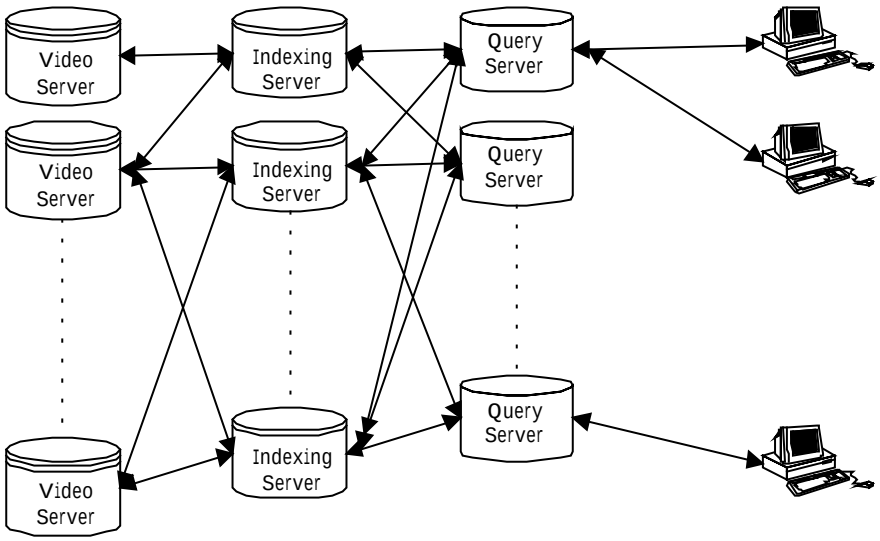

Figure 2: DVL Network

In such system design, each server may be just part of the “middleware” of the DVL Network, providing various services in its specialized field. Moreover, it is also possible to work with some existing systems such as search engines, meta-search engines, web browsers and other tools for handling video documents. To allow multiple servers working in a compatible way, it is useful to have a specification of protocols when various components are involved [1].

In some cases, components can also combined to do some complicated operations. For example, a client application can do the job of query server by itself, such that it can connect to some specific indexing servers. A video server may be combined with

an indexing server, by integrating MPEG-4 standard, it can provide the video indexes and deliver the video using a single server through a single interface.

Video Server

A Video Server is specialized in capturing, storing and delivering video contents. Various sources of video may differ in their physical properties; some may be stored in video tapes or laser disc; some may come in digital formats already; some maybe available in analog form in air broadcasting. They will also differ in their semantics and structure, some of them may be business interview, or TV Programs, News segments, MTV, movies, educational materials and etc.; this will lead to varies in length of video segments, and also different in the techniques to process the videos.

The video server may serve for some particular sources, it is likely that a Content Provider (e.g. a TV company) may own its video server to server their videos, or there may be public video servers that individuals can submit their videos and share it with others. Different video servers may have different requirements and different features, for example, level of security, support of money transaction, streaming ability, and the maximum number of concurrent connections.

Video Storage

Whatever the input video sources are, the Video Server will store the segmented videos in its database in digital formats. This may involve video segmentation and format conversion.

For videos that represent a single entity and import as a file at a time, video segmentation is not necessary. While for videos that consist of multiple entities or those captured as video streams from analog media, video segmentation will be carried out to yield the smaller pieces for storage and retrieval purposes.

For automatic video segmentation, much research has concentrated on segmenting video streams into “shots” using low-level visual features. With such segmentation,

the retrievable units are low-level structures such as clips of video represented by key frames which is not the users usually wanted. Instead, we can carried out semantic segmentation using multimedia cues which return requested information that is long enough to be informative and as short as possible to avoid irrelevant information [2] [3].

Some work on this issue had been carried out, by using a single medium (visual or text) or integrated the cues from different media. Some existing approaches make use of fixed textual phrases or exploit the known format of closed captions. Phrases like “still to come on the news…” and other known captions are used as cues for story boundaries [4] [5]. Some other techniques are aimed at developing a more general approach that does not make assumption about some fixed phrases or captions [6]. Audio, visual, and textual signals are utilized at different stages of processing to obtain story segmentation. For example: audio features are used in separating speech and commercials; anchorperson segments are detected on the basis of speaker recognition techniques; story-level segmentation is achieved using text analysis that determines how blocks of text should be merged to form news stories. This process requires the Video Server to have knowledge about what kind of video it is processing which may need special customization, but the resulting video segments can convey the semantics more effectively and meet the needs of different users.

After the video is segmented, and also for those video clips that do not need to be segmented, media formats conversion may be needed in cases that a format is superior to another in terms of features provided, compression ration, video quality and streaming ability. But this kind of conversion is optional.

Meta-Media Attributes

Some kinds of information are related to but not “within the video”, and some of the information are impossible to be extracted from the video with current technology. These kind of attributes have to be input manually at the time the video is imported into the system, including:

- Production Feature – date of data acquisition, producer, director, performers, roles, production company, production history.
- Media Feature – duration of video, streaming protocol, resolution and frame rate.

- Text Description – events, activity, closed-captions and annotations
- Intellectual Property Information – copyright, licensing, and authentication information.
- Reference – some external references associated with the video

The main “client” of a Video Server is the Indexing Server and the Client Application. Both of them will use these attributes to provide better services. For the Indexing Server, the considering of meta-media attributes will lead to better understanding about the video content, and thus lead to more accurate indexing on both semantics and video object extraction. With some of these attributes like, performers, date, reference, etc., it can provide alternative searching methods for users who are interested not only in video contents. It can also clustering the videos according to different attributes to create categories that may suitable for video-surfers who do not have a particular video in mind. For the Client Application, it may display some of the meta-media attributes that the users may interested in, it may also use the information to connect to reference database to retrieve related information, to optimize the network connection, and to give use a choice of frame rate and resolution.

Beyond these basic meta-media attributes, it is also possible to define some advance attributes for other purpose, for example: signature to identify the video file for security; payment information for money transaction; links to previous/next chapter for TV programs. But the most important issue is to have the corresponding processor for the attributes in Indexing Server, Client Application, or other specialized components.

Video Delivery

Video content must delivered to the user application in order to be “useful”. For a Video Server, the main concern is the number of clients it can serve at the same time, and the quality of services it can provide.

The throughput of a system is limited by its computing power and the network bandwidth. To increase the number serving connections in a time, techniques like load balancing and multicast may be use, which are out of the scope of this paper. On the other hand, traditional protocols used in Internet including HTTP and FTP are implemented base on TCP, which are unsuitable for continuous media such as real time audio and video. And thus there are some new protocols being developed (e.g. RTP) which addressing the streaming requirement of these real-time media data.

These new protocols change the latency for the starting the video from the loading time of whole video files to the loading time of video header and the buffered content. They also exploit the error tolerance nature of video contents to minimize the number of retransmission of data. In fact, these advances in protocol do not change the throughput of the system but increase the utilization of system.

The need to define quality of service (QoS) arises from the realization that users require different quality presentations at different times. When a multimedia presentation is transmitted via a network, it translates into different requirements of network performance. Because the ultimate consumer of the system is human, thus the quality of presentation is a matter of the user's perception, which is limited by the response of the human vision and auditory senses. This perceptual nature of QoS makes it subjective and difficult to quantify precisely.

A complete QoS specification must consider all aspects of the system's presentation, hardware, and software components. Nalin K. Sharda suggest a modeling of parameters in three different perspectives: user perspective, application perspective, and transmission perspective [7]. Various QoS parameters are listed in the QoS Venn diagram below in Figure 3.

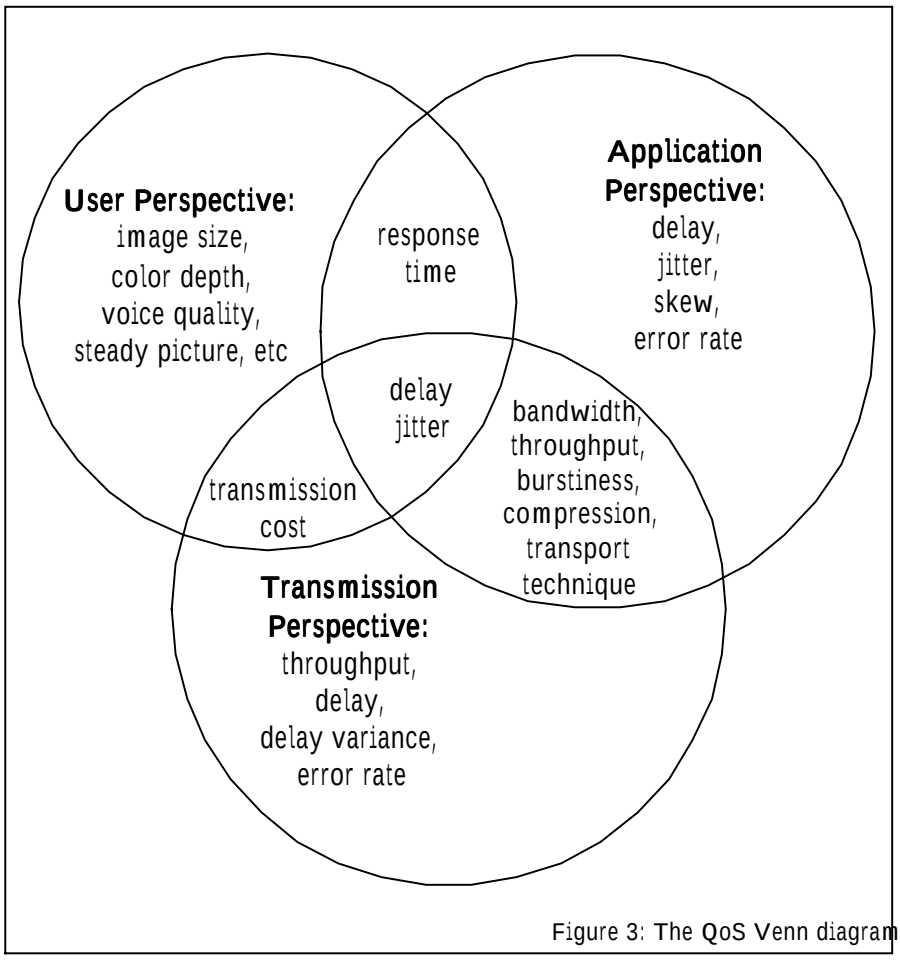

Figure 3: The QoS Venn diagram

In the QoS Venn diagram, the user perspective is a function of the ability of the human senses to distinguish between different-quality presentations, including image size, color depth, voice quality and steadiness of the picture. As high quality usually demand more on the presentation hardware and network bandwidth, a user may like to specify some cost parameters that reflect the actual situation. The application perspective including delay, jitter, skew, and error rate that are directly related to the performance of the application. As respond time is a useful parameter from the user perspective, it lies at the intersection of user and application perspectives. Moreover, network bandwidth, throughput, traffic burstiness, transport services (e.g. connectionless or connection-oriented), and compression techniques lie at the intersection of application and transmission perspectives. For the transmission perspective, the main parameters are throughput, delay, delay variance and error rate. Since a user may want to specify cost parameters such as maximum cost to be incurred per unit time, therefore the transmission costs is placed at the intersection of the user perspective and the transmission perspective.

The network service chosen to deliver the video is decided through a process of

negotiations. Such negotiations processed in layers in a way similar to the models used in network protocols. The three layers used in our example are user perspective layer, application perspective layer, and the transmission perspective layer [7]. For each layer, there will be three processing steps [8]:

- Assessment of the QoS request
- Mapping of the QoS request into QoS parameters
- Negotiating parameter values.

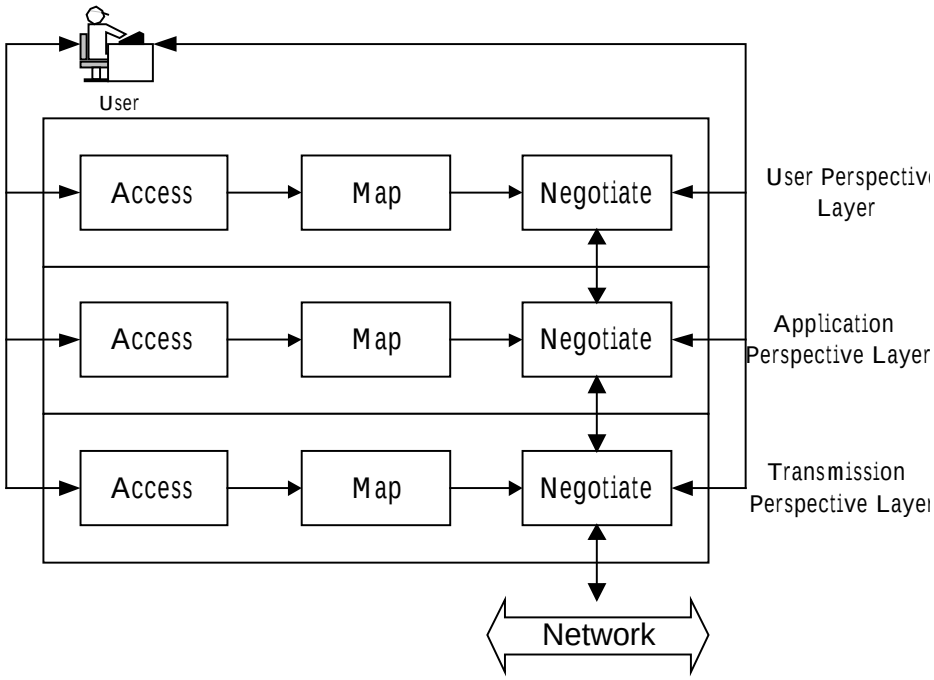

Figure 4: A QoS processing model

The QoS processing model is shown in Figure 4. The user may specify the parameter he want in the user perspective layer, for each layer, the values obtained from upper layer will be transformed to the parameters in current layer, and then negotiate with the subsequent layer. This process repeats until the transmission perspective layer meet the network, where the network systems will granted the access and guarantee the requested parameters, or it will reject the request.

Indexing Server

The Indexing Server is specialized at indexing the video contents for fast query and retrieval. Functionality is depended on the video features that the server is able to extract and index. Specialized Indexing Server may be developed to provide specific features like video caption indexing, company logo indexing, or face recognition. From the complexity level of processing needed, we can classify the index into textual information, physical feature and semantic features.

Textual Information

Raw data are mainly textual information, which includes meta-media attributes provided by the content creator, and perhaps the full-script generated by speech recognition. For textual information, there are many well-developed theories of information retrieval can be applied. Besides fields that only required exact matching (e.g. Producer Information), there are fields like close-caption, annotation, or other textual description, which can be treat as ordinary text documents. As not all words are equally significant for representing the semantics of the document, we will process the text of the documents in the collection to determine the terms to be used as index terms. The document preprocessing includes [9]:

- Lexical analysis - treating of digits, hyphens, punctuation marks, and the case of letters.
- Elimination of stopwords - filtering out words with very low discrimination values for retrieval purposes.
- Stemming - removing affixes and allowing the retrieval of documents containing syntactic variations of query terms.
- Selection of index terms - to determine which words will be used as an indexing element. Usually, the decision on whether a particular word will be used as an index term is related to the syntactic nature of the word. In fact, noun words frequently carry more semantics than adjectives, adverbs, and verbs.
- Construction of term categorization structures - allowing the expansion of the original query with related terms with structures such as a thesaurus.

After the documents are processed, they may be kept in record in the database for

retrieval use.

Physical Features

By exploiting the low-level objects and the associated features in images or videos, object-based query becomes possible. In a feature-based visual query, users may ask the computer to find similar images according to specified features such as color, texture, shape, motion, and spatiotemporal structures of image regions. They may even draw visual sketches to describe the images or videos they have in mind. And it will be useful if users are allowed to specify different weightings for different features

To extract these physical features, video sequences are first decomposed into separate shots. A video shot has a consistent background scene, while the foreground objects may occlude each other, disappear, or reappear. Video shot separation is achieved by scene change detection. Scene change may include abrupt scene change, transitional changes (e.g. dissolve, fade in or out, and wipe). Once the video is separated into basic segments, salient video regions and video objects are extracted. Video object is the fundamental level of indexing. Primitive regions are segmented according to color, texture, edge, or motion measures. As these regions are tracked over time, temporal attributes such as trajectory, motion pattern, and life span are indexed. These low-level regions may also be used to develop a higher level of indexing that includes links to conceptual abstractions of video objects.

In order to increase the resistant to transformation of objects in the video scene (e.g. rotation) which may not be captured at the low-level object segmentation, it is useful to considering some hints from users or content providers. For example in the new MPEG-4 standard, video scenes consist of video objects that can be maintained separately and they usually correspond to the semantic entities. Region segmentation and feature extraction can be applied to analyze these objects and obtain efficient indexes for search and retrieval.

Semantic Features

While the physical feature extraction is a powerful tools for user query, semantic features can lead to more intuitive and direct searching. Automatic classification of semantic concepts in multimedia data is a challenging topic recently [2].

The first approach is to use probabilistic graphic models to identify events, objects, and sites in multimedia signals by computing probabilities such as P(car AND road | segment of multimedia data). The basic idea is to estimate the parameters and structure of a model from a set of labeled multimedia training data.

A multimedia object has a semantic label and summarizes the time sequences of low-level features of multiple modalities in the form of a probability P(semantic label | multimedia sequence). Most multimedia objects fall into one of these categories: sites, objects, and events. Some of the multimedia objects are mainly featured in the video (e.g. a flower), and some are in the audio (e.g. a disco), and some are by both (e.g. a man).

The lifetime of a multimedia object is the duration of input that is used to determine its probability. In general, some multimedia objects (e.g. dinning) live longer that others (e.g. a gunshot). But we can limit all the lifetime of multimedia objects to the duration of the shot. This means we will not able to identify the existent of a multimedia objects across several scenes (e.g. an apple in consecutive scenes but in different direction), and it also ignores event sequences within a shot (e.g., gunshot after a dinner). It still give us a plenty of hints to do the inferences.

By using a hidden Markov Model (HMM), we can investigate what combinations of input features can give us a model of multimedia object [10]. For example, we may found that a color histogram of the input frames and a three-state HMM give a reasonably accurate video model of an explosion multimedia object [11]. Moreover, we can develop a higher level HMM between different multimedia objects instead of only investigating their video features. For example, an sky multimedia object may support the existing of a bird multimedia object, while it will decrease the probability of an existing of a shark multimedia object.

However, exact inference such as computing P(bird | multimedia data) is not feasible in terms of computing complexity, but works in applying approximate inference techniques using Markov chain Monte Carlo method [12] and variational techniques [13] are promising.

Speech Recognition

Speech Recognition is not a video feature to be indexed. But it is use to support the video segmentation, video object identification, and also generate the full-text script

for video which maybe indexed as text component.

Typically, an automatic speech recognition system will perform some analysis of the input sound wave to extract features found to be relevant in a similar process performed passively by the human ear. The system will be designed to recognize whole, isolated words and the features for each input sound wave will be compared with stored versions of a vocabulary to find a best match. The incoming sound wave will then be declared to have been recognized [14]. Stretches longer than single words can be recognized if treated as periods of sound analogous to single words to establish a match with stored templates themselves representing stretches of speech longer than words [15]. Typically such systems achieve around 95% accuracy, though this necessarily declines rapidly if the vocabulary is significantly increased.

Less often and with less success the incoming acoustic signal is segmented and an attempt made to match the segments with stored templates which may represent sub-phoneme or phoneme sized sounds or sometimes short strings of phonemes or partial phonemes. This approach has the merit of incorporating the idea that a large number of phrases can be recognized by the way in which a very small number of segments are combined. A knowledge base is required to determine from the sequence of recognized segments which word or phrase has been input.

Query Server

The Query Server accept user queries from Client Applications, construct formal queries that can be look up in the database of Indexing Server. Then it will gather the results, rank them and return the result set to the Client Application. The Query Server will be required to perform retrieval based on: Boolean Model, Vector Model and Probabilistic Model. And also the query attribute may or may not exist in a specified Indexing Server.

It is likely that it that up the act of “meta query server”, which may forward different kind of queries to different Indexing Server since a single Indexing Server may not able to provide the wide scope of information a user may want to search for. For example, when a query is about movies, than it will search the Indexing Server specialized in indexing movies; or when the query is about government information,

then it can search the government database for best result. In a sense, the Query Server is responsible to dispatch the user query, translate and map the query to the form a Indexing Server will accept, finally scores the results and return the results to the user.

While most functions of a Query Server is well-defined and also well developed, the remaining problem is the standardization of schemes and messaging. In fact, the emerging of MPEG-7 standard has the objective of specifying a standard set of descriptors as well as description schemes for the structure of descriptors and their relationships, which is quite the same as the medium for communication between Video Server, Indexing Server, and Query Server. By extending the set of attributes, we can include the user query in a simple and direct way.

Client Applications

For Client Application, the basic functionality of the client is quite the same (i.e. user query and video playback), but the add-on features may lead to very different application. For example, a medical application may link to a professional database, sharing medical video clips with references of pass records; an individual learning software may integrating video contents with other form of resources, and the assessment scheme as well; for a video editing system, it may even manipulate the video in many possible ways.

For a general purpose DVL Client, maybe there will not be much “new features”, but it will exploit the possibilities in the DVL System. For query, it may implement different query modes, either by keywords, by similar video, or even have a sketch board for user to draw what he want. For presentation it may rank the results in different ways: relevance, date, video quality, or popularity; it may also include thumbnail presentation, pop-up hints for brief description of each video, and it may has ability to link to the references associated with the video.

In fact, the area of usage is very diverse, and it is only limited by the creative imagination of people.

Related Works

Informedia

The Informedia Digital Video Library project is a research initiative at Carnegie Mellon University funded by the NSF, DARPA, NASA and others that studies how multimedia digital libraries can be established and used. The Informedia project has pioneered new approaches for automated video and audio indexing, navigaiton, visualization, search and retrieval and embedded them in a system for use in education, information and entertainment environments. Intelligent, automatic mechanisms are being developed to populate the library. Research in the areas of speech recognition, image understanding, and natural language processing supports the automatic preparation of diverse media for full-content and knowledge based search and retrieval.

MPEG-4

MPEG-4 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) to address the requirement on three fields: Digital television, Interactive graphics applications (synthetic content), and Interactive multimedia (World Wide Web, distribution of and access to content). MPEG-4 provides the standardized technological elements enabling the integration of the production; distribution and content access paradigms by providing standardized ways to:

- Represent units of aural, visual or audiovisual content, called "media objects". These media objects can be of natural or synthetic origin; this means they could be recorded with a camera or microphone, or generated with a computer;
- Describe the composition of these objects to create compound media objects that form audiovisual scenes;
- Multiplex and synchronize the data associated with media objects, so that they can be transported over network channels providing a QoS appropriate for the nature of the specific media objects; and
- Interact with the audiovisual scene generated at the receiver’s end.

MPEG-4 audiovisual scenes are composed of several media objects, organized in a hierarchical fashion. At the leaves of the hierarchy, we find primitive media objects such as: still images, video objects, audio objects and etc. that will be capable of representing both natural and synthetic content types. In addition, MPEG-4 defines the coded representation of objects such as text and graphics; talking synthetic heads and associated text used to synthesize the speech and animate the head; and also, synthetic sound as well.

More generally, MPEG-4 provides a standardized way to describe a scene, allowing for example to:

- Place media objects anywhere in a given coordinate system;
- Apply transforms to change the geometrical or acoustical appearance of a media object;
- Group primitive media objects in order to form compound media objects;
- Apply streamed data to media objects, in order to modify their attributes (e.g. a sound, a moving texture belonging to an object; animation parameters driving a synthetic face);
- Change, interactively, the user’s viewing and listening points anywhere in the scene.

The scene description is built on several concepts from the Virtual Reality Modeling language (VRML) in terms of both its structure and the functionality of object composition nodes and extends it to fully enable the aforementioned features.

MPEG-7

MPEG-7 is also an ISO/IEC standard developed by MPEG (Moving Picture Experts Group). While MPEG-4 provides the standardized technological elements enabling the integration of the production, distribution and content access paradigms of the fields of digital television, interactive graphics and interactive multimedia. MPEG-7, formally named "Multimedia Content Description Interface", aims to create a standard for describing the multimedia content data that will support some degree of interpretation of the information’s meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes shall support as broad a range of applications as possible.

Conclusion

In this paper, we have reviewed the system architecture of a Digital Video Library Network. We have also introduced different technologies used in the components, focusing on the multimedia indexing and retrieval.

We can see that, by the sharing and reuse of video content, and also the specialized modules which may serve professional uses, the power of Digital Video Library is increased by a large degree with the networked implementation.

Moreover, MPEG-4 and MPEG-7 is also addressing the indexing and retrieval of DVL. Although there is not many success implementation currently, the potential power of these two technologies can not be under estimated.

References

[1] R. Kahn and R. Wilensky. A framework for distributed digital object services. Technical Report cnru.dlib/tn95-01, CNRI, May 1995. http://www.cnri.reston.va.us/k-w.html.

[2] S.F. Chang, Q. Huang, A. Puri, B. Shahraray and T. Huang. Multimedia search and retrieval. Multimedia Systems, Standards, and Networks, MARCEL DEKKER, 2000.

[3] A. Hauptmann, M. Witbrock. Story segmentation and detection of commercials in broadcast news video. Proceedings of Advances in Digital Libraries Conference, Santa Barbara, April 1998.

[4] A. Merlino, D. Morey, D. Maybury. Broadcast news navigation using story segmentation. Proceedings of ACM Multimedia, November 1997.

[5] B. Shahraray, D. Gibbon. Efficient archiving and content-based retrieval of video information on the Web. AAAI Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora, Stanford, CA, March 1997, pp 133-136

[6] Q. Huang, Z. Liu, A. Rosenberg. Automated semantic structure reconstruction and representation generation for broadcast news. Proceedings SPIE, Storage and Retrieval for Image and Video Databases VII, San Jose, CA, January 1999.

[7] Nalin K. Sharda, Multimedia Information Networking, Prentice Hall, 1999

[8] A. Vogel et al., “Distributed Multimedia and QOS: A survey, “ IEEE Multimedia, vol. 2, no. 2, Summer 1995, pp. 10-19.

[9] B. Yates, R Neto, Modern Information Retrieval, Addison Wesley,

[10] A. Jaimes, F-F Chang. Model based image classification for content-based retrieval. SPIE Conference on Storage and Retrieval for Image and Video Database, San Jose, CA, January 1999.

[11] T.T. Kristjansson, B.J. Frey, T.S. Huang. Event-coupled higgen Markov models. Submitted to Adv Neural Inform Process Syst.

[12] G.E. Hinton, T.J. Sejnowski. Learning and relearning in Boltzmann machines. In: DE Rumelhart, and JL McClelland, eds. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. vol I, Cambridge, MA: MIT Press, 1986, pp 282-317.

[13] M.I. Jordab, Z. Ghahramani, T.S. Jaakkola, L.K. Saul. An introduction to variational methods for graphic models. In: MI Jordan, ed. Learning and Inference in Graphic Models. Norwell, MA: Kluwer Academic Publichers,

1988.

[14] Moore, R. K., Overview of speech input. In Proceedings of the 1st international conference on speech technology (J. N. Holmes, editor), pp. 25-38. Bedford: IFS (Publications) Ltd. Amsterdam: North Holland.

[15] Bridle, J. S., Brown, M. D. and Chamberlain, R. M., An algorithm for connected word recognition. In Automatic speech analysis and recognition (J.P. Haton, editor), pp.191-204. Dordrecht, Holland: D. Reidel.

[16] Stephen W. K. Fu, C. H. Lee, Orville L. Clubb, A Survey on Chinese Speech Recognition, 23 November, 1995.

[17] M. A. A. Tatham, An Integrated Knowledge Base for Speech Synthesis and Automatic Speech Recognition, Journal of Phonetics (1985) 13, pp. 175-188, Academic Press Inc. (London) Limited.