

## Chapter XIII

# A Framework for Indexing Personal Videoconference

Jiqiang Song, Chinese University of Hong Kong, Hong Kong

Michael R. Lyu, Chinese University of Hong Kong, Hong Kong

Jenq-Neng Hwang, University of Washington, USA

### ABSTRACT

*The rapid technical advance of multimedia communication has enabled more and more people to enjoy videoconferences. Traditionally, the personal videoconference is either not recorded or only recorded as ordinary audio and video files that only allow linear access. Moreover, in addition to video and audio channels, other videoconferencing channels, including text chat, file transfer, and whiteboard, also contain valuable information. Therefore, it is not convenient to search or recall the content of videoconference from the archives. However, there exists little research on the management and automatic indexing of personal videoconferences. The existing methods for video indexing, lecture indexing, and meeting support systems cannot be applied to personal videoconference straightforwardly. This chapter discusses important issues unique to personal videoconference and proposes a comprehensive framework for indexing personal videoconference. The framework consists of three modules: videoconference archive acquisition module, videoconference archive indexing module, and indexed videoconference accessing module. This chapter will elaborate on the design principles and implementation methodologies of each module, as well as the*

*intra- and inter-module data, and control flows. Finally, this chapter presents a subjective evaluation protocol for personal videoconference indexing.*

## INTRODUCTION

Videoconference is an advanced type of meeting approach that employs real-time video communication technology to enable participants in different geographical locations to see and talk to each other. In fact, today's videoconference employs richer communication media than audio/video, such as text chat, file transfer, whiteboard, and shared applications. Therefore, it should be more precisely called a "multimedia conference." By convention, we still use the term "videoconference" in this chapter. A few years ago, videoconference was only an expensive option for big companies' business operations due to the requirement of special hardware and communication lines. With the growth of Internet bandwidth and the development of multimedia communication technologies in past years, videoconference has become more and more popular in business operations (Sprey, 1997). Furthermore, with affordable video and audio capture devices, the advanced low bit-rate coding, and pure-software videoconferencing tools, home users can also enjoy visual communications at 56Kbps or lower (Deshpande & Hwang, 2001). For example, video-based distance learning has benefited significantly from videoconferencing techniques.

A videoconference can be classified as either *personal videoconference* or *group videoconference* based on the number of participants at each geographical location. A videoconference is classified as a personal videoconference on the condition that there is only one participant at each location; otherwise, it is a group videoconference. A personal videoconference is usually held among several participants, and each participant sits in front of a computer equipped with a camera, a microphone, and a speaker (or earphone). A group videoconference is often held in a multimedia conference room, where more than one camera and microphones are installed. Existing research on videoconference indexing all focuses on group videoconferences, leading to meeting support systems. Since personal videoconferences are becoming more and more popular recently and the characteristics of personal videoconferences are different from that of group videoconferences, this chapter aims to propose a framework for indexing personal videoconferences.

A participant of a personal videoconference usually wishes to save the content of the conference for the later reference. However, current videoconferencing systems provide little support on this aspect. Even if the streaming video and audio information can be recorded as ordinary video and audio files, these files occupy too much space and do not support non-linear access, so it is not easy to recall the details of a videoconference without watching it again. Therefore, it is very difficult to manage and search those videoconference records, revealing the timely importance of the research on indexing personal videoconferences.

The rest of this chapter is organized as follows. The next section reviews the background of multimedia indexing. Then, we discuss the characteristics of personal videoconference indexing, followed by a proposed comprehensive framework for building a personal videoconference archive indexing system. A subjective evaluation protocol is then presented in the next section. Finally, we draw our conclusions.

## BACKGROUND

When talking about videoconference archive indexing, one cannot ignore what has been researched in general video indexing because video — containing both visual and aural information — is the major medium during a videoconference. Before the wide deployment of videoconferencing, research on general video indexing had been conducted for several years (Madrane & Goldberg, 1994; Sawhney, 1993). Since the content of a video clip is hidden in its visual and aural information, which is not directly searchable like text, the main purpose of video indexing is to support content-based video retrieval (CBVR). CBVR is more complicated than the content-based image retrieval (CBIR) since a video clip is a time-varying image/audio sequence (Zhang, Wang, & Altunbasak, 1997). Therefore, the temporal semantic structure of a video clip also implies the video content. The success of video indexing research, not yet accomplished, will result in a digital video library featuring full-content and knowledge-based search and retrieval. A well-known effort in this field is the Informedia project (Wactlar Kanade, Smith, & Stevens, 1996) undertaken at Carnegie Mellon University.

Video indexing has been studied in multifarious ways, producing a lot of useful techniques for multimedia information retrieval and indexing. Early research mostly draws on single or dual modality of video content analysis. Liang, Venkatesh, and Kieronsak (1995) focused on the spatial representation of visual objects. Ardizzone, La Cascia, Di Gesu, & Valenti (1996) utilized both global visual features (e.g., color, texture, and motion) and local visual features (like object shape) to support the content-based retrieval. Chang, Zeng, Kamel, and Alonso (1996) integrated both image and speech analyses for news or sports video indexing. Ardizzone and La Cascia (1996), Wei, Li, and Gertner (1999), Zeng, Gao, and Zhao (2002), and Hsu and Teng (2002) proposed that motion is also an important clue for video indexing. To take advantage of video encoding features, some video indexing work is performed in the compressed domain (Chang, 1995). Later, multimodal information was explored for video indexing (Hauptmann & Wactlar, 1997). Li, Mohan, and Smith (1998) and Lebourgeois, Jolion, and Awart (1998) presented multimedia content descriptions of video indexation. Abundant multimodal video indexing methods have also been proposed (Albiol, Torres, & Delp, 2002; Lyu, Yau, & Sze, 2002; Tsekeridou & Pitas, 1998; Viswanathan, Beigi, Tritschler, & Maali, 2000). They extracted information in many aspects, including face, speaker, speech, keyword, etc.

On the other hand, high-level indexing aided by domain-specific knowledge has also attracted much interest. Dagtas, Al-Khatib, Ghafoor, and Khokhar (1999) proposed a trail-based model utilizing object motion information. Hidalgo and Salembier (2001) segmented and represented foreground key regions in video sequences. Maziere, Chassaing, Garrido, and Salembier (2000) and Hwang and Luo (2002) conducted object-based video analysis. Ben-Arie, Pandit, and Rajaram (2001) and Wang, Ma, Zhang, and Yang (2003) performed view-based human activity recognition and people similarity-based indexing, respectively.

Recently, researchers focus on understanding the semantic structure of video sequences, that is, story, scene, and shot. Segmenting video into shots is a fundamental task for this purpose. Segmentation is the partitioning of continuous media into homogenous segments. There exist many ways for shot segmentation, for example, by camera motion (Corridoni & Del Bimbo, 1996), using the human face (Chan, Lin, Tan, &

Kung, 1996), analyzing image basic features (Di Lecce et al., 1999), and counting the percentage of moving pixels against the background (Kang & Mersereau, 2002). It is also important to detect gradual shot transitions (Bescos, Menendez, Cisneros, Cabrera, & Martinez, 2000). The segmented shots are annotated for subsequent searches (Ito, Sato, & Fukumura, 2000; Wilcox & Boreczky, 1998) and for automatically summarizing the title (Jin & Hauptmann, 2001). Based on the semantic structure analysis, various semantic indexing models or schemes have been proposed (Del Bimbo, 2000; Gao, Ko, & De Silva, 2000; Gargi, Antani, & Kasturi, 1998; Iyengar, Nock, Neti, & Franz, 2002; Jasinschi et al., 2001a; Jasinschi et al., 2002; Luo & Hwang, 2003; Naphade & Huang, 2000; Naphade, Kristjansson, Frey, & Huang, 1998). Most of them employ probabilistic models, for example, Hidden Markov Models (HMMs) and/or Dynamic Bayesian networks (DBNs), to represent the video structure.

The extensive research on video indexing has produced the following valuable techniques to analyze the video and audio archives of a videoconference.

- *Key Frame Selection.* One basic means to remove the redundancy of video sequences is to represent them by key frames. Some methods work in the MPEG compressed domain (Calic & Lzquierdo, 2002; Kang, Kim, & Choi, 1999; Tse, Wei, & Panchanathan, 1995), whereas others work in the uncompressed domain (Diklic, Petkovic, & Danielson, 1998; Doulamis, Doulamis, Avrithis, & Kollias, 1998; Kim & Park, 2000).
- *Face Detection and Recognition.* Since a videoconference is always human-centric, human faces are principal objects in the video. Sato and Kanade (1997) associated human faces with corresponding closed-caption text to name the faces. Arika, Suiyama, and Ishikawa (1998) indexed video by recognizing and tracking faces. Gu and Bone (1999) detected faces by spotting skin color regions. Tsapatsoulis, Avrithis, and Kollias, (2000) and Mikolajczyk, Choudhury, and Schmid (2001) proposed temporal face detection approaches for video sequences. Eickeler, Walhoff, Lurgel, and Rigoll (2001) and Acosta, Torres, Albiol, and Delp (2002) described content-based video indexing system using both face detection and face recognition.
- *Speech Recognition.* Speech of a videoconference contains the most information. In fact, there are many video-indexing methods only focusing on the audio track. Hauptmann (1995) analyzed the uses and limitations of speech recognition in video indexing. The audio track can also be segmented for acoustic indexing (Young, Brown, Foote, Jones, & Sparck-Jones, 1997). Barras, Lamel, and Gauvain (2001) proposed an approach to obtain the transcript of the compressed audio.
- *Speaker Identification.* Knowing who is speaking greatly enhances the videoconference indexing. Speaker identification can be performed in video or audio only, or using audio and video correlation (Cutler & Davis, 2000; Nam, Cetin, & Tewfik, 1997; Wilcox, Chen, Kimber, & Balasubramanian, 1994).
- *Keyword Spotting.* Other than recognizing every word in the speech to obtain the transcript of the speech for later textual analysis, one can use a domain-specific keyword collection to spot keywords in the speech for the indexing purpose (Dharanipragada & Roukos, 1996; Gelin & Wellekens, 1996).

- *Video Text Detection.* Since plain text is no doubt the most convenient medium for indexing and searching, one may try to extract as much as possible textual information from audio and video. Thus, video text detection has always been emphasized (Cai, Song, & Lyu, 2002; Hua, Yin, & Zhang, 2002; Jain & Yu, 1998; Kim, Kim, Jung, & Kim, 2000; Li & Doermann, 1998).
- *Topic Classification.* Locating the topic-switching point and summarizing the topic for a conference session are critical to deliver the accurate, both in time and topic, results to content-based searches. Jasinschi, Dimitrova, McGee, Agnihotri, and Zimmerman (2001b) integrate multimedia processing to segment and classify topics. Ngo, Pong, and Wang (2002) detect slide transitions for topic indexing

In addition to the research on video indexing, other research related to videoconference archive indexing includes lecture indexing systems and meeting support systems. Lecture indexing focuses on two points: (1) how to collect media archives of attractive lectures automatically, and (2) how to access these recorded archives. For the first point, the problem is subdivided into “what should be captured?” and “how it should be captured?” Since a lecture involves one or more speakers and an audience, who may interact with each other from time to time, both speakers and audience should be captured. Besides the verbal cues, non-verbal communication cues, e.g., body posture, gestures, and facial expressions, also play an important role during the interaction (Ju, Black, Minneman, & Kimber, 1997). Slides and handwritings on the whiteboard cannot be missed either. To capture the multimodal information, one or more cameras and microphones are necessary to construct an intelligent lecture room (Joukov & Chiueh, 2003; Kameda, Nishiguchi, & Minoh, 2003; Rogina & Schaaf, 2002; Stewart, Wolf, & Heminje, 2003). Kameda et al. (2003) even utilized ultrasonic equipment to capture the movement trajectory of the speaker. One principle of the capturing process is to keep it minimally intrusive. For the second point, these systems can be divided into two groups by providing static functionality or dynamic functionality, according to Stewart et al. (2003). The systems with static functionality assume that the media archives, once captured and produced, become frozen assets for later playback use. On the other hand, dynamic systems provide some extended functionality so the media assets can be edited and reused in addition to presentation.

Meeting support systems focus on real-life meetings or group videoconferences. These systems develop special techniques or devices to free participants from paper-based note-taking, to record meeting activities, and to provide post-meeting information sharing. These systems also follow the principle of “minimally intrusive” to record meetings. Chiu, Kapuskar, Reitmeier, and Wilcox (2000) described the multimedia conference room in FX Palo Alto Laboratory, which is well equipped with three room cameras, one videoconference camera, one document camera, ceiling microphones, rear projector screen, whiteboard, and wireless pen computers. The indices of a captured meeting can be classified into *direct indices* and *derived indices*. Direct indices are created online during the meeting capturing, whereas derived indices require further, usually offline, analyses.

Direct indices include meeting activities and participants’ comments. Ginsberg and Ahuja (1995) explored various ways to visualize meeting activities, such as joining or leaving a meeting, using whiteboard, and more. There exist many tools to record

participants' comments. NoteLook (Chiu et al., 1999) allows participants to take handwritten notes for indexing audio and video. LiteMinutes (Chiu, Boreczsky, Girgensohn, & Kimber, 2001) supports typing notes or minutes on a laptop, where each line of text acts as an index. TeamSpace (Geyer, Richter, & Abowd, 2003) emphasizes tracing those domain-specific artifacts that connect several meetings.

Creating derived indices focuses mostly on audio, video, and slides. There are many kinds of analyses that can be done on audio streams, for example, speech recognition, speaker identification, keyword spotting, and interaction pattern classification. The Jabber system (Kazman, Al-Halimi, Hunt, & Mantei, 1996; Kazman & Kominek, 1999) proposes four paradigms for indexing videoconferences by audio. A speech recognizer is used to obtain a partial transcript of the meeting, and topics or themes are identified using lexical chaining. In addition, participants' interaction patterns, for example, discussion or presentation, are classified. Kristjansson, Huang, Ramesh, and Juang (1999) also described a unified framework for indexing and gisting spoken interactions of people. Foote, Boreczsky, and Wilcox (1999) introduced an approach to find presentations in recorded meetings by correlating the slide intervals in video streams and the speaker-spotting results in audio streams. Gross et al. (2000) developed a system that does speaker identification, speech recognition, action item recognition, and auto summarization. The eMeeting system (Leung, Chen, Hendriks, Wang, & Sahe, 2002) provides a slide-centric recording and indexing system for interactive meetings.

The review of literature indicates that personal videoconference indexing has seldom been addressed, except for our preliminary study (Song, Lyu, Hwang, & Cai, 2003). Since personal videoconferencing is booming and its indexing bears different characteristics from existing video library indexing systems, lecture indexing systems, and meeting support systems, this chapter will propose a comprehensive framework for building a Personal Videoconference Archive Indexing System (PVAIS).

## CHARACTERISTICS OF A PVAIS

This section describes the characteristics of a PVAIS in three aspects: archive acquisition, archive indexing, and indexed archive accessing and presenting.

In the archive acquisition process, the principle of "minimally intrusive" applies. The scenario of personal videoconferencing is that each participant sits in front of a computer, using pure-software videoconferencing client to communicate with each other. Different from video indexing, where only audio and video are available, a PVAIS considers six communication channels, including five medium channels and one conference control channel. The five medium channels are the audio channel, video channel, text chat channel, file transfer channel, and whiteboard channel. The conference control channel contains member information and conference coordination information. Different from lecture indexing and group videoconference indexing, where additional capturing equipment is used, a PVAIS is restricted to software-based capturing because there is no room for additional equipment between a participant and his/her computer.

The archive indexing process of a PVAIS should also be transparent to the user. Since a personal videoconference is not as well-organized as a broadcasting video program, the semantic video structure is not emphasized in a PVAIS. Instead, a PVAIS focuses on the conference events from the user's viewpoint. Such conference events

include both media events (e.g., speaker changed, topic switched, text transmitted, whiteboard updated, and file transferred) and control events (e.g., member joined or left, channel created or closed). In addition to these conference events, a PVAIS also emphasizes the multimodal derived indices to provide content-based searches. As a personal indexing system, a PVAIS can automatically create and maintain a contact list for the user.

The access to the indexed videoconference archives is restricted to the authenticated users only. A PVAIS provides an interface for the user to search and review the indexed videoconference archives. The user also needs to manage and edit the indexed items via this interface. The user can conduct searches with various criteria based on content of interest or videoconferencing events through the interface. A PVAIS supports synchronized presentation of multimedia searching results. During the presentation, a PVAIS allows the user to pause at any time and add annotations or bookmarks.

## FRAMEWORK FOR A PVAIS

Figure 1 shows the top-level view of the framework of a PVAIS, which consists of three separate processing modules with a linear processing order:

- *Videoconference Archive Acquisition Module.* This module works together with the personal videoconferencing terminal to extract and save the raw content archives from all communication channels *in real time*. This module can be either embedded in the terminal or separated from it.
- *Videoconference Archive Indexing Module.* This module works *offline* after the videoconference finishes. It takes the raw videoconference archives as input, analyzes the videoconferencing events, integrates information from multiple channels to produce derived indices, and finally outputs the indexed videoconference archives.
- *Indexed Videoconference Accessing Module.* This module serves as the interface to users. It provides the functions of managing indexed videoconferences, searching for content of interest among them, and presenting the selected videoconference.

This section further elaborates the design principles and implementation methodologies for each module. First, we briefly introduce the knowledge of videoconferencing system. Generally, a videoconferencing system could be implemented in various archi-

Figure 1. Top-level view of the framework of a PVAIS

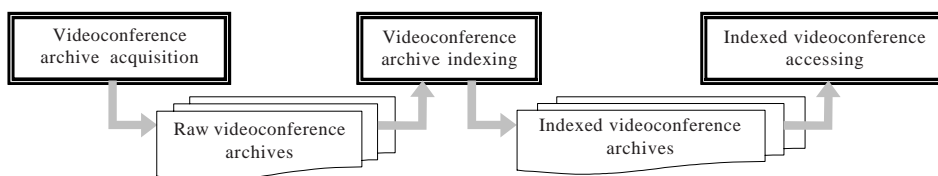
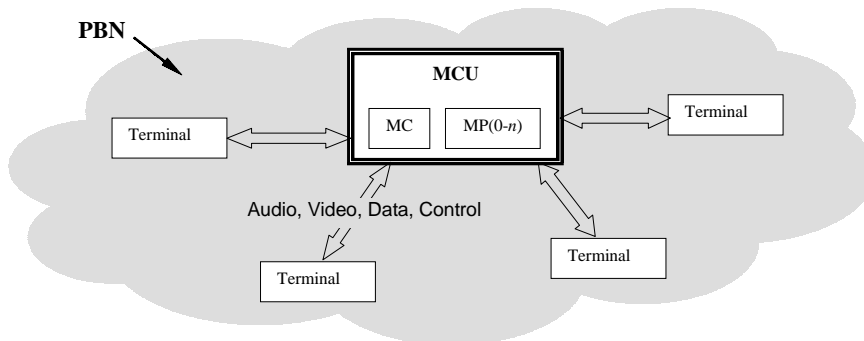


Figure 2. A centralized architecture of H.323 videoconference system



tures, such as the Client/Server architecture for intranet-based videoconferencing and the H.323 architecture for Internet-based videoconferencing. To achieve the best compatibility, this framework should be designed to cooperate with the videoconferencing systems compliant to the most widely adopted ITU-T H.323 Recommendation (ITU-T, 2001), which is quite comprehensive for multimedia communication systems. Nevertheless, this framework can be easily tailored to fit other videoconferencing architectures.

Figure 2 illustrates a typical centralized architecture of H.323 videoconferencing system over a Packet Based Network (PBN). It consists of one Multipoint Control Unit (MCU) and several (at least two) *Terminals*. A terminal is a videoconferencing client in one location. Participants of a videoconference communicate with each other using their local terminals. An MCU contains one Multipoint Controller (MC) and  $n$  ( $n \geq 0$ ) Multipoint Processors (MPs). The responsibility of an MC is to coordinate the videoconference, while that of an MP is to process audio/video streams when necessary, such as video switching and audio mixing. The communications between a terminal and the MCU may include audio, video, data, and control channels. A PVAIS, particularly the videoconference archive acquisition module, is only concerned with the terminal.

## Videoconference Archive Acquisition Module

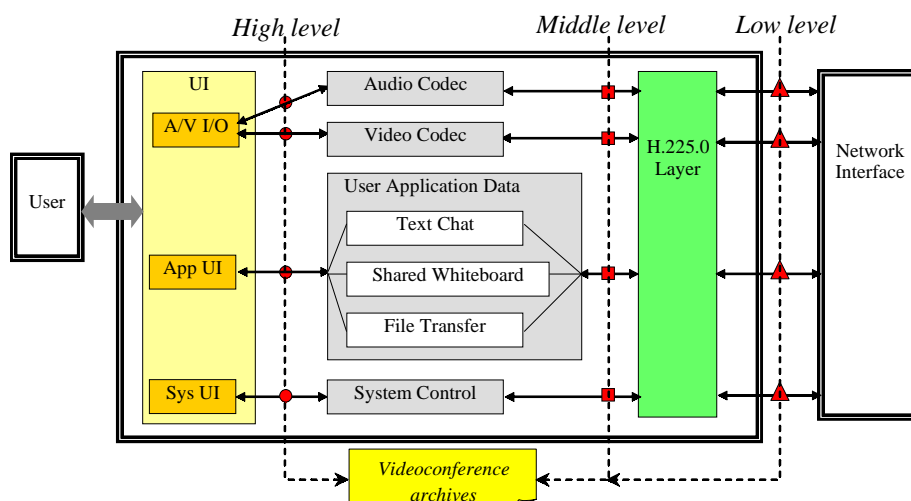
Since a personal videoconference terminal is pure software with only simple media capture and playback devices attached (like WebCam, microphone, and speaker), the videoconference archive acquisition module should also be pure software to be minimally intrusive. Thus, a thorough understanding of the structure of terminal is necessary.

As shown in Figure 3, the central double-bordered rectangle encloses the elements of a videoconferencing client, which consists of a User Interface (UI) and an H.323 terminal model. The UI part provides the interface for audio/video capture/playback equipments, for user applications, and for system controls. The H.323 terminal model contains audio/video codecs, user data path, system control unit, and H.225.0 layer (ITU-T, 2003a), which is for media stream packetization and synchronization.

There are four types of communication channels: audio, video, data, and control. The audio and video channels transmit incoming/outgoing video and audio information, respectively. The data channel carries information streams for user applications, such



Figure 3. Terminal structure and three interception levels



as text chat, shared whiteboard, and file transfer. The control channel transmits system controls information, including H.245 control (ITU-T, 2003b), call control, and Registration, Admission, and Status (RAS) control. Note that the transportation protocols for these channels are different. Since real-time audio and video transmission are extremely sensitive to delays and jitters, but insensitive to the occasional loss of one or two packets, the reliable Transportation Control Protocol (TCP) is not suitable to transmit audio and video due to the delays introduced during the connection-setup routine and the acknowledgment routines. Therefore, User Datagram Protocol (UDP) together with Real-time Transportation Protocol/Real-time Transportation Control Protocol (RTP/RTCP) is used for audio and video channels. In contrast to audio and video, data and control information needs very reliable, accurate transmission, but they are not sensitive to a few delays. Thus, TCP is most suitable for data and control channels.

## Where to Extract the Information

“Try to introduce the *least* delay into the terminal” is an important principle for extracting videoconference content in real time. This principle implies that the extraction process should not be too complex. This section describes three levels of extraction methods of the videoconference archive acquisition module: high-level, middle-level and low-level, as shown in Figure 3. Both the high-level extraction and the middle-level extraction are embedded in the videoconferencing client, whereas the low-level extraction is separated from the client.

The high-level extraction takes place between the UI and the information codecs. The interception points are marked with disks in Figure 3. For the audio channel and the video channel, the extraction will get uncompressed information. For the data channel and the control channel, the extraction can get the semantic operations from the UI directly.

The middle-level extraction is situated before the H.225.0 module — that is, before the information is packetized. The interception points are marked with squares in Figure 3. For the audio channel and the video channel, the extraction can utilize the features of codecs to obtain low-redundancy information, for example, taking I-frames from H.263-encoded video streams as candidate key frames. For the data channel and the control channel, the extraction should be aware of the structure of message stacks to retrieve the information.

The low-level extraction is separated from the videoconferencing client, situated before the network interface. The interception points are marked with triangles in Figure 3. The extraction process runs as a Daemon monitoring the IP transportation ports of the computer. When the communication of the videoconferencing client is detected, the extraction will unpacketize the bitstream to retrieve the information.

Comparing the three levels of extraction methods, we realize that from high-level extraction to low-level extraction, the implementation complexity increases dramatically. However, high complexity does not guarantee the commensurate enhancement to the efficiency. Therefore, to minimize the complexity of the videoconference archive acquisition module, it is recommended that the high-level or middle-level implementation methods is chosen if the existing videoconferencing client could be modified to or be replaced with an information-acquisition-enabled client.

## How to Extract the Information

There are two important principles for storing videoconference content into archives:

- Try to reduce the information redundancy as much as possible.
- All recorded events must be labeled with a timestamp.

To index the content of a videoconference, the information in all the four channels should be extracted and stored. These channels could be further divided into logical channels according to the user's point of view, as follows: *video\_in*, *video\_out*, *audio\_in*, *audio\_out*, *text\_chat*, *whiteboard*, *file\_in*, *file\_out*, and *control*. Some useful events in each channel are defined in Table 1.

Most of the above events do not take much time to detect except some events in video channels. Since video analysis is usually not fast enough, those events demanding complex analysis (e.g., face detection, slide classification, gesture, and head motion detection) should not be detected in real time. Usually, only the scene change event, resulting in key frames, is detected in real time. Other events will be recovered in the offline indexing by analyzing key frames.

Figure 4 shows the paradigm of storing the extracted information in each channel into the corresponding archive. The extraction processes of all logic channels begin simultaneously when the videoconference starts. The operations in each logic channel are depicted as follows.

The extraction processes for the *Video\_in* channel and those for the *Video\_out* channel are similar. The *scene change* events are detected in real time. Let  $f(t)$  denote the function of the video content feature, which varies with time  $t$ . Thus, the changes of video content will be detected in  $f'(t)$  as peaks, and the valleys right after each peak can be

Table 1. Event definition for logic channels

Logic Channel	Events
<i>video_in</i>	Scene change, face appeared, slide appeared, gesture made, head movement
<i>video_out</i>	Scene change, face appeared, slide appeared, gesture made, head movement
<i>audio_in</i>	Speech started, silence started, speaker changed
<i>audio_out</i>	Speech started, silence started
<i>text_chat</i>	Text sent, text received, chat type changed
<i>whiteboard</i>	Whiteboard updated
<i>file_in</i>	File sent
<i>file_out</i>	File received
<i>control</i>	Member joined, member left, channel created, channel closed

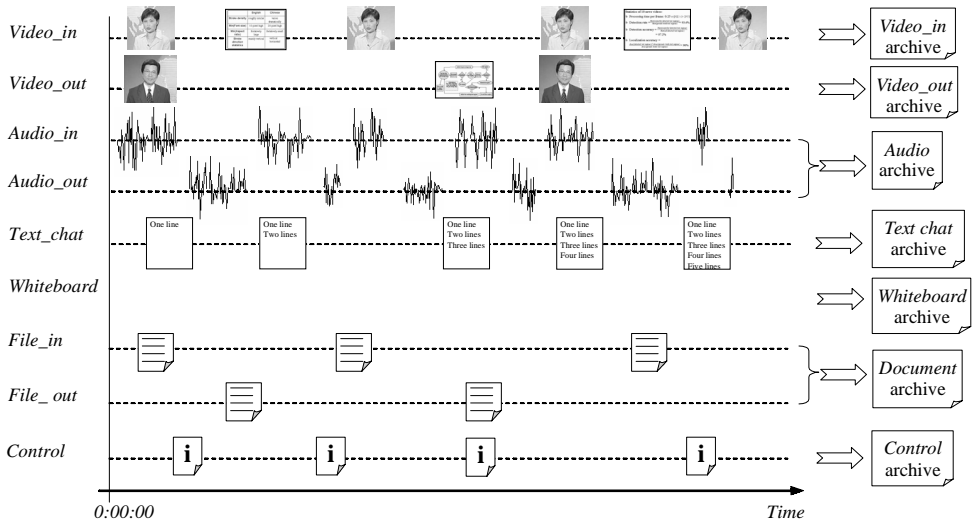
selected as key frames. Note that  $f(t)$  should consider not only the statistic distribution, but also the spatial distribution of colors to discriminate the change between slides, such as the definition in Dirfaux (2000). The resulting archive — the *Video\_in* archive or *Video\_out* archive — consists of one text-based index file and a number of key-frame pictures. The index file records the timestamp of each event and the location of the corresponding key frame picture. To preserve the details of the slide pictures, we should employ lossless compression methods, such as TIFF, to store the key-frame pictures.

The audio streams in the *Audio\_in* channel and the *Audio\_out* channel are first mixed into one stream. Then, silence detection is applied to the mixed stream. Thus, only the speech segments will be stored. The *speaker changed* events are detected in real time by comparing the current vocal feature with the last one. Usually, the vocal feature is modeled by a Gaussian Mixture Model (GMM). The *Audio* archive also includes a text-based index file that records the timestamp for each event and the location of the corresponding audio segment.

The *Text\_chat* archive is just a text file containing the timestamp and the corresponding information of each event. For example, for a *text sent* event, the corresponding information includes the sender's username and the textual content. For a *chat type changed* event, the corresponding information is the new chat type, that is, either public chat or private chat.

The *Whiteboard* archive consists of a text-based index file and a number of whiteboard snapshot pictures. The index file records the timestamp for each *whiteboard updated* event and the location of the corresponding snapshot pictures. Since the *Whiteboard* channel contains handwritten texts and graphics, the update of this channel happens not at a point in time but in a period of time. To detect when the update begins and finishes, Song et al. (2003) proposed a method to monitor the *Whiteboard* by comparing sampling images. One can also obtain this information by monitoring the data transfer in this channel. The whiteboard snapshots can be saved as grayscale/binary images with lossless compression methods, such as TIFF/JBIG.

Figure 4. Paradigm of storing the videoconference archives



On each file exchange, the extraction process will copy the sent/received files to the directory storing the videoconference archives. The file exchange information in the *File\_in* channel and the *File\_out* channel is stored in one *Document* archive, which consists of a text-based index file and a number of exchanged files. The index file records the timestamp and the corresponding information of each event. For instance, for the *file sent* event, the corresponding information includes the recipient's user name and the location of the copy.

The *Control* archive is a text file containing the timestamp and the corresponding information of each event. For a *member joined/left* event, the corresponding information is the involved member's user name. For a *channel created/closed* event, the corresponding information is the involved channel type.

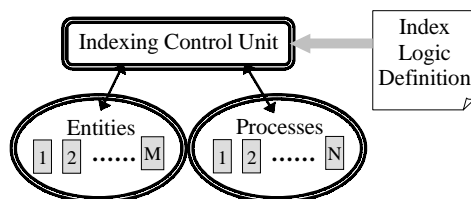
## Videoconference Archive Indexing Module

The videoconference archive indexing module is automatically started after the videoconference finishes. This module should be transparent to users.

## LEP Indexing Architecture

Since this module integrates many content indexing functionalities, its architecture should be carefully designed to ensure the development flexibility, the runtime stability, and the maintenance convenience. Therefore, the framework of a PVAIS employs an Logic of Entity and Process (LEP) indexing architecture, as shown in Figure 5, which consists of one *Indexing Control Unit*, one *Indexing Logic Definition*, 1~M *Entities*, and 1~N *Processes*.

Figure 5. LEP indexing architecture



*Entities* include raw videoconference archives produced by the videoconference archive acquisition module and indexed videoconference archives generated by performing indexing functions on raw videoconference archives.

A *Process* implements a content indexing function that takes one or more raw videoconference archives and/or indexed videoconference archives as input and outputs a new indexed archive or updates an existing indexed archive. A process is a stand-alone executable file, separating from the main control and other processes.

The *Indexing Logic Definition* (ILD) describes the function of each process and specifies the input entities and output entities of each process. Therefore, the relationship among all processes and entities — the indexing logic — is defined. The indexing logic determines the priority of each process, that is, the process generating the input entities of the current process should be accomplished before starting the current process. The ILD also defines the format of the resulting Extensible Markup Language (XML) index file.

The *Indexing Control Unit* (ICU) plays the role of main controller in the LEP indexing architecture. It reads the indexing logic from the ILD, checks the availability of each entity, coordinates the execution of each process, and finally generates the resulting XML index file. ICU is responsible for making the whole videoconference archive indexing module work smoothly and automatically. Once an entity become available, it should register to the ICU. When all the input entities of a process become available, the ICU will create a thread to start the process. When a process finishes, no matter whether it succeeds or fails, it will notify the ICU. Moreover, the ICU should check the status of every process from time to time to detect any abrupt termination caused by unexpected exceptions.

Since the ICU and the processes are separately executable files that are loosely coupled by the ILD in the LEP architecture, the *robustness*, *extensibility*, and *maintainability* of the videoconference archive indexing module are greatly enhanced. Therefore, adding, removing, or updating a process only affects the involved process and the ILD.

- *Robustness*: The failure of one function will not lead to the failure of the whole indexing module.
- *Extensibility*: To add new functions implemented as stand-alone processes, one only needs to edit the ILD to link the new processes in.
- *Maintainability*: To merely upgrade one function, just replace the involved process with the upgraded version, without updating the whole indexing module.

Only when the requirement of input entities has been changed is it necessary to edit the ILD. To remove one function, simply modify the ILD to drop this function.

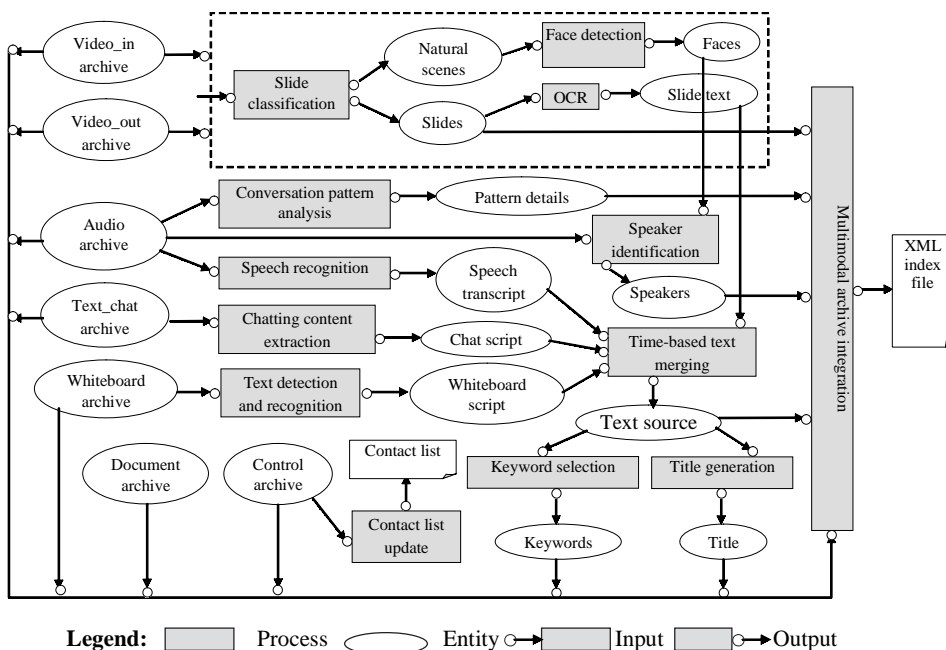
## Indexing Logic Definition of a PVAIS

How to define the indexing logic is the core of the videoconference archive indexing module. The ILD of a PVAIS is illustrated in Figure 6, where entities are drawn as white ellipses and processes as gray rectangles. The ILD defines twelve processes totally, which are described in the processing order as follows.

At the very beginning, only the seven raw videoconference archives are available (see Section 4.1.2). There are six processes whose input entities are ready. They are slide classification, conversation pattern analysis, speech recognition, chatting content extraction, text detection and recognition, and contact list update. These five processes can be started simultaneously.

Note that the *Video\_in* archive and the *Video\_out* archive will go through the same processes enclosed in the top dashed rectangle separately. The slide classification process takes a video archive (either *Video\_in* archive or *Video\_out* archive) as input, and classifies the key frames in the video archive into two classes: nature scenes and slides. Foote et al. (1999) detected slides by first downsizing the key frame to a 64×64 grayscale representation, then performing the DCT transform on the downsized frame, and finally feeding the 100 principle transform coefficients to a trained diagonal-covariance Gaussian model for the classification. For a simpler implementation, the slide classification can also be accomplished by analyzing both the maximum peak of color

Figure 6. Index logic definition of a PVAIS



histogram and the absolute difference in entropy between horizontal lines in a key frame picture (Leung et al., 2002).

The conversation pattern analysis process takes the *Audio* archive as input, analyzes speeches, silence and *speaker changed* events, and finally divides the whole timeline into segments according to conversation patterns, for example, presentation, discussion, and argument. Kazman et al. (1996) identify three salient measures to classify conversation patterns: (1) who is speaking and for how long, (2) the length of pause, and (3) the degree of overlap between speakers.

The speech recognition process also takes the *Audio* archive as input and produces the *Speech transcript* archive, which composes the majority of text source. Every word in the speech transcript is time-stamped. Since the accuracy of the speech recognizer will determine whether the speech transcript is useful or useless, this process is critical to the system performance. The speech recognizer can be implemented in two ways: (1) employ or improve existing algorithms for large-vocabulary, speaker-independent speech recognition (Barras et al., 2001; Hauptmann, 1995), or (2) utilize commercial speech recognition engines, such as IBM ViaVoice® and Microsoft SpeechAPI®.

The chatting content extraction process extracts the chatting texts and the corresponding time-stamps from the *Text\_chat* archive to yield the *Chat script* archive. This process can be easily implemented since the structure of the *Text\_chat* archive is known.

The text detection and recognition process takes the *Whiteboard* archive as input, detects text from the whiteboard snapshot pictures, recognizes the text (if any), and stores the recognized text and the corresponding time-stamps into the *Whiteboard script* archive. Since the whiteboard snapshot may contain both graphics and text, the first step is to segment text from graphics. This can be done by a connected-component based method (Fletcher & Kasturi, 1988). Then, an off-line, handwritten text-recognition method (Vinciarelli, Bengio, & Bunke, 2003) is applied to the segmented text image to obtain the text.

The contact list update process checks the *Member joined* events in the *Control* archive. If a member has not yet been recorded in the contact list, this process will add this member into the contact list. Note that the contact list is a global archive corresponding to the whole PVAIS, not belonging to a specific videoconference.

After the slide classification process finishes, its output entities (i.e., natural scenes and slides) become available. Then, the face detection process and the Optical Character Recognition (OCR) process can be started.

The face detection process takes the natural scenes as input and detects faces from them. The detected face regions and their corresponding time-stamps are stored as a new *faces* archive. When the *faces* archive becomes available, the speaker identification process can be started to identify speakers in every time period by integrating aural and visual information (Cutler & Davis, 2000). The implementation of these two processes involves face detection and recognition techniques, which are easily attainable in the literature. Eickeler et al. (2001) detect faces by Neural Network and then recognize faces using pseudo-2-D HMMs and a k-means clustering algorithm. Acosta et al. (2002) proposed a face detection scheme based on a skin detection approach followed by segmentation and region grouping. Their face recognition scheme is based on Principal Component Analysis (PCA). To diminish the limitation that any single face recognition algorithm cannot handle various cases well, Tang, Lyu, and King (2003) presented a face

recognition committee machine that assembles the outputs of various face recognition algorithms to obtain a unified decision with improved accuracy.

The OCR process takes the *Slides* archive as input, binarizes the slide pictures, and recognizes text from them. The recognized text is stored as a *Slide text* archive. The major task of this process is to identify text regions in the slide pictures. Cai et al. (2002) proposed a robust approach to detect and localize text on complex background in video images. This algorithm utilizes invariant edge features to detect and enhance text areas, and then localizes text strings with various spatial layouts by a coarse-to-fine localization scheme.

Once the *Speech transcript* archive, the *Chat script* archive, the *Whiteboard script* archive, and the *Slide text* archive are all ready, the time-based text merging process will be started to merge these four archives into one *Text source* archive according to their associated time-stamp information. Therefore, the *Text source* archive integrates all the textual information obtainable from videoconference archives. When the *Text source* archive becomes available, the keyword selection process and the title generation process will be started.

The keyword selection process takes the *Text source* archive as input and produces keywords on two levels: global and local. The global keywords are selected in the scope of the whole videoconference, representing the overall subject of the conference. On the other hand, the local keywords are clustered in a limited time period; therefore, they only indicate the topic of this period. Providing two-level keywords enhances the flexibility in supporting the content-based retrieval. Global keywords enable the quick response when searching videoconferences, while local keywords are more powerful in seeking a point of interest in a videoconference. The Neural Network-based text clustering algorithm (Lagus & Kaski, 1999) can be employed to select both global keywords and local keywords.

The title generation process also takes the *Text source* archive as input and generates a title for this videoconference by employing language-processing techniques. Jin and Hauptmann (2001) proposed a novel approach for title word selection. They treat this task as a variant of an Information Retrieval problem. A good representation vector for title words is determined by minimizing the difference between human-assigned titles and machine-generated titles over the training examples.

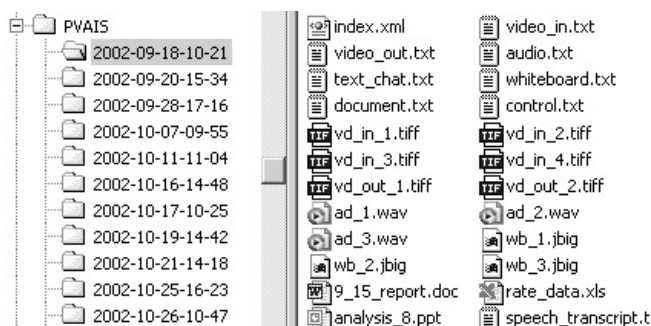
In addition to employing existing algorithms, one may also need to develop some specific multimedia processing techniques. In this case, one can find the fundamentals of digital image processing, digital video processing, and speech analysis in the books of Castleman (1996), Tekalp (1995), and Rabiner and Juang (1993), respectively.

Finally, when all the entities are ready, the multimodal archive integration process is started. It takes thirteen archives as input to construct an XML index file, which structurally organizes all index information from these archives and serves as an interface to the search engine.

All indexed videoconferences are stored in the home directory of a PVAIS. The indexed archives of the same videoconference are stored in the same subdirectory of the home directory, as shown in Figure 7. Using the starting date and time of the videoconference as the directory name is a good way to avoid conflicts.



Figure 7. Storage structure of PVAIS



## Indexed Videoconference Accessing Module

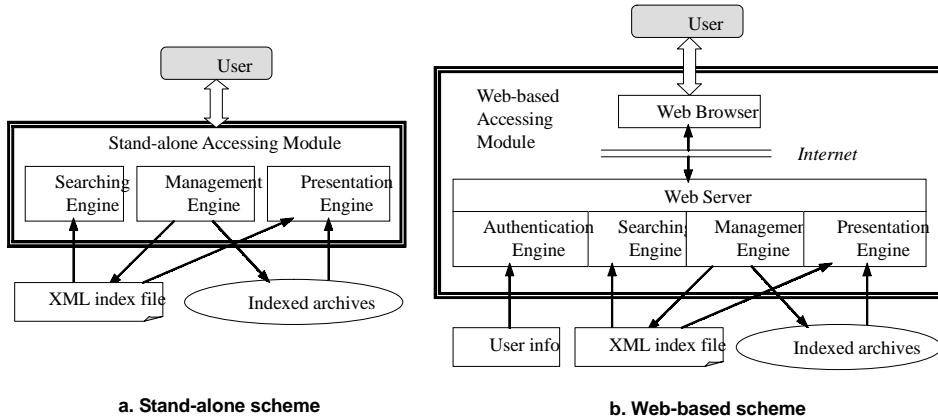
The indexed videoconference accessing module provides a user with an interface to manage, search, and review all indexed conferences. The search function allows the user to search the videoconference of interest by a variety of criteria. The management function can be divided into two levels: conference-oriented management and content-oriented management. Conference-oriented management functions apply to the whole conference, such as classifying the type of a conference (e.g., private or business) or deleting a conference. Content-oriented management functions only apply to specific content of a conference, such as editing keywords or title. The review function supports the synchronized presentation of a videoconference. During the presentation, it also enables the user to pause the presentation and attach annotations or bookmarks to the current time-stamp.

There are two types of implementation schemes for this module: the stand-alone scheme and the Web-based scheme, as shown in Figure 8. The former restricts the user to accessing the indexed videoconferences from the computer in which the indexed videoconferences are stored, while the latter allows the user to access the indexed videoconferences from any computer via the Internet.

According to the stand-alone scheme (Figure 8a), this module is implemented as a stand-alone application that integrates three engines: a searching engine, a management engine, and a presentation engine. These three engines handle the user's requests and return the results to the user. The searching engine searches the index files of all videoconferences for the content specified by the user, and then returns the videoconferences satisfying the user's searching criteria. The management engine maintains the indexed videoconferences according to the user's command. It will affect both the XML index file and the involved indexed archives. When the user selects a videoconference to review, the presentation engine reads synchronization information from the XML index file to control the display of multimedia information stored in the indexed archives.

In the Web-based scheme (Figure 8b), the user interface and the functionality engines are separated. The Web server is situated in the same computer in which the indexed videoconferences are stored. The user can access the indexed videoconferences

Figure 8. Two implementation schemes of the indexed videoconference accessing module



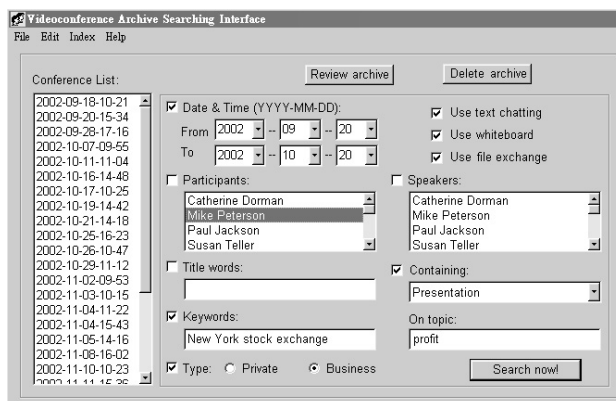
through any ordinary Web browser. The functionality engines are coupled with the Web server. In addition to the three engines discussed in the stand-alone scheme, the Web-based scheme requires an authentication engine to verify the user's identity because a PVAIS is a personal system. In this scheme, Synchronized Multimedia Integration Language (SMIL) can be employed to provide synchronized multimedia presentation in the Web browser.

The choice between the stand-alone scheme and the Web-based scheme depends on the user's habit. For those users who do videoconferencing on laptops and always take their laptops with them, the stand-alone scheme is suitable. For other users, especially those who wish to share their indexed videoconferences, the Web-based scheme is better. From the development point of view, the Web-based scheme is more advanced at the cost of demanding more complex development efforts. In fact, the Web-based scheme contains the stand-alone scheme; therefore, one can first implement a stand-alone accessing module as a prototype, and then extend it to the Web-based accessing module.

We built a prototype system of a PVAIS whose indexed videoconference accessing module is based on the stand-alone scheme. The searching interface and the review interface are shown in Figure 9 and Figure 10, respectively.

The searching interface (Figure 9) is the initial interface of the stand-alone application. Initially, the conference list contains all indexed videoconferences stored in the home directory of a PVAIS. The list of participants and the list of speakers are both loaded from the contact list. Other items are the searching criteria to be set by the user. For example, the user wants to find a business videoconference about "New York stock exchange." He only remembers the approximate date of the videoconference, but forgets who participated in the videoconference. However, he is sure that someone gave a presentation on "Profit" in the videoconference, and that speaker used text chat, whiteboard, and file exchange to communicate. Therefore, the user sets his searching criteria as shown in Figure 9. After the "Search now!" button is pressed, the searching engine will search the XML index files of all indexed videoconferences and then update

Figure 9. Searching interface of the accessing module of PVAIS



the conference list with those satisfying the searching criteria. The user may select a videoconference from the list and press the “Review archive” button to review the videoconference (Figure 10). The other button, “Delete archive,” is used to delete a selected videoconference.


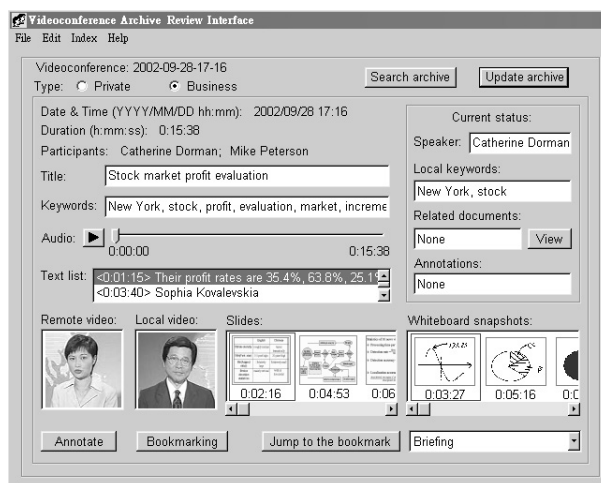
The review interface (Figure 10) provides the management functions and the synchronized presentation of an indexed videoconference. The user can set the conference type or edit those textual indexing items, such as title, keywords, and speaker name. On pressing the “Update archive” button, the updated information will be saved to the XML index file and archives. This interface displays rich information (both static and dynamic) of a videoconference. The static information includes date and time, duration, participants, title, and global keywords. The dynamic information is presented during the synchronized playback. When the user presses the  button to play the audio archive

Figure 10. Review interface of the accessing module of PVAIS



or drags the slide of audio timeline, the contents in the “Current status” frame, “Text list,” “Remote video,” “Local video,” “Slides,” and “Whiteboard snapshots” will be automatically scrolled and highlighted according to the current time-stamp. The user may also change the key frame in the “Remote video” and “Local video” windows or change the highlighted contents in the text list, the slide list, or the whiteboard snapshot list to change the current time-stamp. At any time during the playback, the user can press the “Annotation” button to write comments or press the “Bookmarking” button to insert a bookmark associated with the current time-stamp. These two operations will automatically pause the playback until the operations are finished. The existing annotations are shown in the “Current status” frame. The “Jump to the bookmark” button allows the user to quickly change to the time-stamp associated with the selected bookmark. The annotation and bookmark information will be saved after pressing the “Update archive” button. The function of the “Search archive” button is to finish the review and go back to the searching interface.

## SUBJECTIVE EVALUATION PROTOCOL

This section discusses how to evaluate the performance of a videoconference indexing system. Since the ultimate goals of such systems are to augment people’s memory and to accelerate the content-based searching, the evaluation should be conducted in terms of *recall capability* and *search capability*. The recall capability demonstrates how the system helps the user to recall the information in the videoconference, from the overall level to the detailed level. The search capability indicates how the system supports various means to let the user locate the content of interest quickly and correctly. Since it is still difficult to develop an objective evaluation protocol to represent these high-level criteria, this section defines a subjective evaluation protocol. Table 2 shows the aspects evaluated by this protocol. The upper limit of scoring for each aspect represents the weight of this aspect in its related capability.

The evaluation of recall capability considers eight aspects, as shown in the left half of Table 2. The “Subject” aspect checks whether the title and global keywords outline the videoconference. The “Details” aspect further examines whether the important details can be retrieved from the archives. The “Participant” aspect focuses on the

Table 2. Aspects of subjective evaluation

Recall capability			Search capability		
Aspect	Symbol	Scoring	Aspect	Symbol	Scoring
Subject	$R_1$	0 ~ 6	Time	$S_1$	0 ~ 6
Details	$R_2$	0 ~ 10	Topic	$S_2$	0 ~ 10
Participant	$R_3$	0 ~ 6	Participant	$S_3$	0 ~ 6
Key frame accuracy	$R_4$	0 ~ 8	Visual pattern	$S_4$	0 ~ 8
Speech fidelity	$R_5$	0 ~ 8	Aural pattern	$S_5$	0 ~ 8
Supporting tools	$R_6$	0 ~ 6	Textual pattern	$S_6$	0 ~ 10
Presentation	$R_7$	0 ~ 10	Conference event	$S_7$	0 ~ 6
Extensibility	$R_8$	0 ~ 6	Nonlinear access	$S_8$	0 ~ 10
<b>Overall</b>	$R = \sum R_i$	0 ~ 60	<b>Overall</b>	$S = \sum S_i$	0 ~ 64

completeness of participant's information, including joining or leaving. The "Key frame accuracy" aspect and the "Speech fidelity" aspect evaluate the effect of removing the redundancy in video and audio streams. The "Supporting tools" aspect checks whether the content in other communication tools, for example, text chat, whiteboard, and file exchange, is well indexed. The "Presentation" aspect pays attention to the presentation modes of all types of media and the capability of synchronized presentation. The "Extensibility" aspect checks whether the recall capability can be extended by user's interactions, such as editing, annotation, and bookmarking. The overall recall capability ( $R$ ) sums up the scores of the above eight aspects.

The evaluation of search capability also considers eight aspects, as shown in the right half of Table 2. For the top seven aspects, we evaluate how the system supports searching by these aspects. For the "Nonlinear access" aspect, we check whether the system can automatically locate the point of interest in the searching result according to the user's searching criteria. The overall searching capability ( $S$ ) sums up the scores of the above eight aspects.

After obtaining the scores for all evaluation aspects, the performance (denoted by  $P$ ) of a videoconference indexing system is calculated as follows:

$$P = \alpha \cdot \frac{R}{60} + (1 - \alpha) \cdot \frac{S}{64}$$

Where  $\alpha$  is to adjust the weights of the recall capability and the searching capability. The default value of  $\alpha$  is 0.5. Thus,  $P$  ranges from 0 to 1. The higher  $P$  is, the better performance the videoconference indexing system achieves.

## CONCLUSIONS

This chapter focuses on the videoconference archive indexing, which bears different characteristics from existing research on video indexing, lecture indexing, and meeting support systems. We proposed a comprehensive personal videoconference indexing framework, i.e., PVAIS that consists of three modules: videoconference archive acquisition module, videoconference archive indexing module, and indexed videoconference accessing module. This chapter elaborated the design principles and implementation methodologies of each module, as well as the intra- and inter-module data and control flows. Based on the PVAIS framework, one can easily develop a videoconference indexing system according to his/her specific requirements. Finally, this chapter presented a subjective evaluation protocol for personal videoconference indexing.

## ACKNOWLEDGMENT

The work described in this chapter was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E).

## REFERENCES

- Acosta, E., Torres, L., Albiol, A., & Delp, E. (2002). An automatic face detection and recognition system for video indexing applications. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Vol. 4, IV-3644-IV-3647.
- Albiol, A., Torres, L., & Delp, E.J. (2002). Video preprocessing for audiovisual indexing. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Vol.4, IV-3636-IV-3639.
- Ardizzo, E., La Cascia, M., Di Gesu, V., & Valenti, C. (1996). Content-based indexing of image and video databases by global and shape features. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR'96)*, Vol.3, 140-144.
- Ardizzone, E., & La Cascia, M. (1996). Video indexing using optical flow field. In *Proceedings of the International. Conference on Image Processing (ICIP'96)*, Vol.3, 831-834.
- Ariki, Y., Suiyama, Y., & Ishikawa, N. (1998). Face indexing on video data-extraction, recognition, tracking and modeling. In *Proceedings of the Third IEEE International. Conference on Automatic Face and Gesture Recognition*, 62-69.
- Barras, C., Lamel, L., & Gauvain, J.-L. (2001). Automatic transcription of compressed broadcast audio. In *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Vol. 1, 265-268.
- Ben-Arie, J., Pandit, P., & Rajaram, S. (2001). View-based human activity recognition by indexing and sequencing. In *Proceedings of the International. Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Vol. 2, II-78-II-83.
- Bescos, J., Menendez, J.M., Cisneros, G., Cabrera, J., & Martinez, J.M. (2000). A unified approach to gradual shot transition detection. In *Proceedings of the Intl. Conf. on Image Processing (ICIP'00)*, Vol. 3, 949-952.
- Cai, M., Song, J., & Lyu, M.R. (2002). A new approach for video text detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'02)*, Vol. 1, 117-120.
- Calic, J., & Lzquierdo, E. (2002). A multiresolution technique for video indexing and retrieval. *Proceedings of the International Conference on Image Processing (ICIP'02)*, Vol. 1, 952-955.
- Castleman, K.R. (1996). *Digital image processing*. Englewood Cliffs, NJ: Prentice Hall.
- Chan, Y., Lin, S.-H., Tan, Y.-P., & Kung, S.Y. (1996). Video shot classification using human faces. In *Proceedings of the International. Conference on Image Processing (ICIP'96)*, Vol. 3, 843-846.
- Chang, S.-F. (1995). Compressed-domain techniques for image/video indexing and manipulation. In *Proceedings of the International. Conference on Image Processing (ICIP'95)*, Vol. 1, 314-317.
- Chang, Y.-L., Zeng, W., Kamel, I., & Alonso, R. (1996). Integrated image and speech analysis for content-based video indexing. In *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems (ICMCS'96)*, 306-313.
- Chiu, P., Boreczky, J., Girgensohn, A., & Kimber, D. (2001). LiteMinutes: An Internet-based system for multimedia meeting minutes. In *Proceedings of the 10th World Wide Web Conference*, 140-149.

- Chiu, P., Kapuskar, A., Reitmeier, S., & Wilcox, L. (1999). NoteLook: Taking notes in meeting with digital video and ink. In *Proceedings of the ACM International Conference on Multimedia*, Vol. 1, 149-158.
- Chiu, P., Kapuskar, A., Reitmeier, S., & Wilcox, L. (2000). Room with a rear view. Meeting capture in a multimedia conference room. *IEEE Multimedia*, 7(4), 48-54.
- Christel, M., Kanade, T., Mauldin, M., Reddy, R., Stevens, S., & Wactlar, H. (1996). Techniques for the creation and exploration of digital video libraries. In B. Furht (Ed.), *Multimedia Tools and Applications (Vol. 2)*. Boston, MA: Kluwer Academic Publishers.
- Corridoni, J.M., & Del Bimbo, A. (1996). Structured digital video indexing. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR'96)*, Vol.3, 125-129.
- Cutler, R., & Davis, L. (2000). Look who's talking: Speaker detection using video and audio correlation. In *Proceedings of the International Conference on Multimedia and Expo (ICME'00)*, Vol. 3, 1589-1592.
- Dagtas, S., Al-Khatib, W., Ghafoor, A., & Khokhar, A. (1999). Trail-based approach for video data indexing and retrieval. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, Vol. 2, 235-239.
- Del Bimbo, A. (2000). Expressive semantics for automatic annotation and retrieval of video streams. In *Proceedings of the International Conference on Multimedia and Expo (ICME'00)*, Vol. 2, 671-674.
- Deshpande, S.G., & Hwang, J.-N. (2001). A real-time interactive virtual classroom multimedia distance learning system. *IEEE Trans. on Multimedia*, 3(4), 432-444.
- Dharanipragada, S., & Roukos, S. (1996). A fast vocabulary independent algorithm for spotting words in speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Vol. 1, 233-236.
- Diklic, D., Petkovic, D., & Danielson, R. (1998). Automatic extraction of representative key frames based on scene content. In *Proceedings of the 32nd Asilomar Conference on Signals, Systems & Computers*, Vol. 1, 877-881.
- Di Lecce, V., Dimauro, G., Guerriero, A., Impedovo, S., Pirlo, G., & Salzo, A. (1999). Image basic features indexing techniques for video skimming. In *Proceedings of the International Conference on Image Analysis and Processing*, 715-720.
- Dirfaux, F. (2000). Key frame selection to represent a video. In *Proceedings of the International Conference on Image Processing (ICIP'00)*, Vol. 2, 275-278.
- Doulamis, N.D., Doulamis, A.D., Avrithis, Y.S., & Kollias, S.D. (1998). Video content representation using optimal extraction of frames and scenes. In *Proceedings of the International Conference on Image Processing (ICIP'98)*, Vol. 1, 875-879.
- Eickeler, S., Wallhoff, F., Lurgel, U., & Rigoll, G. (2001). Content based indexing of images and video using face detection and recognition methods. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Vol. 3, 1505-1508.
- Fletcher, L.A., & Kasturi, R. (1988). A robust algorithm for text string separation from mixed text/graphics images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 10(6), 910-918.
- Foote, J., Boreczsky, J., & Wilcox, L. (1999). Finding presentations in recorded meetings using audio and video features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, Vol. 6, 3029-3032.

- Gao, Q., Ko, C.C., & De Silva, L.C. (2000). A universal scheme for content-based video representation and indexing. In *Proceedings of the 2000 IEEE Asia-Pacific Conference on Circuits and Systems*, 469-472.
- Gargi, U., Antani, S., & Kasturi, R. (1998). VADIS: A Video Analysis, Display and Indexing System. In *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR'98)*, 965-965.
- Gelin, P., & Wellekens, C.J. (1996). Keyword spotting enhancement for video soundtrack indexing. In *Proceedings of the 4th International Conference on Spoken Language*, Vol. 2, 586-589.
- Geyer, W., Richter, H., & Abowd, G.D. (2003). Making multimedia meeting records more meaningful. In *Proceedings of the International Conference on Multimedia and Expo (ICME'03)*, Vol. 2, 669-672.
- Ginsberg, A., & Ahuja, S. (1995). Automating envisionment of virtual meeting room histories. In *Proceedings of the ACM International. Conference on Multimedia*, 65-75.
- Gross, R., Bett, M., Yu, H., Zhu, X., Pan, Y., Yang, J., & Waibel, A. (2000). Towards a multimodal meeting record. In *Proceedings of the International. Conference on Multimedia and Expo (ICME'00)*, Vol. 3, 1593-1596.
- Gu, L., & Bone, D. (1999). Skin colour region detection in MPEG video sequences. In *Proceedings of the International Conference on Image Analysis and Processing*, 898-903.
- Hauptmann, A.G. (1995). Speech recognition in the Informedia Digital Video Library: Uses and limitations. In *Proceedings of the 7th International Conference on Tools with Artificial Intelligence*, 288-294.
- Hauptmann, A.G., & Wactlar, H.D. (1997). Indexing and search of multimodal information. In *Proceedings of the IEEE International. Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Vol. 1, 195-198.
- Hidalgo, J.R., & Salembier, P. (2001). Robust segmentation and representation of foreground key regions in video sequences. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Vol. 3, 1565-1568.
- Hsu, C.-T., & Teng, S.-J. (2002). Motion trajectory based video indexing and retrieval. In *Proceedings of the International Conference on Image Processing (ICIP'00)*, Vol. 1, 605-608.
- Hua, X.-S., Yin, P., & Zhang, H.-J. (2002). Efficient video text recognition using multiple frame integration. In *Proceedings of the International Conference on Image Processing (ICIP'02)*, Vol. 2, II-397 -II-400.
- Hwang, J.-N., & Luo, Y. (2002). Automatic object-based video analysis and interpretation: A step toward systematic video understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Vol. 4, IV-4084-IV-4087.
- Ito, H., Sato, M., & Fukumura, T. (2000). Annotation and indexing in the video management system (VOM). In *Proceedings of the 2000 IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 2, 834-839.
- ITU-T Recommendation H.323 Draft v4, (2001). Packet-based multimedia communications systems. Retrieved February 2001: <http://www.itu.int/rec/recommendation.asp>



- ITU-T Recommendation H.225.0 (2003a). Call signaling protocols and media stream packetization for packet-based multimedia communication systems. Retrieved July 2003: <http://www.itu.int/rec/recommendation.asp>
- ITU-T Recommendation H.245 (2003b). Control protocol for multimedia communication. Retrieved August 2001: <http://www.itu.int/rec/recommendation.asp>
- Iyengar, G., Nock, H., Neti, C., & Franz, M. (2002) Semantic indexing of multimedia using audio, text and visual cues. In *Proceedings of the International Conference on Multimedia and Expo (ICME'02)*, Vol. 2, 369-372.
- Jain, A.K., & Yu, B. (1998). Automatic text location in images and video frames. In *Proceedings of the 14th International Conference on Pattern Recognition (ICPR'98)*, Vol. 2, 1497-1499.
- Jasinschi, R.S., Dimitrova, N., McGee, T., Agnihotri, L., & Zimmerman, J. (2001a). Video scouting: An architecture and system for the integration of multimedia information in personal TV applications. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Vol. 3, 1405-1408.
- Jasinschi, R.S., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J., & Li, D. (2001b). Integrated multimedia processing for topic segmentation and classification. In *Proceedings of the International Conference on Image Processing (ICIP'01)*, Vol. 3, 366-369.
- Jasinschi, R.S., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J., Li, D., & Louie, J. (2002). A probabilistic layered framework for integrating multimedia content and context information. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Vol. 2, 2057-2060.
- Jin, R., & Hauptmann, A. (2001). Learning to select good title words: A new approach based on reversed information retrieval. In *Proceedings of the International Conference on Machine Learning (ICML'01)*, 242-249.
- Joukov, N., & Chiueh T.-C. (2003). Lectern II: A multimedia lecture capturing and editing system. In *Proceedings of the International Conference on Multimedia and Expo (ICME'03)*, Vol. 2, 681-684.
- Ju, S.X., Black, M.J., Minneman, S., & Kimber, D. (1997). Analysis of gesture and action in technical talks for video indexing. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'97)*, 595-601.
- Kameda, Y., Nishiguchi, S., & Minoh, M. (2003). Carmul: Concurrent automatic recording for multimedia lecture. In *Proceedings of the International Conference on Multimedia and Expo (ICME'03)*, Vol. 2, 677-680.
- Kang, E.K., Kim, S.J., & Choi, J.S. (1999). Video retrieval based on key frame extraction in compressed domain. In *Proceedings of the International Conference on Image Processing (ICIP'99)*, Vol. 3, 260-264.
- Kang, J., & Mersereau, R.M. (2002). An effective method for video segmentation and sub-shot characterization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Vol. 4, IV-3652-IV-3655.
- Kazman, R., Al-Halimi, R., Hunt, W., & Mantei, M. (1996). Four paradigms for indexing video conferences. *IEEE Multimedia*, 3(1), 63 -73.
- Kazman, R., & Kominek, J. (1999). Supporting the retrieval process in multimedia information systems. In *Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*, Vol. 6, 229-238.

- Kim, E.Y., Kim, K.I., Jung, K., & Kim, H.J. (2000). A video indexing system using character recognition. In *Digest of Technical Papers of International Conference on Consumer Electronics*, 358-359.
- Kim, S.H., & Park, R.-H. (2000). A novel approach to scene change detection using a cross entropy. In *Proceedings of the International Conference on Image Processing (ICIP'00)*, Vol. 3, 937-940.
- Kristjansson, T., Huang, T.S., Ramesh, P., & Juang, B.H. (1999). A unified structure-based framework for indexing and gisting of meetings. In *Proceedings of the International Conference on Multimedia Computing and Systems*, Vol. 2, 572-577.
- Lagus, K., & Kaski, S. (1999). Keyword selection method for characterizing text document maps. In *Proceedings of the International Conference on Artificial Neural Networks*, Vol. 1, 371-376.
- Lebourgeois, F., Jolion, J.-M., & Awart, P.C. (1998). Towards a description for video indexation. In *Proceedings of the 14th Intl. Conf. on Pattern Recognition (ICPR'98)*, Vol. 1, 912-915.
- Leung, W.H., Chen, T., Hendriks, F., Wang, X., & Shae, Z.-Y. (2002). eMeeting: A multimedia application for interactive meeting and seminar. In *Proceedings of the Global Telecommunications Conference*, Vol. 3, 2994-2998.
- Li, C.-S., Mohan, R., & Smith, J.R. (1998). Multimedia content description in the InfoPyramid. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, Vol. 6, 3789-3792.
- Li, H., & Doermann, D. (1998). Automatic identification of text in digital video key frames. In *Proceedings of the 14th International Conference on Pattern Recognition (ICPR'98)*, Vol. 1, 129-132.
- Liang, R.Z., Venkatesh, S., & Kieronska, D. (1995). Video indexing by spatial representation. In *Proceedings of the 3rd Australian and New Zealand Conference on Intelligent Information Systems (ANZIIS'95)*, 99-104.
- Luo, Y., & Hwang, J.N. (2003). Video sequence modeling by dynamic Bayesian networks: A systematic approach from fine-to-coarse grains. In *Proceedings of the International Conference on Image Processing (ICIP'03)*, Barcelona, Spain, September.
- Lyu, M.R., Yau, E., & Sze, K.S. (2002). A multilingual, multimodal digital video library system. In *Proceedings of the Joint Conference on Digital Libraries (JCDL 2002)*, 145-153.
- Madrane, N., & Goldberg, M. (1994). Towards automatic annotation of video documents. In *Proceedings of the 12th IAPR International Conference on Computer Vision & Image Processing*, Vol. 1, 773-776.
- Maziere, M., Chassaing, F., Garrido, L., & Salembier, P. (2000). Segmentation and tracking of video objects for a content-based video indexing context. In *Proceedings of the International Conference on Multimedia and Expo (ICME'00)*, Vol. 2, 1191-1194.
- Mikolajczyk, K., Choudhury, R., & Schmid, C. (2001). Face detection in a video sequence - A temporal approach. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Vol. 2, II-96-II-101.
- Nam, J., Cetin, E., & Tewfik, A.H. (1997). Speaker identification and video analysis for hierarchical video shot classification. In *Proceedings of the International Conference on Image Processing (ICIP'97)*, Vol. 2, 550-553.

- Naphade, M.R., & Huang, T.S. (2000). Inferring semantic concepts for video indexing and retrieval. In *Proceedings of the International Conference on Image Processing (ICIP'00)*, Vol. 3, 766-769.
- Naphade, M.R., Kristjansson, T., Frey, B., & Huang, T.S. (1998). Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In *Proceedings of the International Conference on Image Processing (ICIP'98)*, Vol. 3, 536-540.
- Ngo, C.-W., Pong, T.-C., & Huang, T.S. (2002). Detection of slide transition for topic indexing. In *Proceedings of the International Conference on Multimedia and Expo (ICME'02)*, Vol. 2, 533-536.
- Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.
- Rogina, I., & Schaaf, T. (2002) Lecture and presentation tracking in an intelligent meeting room. In *Proceedings of the 4th Intl. Conf. on Multimodal Interfaces*, 47-52.
- Sato, S., & Kanade, T. (1997). NAME-IT: Association of face and name in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, 368-373.
- Sawhney, H.S. (1993). Motion video annotation and analysis: An overview. In *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, Vol. 1, 85-89.
- Song, J., Lyu, M., Hwang, J.-N., & Cai, M. (2003). PVCAIS: A personal videoconference archive indexing system. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME'03)*, 117-120.
- Sprey, J.A. (1997). Videoconferencing as a communication tool. *IEEE Trans. on Professional Communication*, 40(1), 41-47.
- Stewart, A., Wolf, P., & Heminje, M. (2003). Media and metadata management for capture and access systems in electronic lecturing environments. In *Proceedings of the International Conference on Multimedia and Expo (ICME'03)*, Vol. 2, 685-688.
- Sun, M.T., Wu, T.-D., & Hwang, J.-N. (1998). Dynamic bit allocation in video combining for multipoint conferencing. *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, 45(5), 644-648.
- Tang, H.-M., Lyu, M.R., & King, I. (2003). Face recognition committee machine. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, 837-840.
- Tekalp, A.M. (1995). *Digital video processing*. Upper Saddle River, NJ: Prentice Hall.
- Tsapatsoulis, N., Avrithis, Y., & Kollias, S. (2000). Efficient face detection for multimedia applications. In *Proceedings of the International Conference on Image Processing (ICIP'00)*, Vol. 2, 247-250.
- Tse, K., Wei, J., & Panchanathan, S. (1995). A scene change detection algorithm for MPEG compressed video sequences. In *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, Vol. 2, 827-830.
- Tsekeridou, S., & Pitas, I. (1998). Speaker dependent video indexing based on audio-visual interaction. In *Proceedings of the International Conference on Image Processing (ICIP'98)*, Vol. 1, 358-362.
- Vinciarelli, A., Bengio, S., & Bunke, H. (2003). Off-line recognition of large vocabulary cursive handwritten text. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*, 1101-1105.

- Viswanathan, M., Beigi, H.S.M., Tritschler, A., & Maali, F. (2000). Information access using speech, speaker and face recognition. In *Proceedings of the International Conference on Multimedia and Expo (ICME'00)*, Vol. 1, 493-496.
- Wactlar, H., Kanade, T., Smith, M., & Stevens, S. (1996). Intelligent access to digital video: The Informedia Project. *IEEE Computer: Digital Library Initiative special issue*, 29(5), 46-52.
- Wang, P., Ma, Y.-F., Zhang, H.-J., & Yang, S. (2003). A people similarity based approach to video indexing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, Vol. 3, III-693-III-696.
- Wei, J., Li, Z.-N., & Gertner, I. (1999). A novel motion-based active video indexing method. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, Vol. 2, 60-465.
- Wilcox, L., & Boreczky, J. (1998). Annotation and segmentation for multimedia indexing and retrieval. In *Proceedings of the 31st International Conference on System Sciences*, Vol. 2, 259-266.
- Wilcox, L., Chen, F., Kimber, D., & Balasubramanian, V. (1994). Segmentation of speech using speaker identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, Vol. 1, 161-164.
- Young, S.J., Brown, M.G., Foote, J.T., Jones, G.J.F., & Sparck-Jones, K. (1997). Acoustic indexing for multimedia retrieval and browsing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Vol. 1, 199-202.
- Zhang, H.J., Wang, J.Y.A., & Altunbasak, Y. (1997). Content-based video retrieval and compression: A unified solution. In *Proceedings of the International Conference on Image Processing (ICIP'97)*, Vol. 1, 13-16.
- Zeng, W., Gao, W., & Zhao, D. (2002). Video indexing by motion activity maps. In *Proceedings of the International Conference on Image Processing (ICIP'02)*, Vol. 1, 912-915.