

PVCAIS: A PERSONAL VIDEOCONFERENCE ARCHIVE INDEXING SYSTEM

Jiqiang Song¹, Michael Lyu¹, Jenq-Neng Hwang², and Min Cai¹

¹Dept. Computer Science & Engineering, Chinese University of Hong Kong, Hong Kong
{jqsong, lyu, mcai}@cse.cuhk.edu.hk

²Dept. Electrical Engineering, University of Washington, USA
hwang@ee.washington.edu

ABSTRACT

With the rapid deployment of videoconference, the fast-accumulated personal videoconference archives need to be effectively indexed. This paper proposes a well-designed indexing system - PVCAIS that integrates many multimedia-indexing techniques to manage personal videoconference archives. Firstly, the contents of video, audio, text and whiteboard communications are stored after removing the redundancies. Next, more information, e.g., participants, title, keywords and slides, is extracted by face detection and recognition, speech recognition, OCR, automatic title generation, keyword selection, etc. Then, an XML index file containing the summary of the videoconference is also generated. The whole indexing process is automatic except that the face of a new contact needs to be interactively identified for only once. Finally, we demonstrate a graphical user interface which allows the user to search and browse the indexed videoconference archives conveniently.

1 INTRODUCTION

With the rapid growth of Internet bandwidth, videoconference becomes more and more popular in business activities [1]. With affordable video and audio capture devices, the low bit-rate coding, and new communication technology, home users can also enjoy visual communications at 56Kbps or lower [2]. Furthermore, the video-based distance learning has already served as an important mean in the education field [2]. The participant of a videoconference usually wishes to keep the video archive for the later reference. Therefore, we can imagine that, in the near future, a person will accumulate many video archives in his/her work, study and daily life. However, since the videoconference archives are stored as normal video and audio files, only the file name is a place to indicate the subject. But, not everyone is glad to compose a suitable file name. Moreover, it is not easy to recall the details of a conference without watching it again. Since the visual and aural information is not directly searchable like text, it is also difficult to search out those archives with content of interest by normal searching engines. Therefore, the research of indexing personal videoconference archives is of timely importance. However, current efforts on multimedia indexing still focus on digital video libraries [3] or distance learning. Reports on indexing videoconference archives have not been found so far as we investigated.

This paper applies many multimedia-indexing techniques to personal videoconference archives to generate a searchable content summary automatically. These techniques are well integrated orienting to the characteristics of videoconference. Instead of using traditional databases, we design an XML-based repository to integrate visual, aural, text and graphical information structurally, which facilitates searching and browsing the summary of videoconference archives. This paper presents the architecture of PVCAIS (Personal VideoConference Archive Indexing System) and demonstrates its user interface to search and browse indexed videoconference archives.

2 RELATED WORK

There are two classes of related techniques: videoconferencing techniques and multimedia-indexing techniques.

Most videoconferencing systems are based on the ITU-T H.323 framework. Videoconferencing techniques, such as *vic* for video, *vat* or *rat* for audio, *wb* for whiteboard and *nte* for text [4,5], create incoming and outgoing channels for video and audio, share information via whiteboard, and exchange text messages. However, since these techniques are beyond the scope of this paper, we assume that the user already has a videoconferencing client which communicates with the server through all of the above channels.

Multimedia-indexing techniques arise from the research of digital video libraries [3]. The low-level image-based processing, such as key-frame selection, shot detection and scene detection, is essential. The most common, also the most valuable, technique is to find a set of keywords by some data mining algorithms from the text source of a video to describe the video content [6,7]. The text source comes from the script of subtitle, the results of speech recognition [8] and/or video text recognition [9]. Since many video clips record human activities, face detection [10] and face recognition [11] also play an important role in multimedia indexing, especially for celebrities' faces and named faces [12]. Frankly speaking, for building a public video library, the artificial intelligence techniques (e.g., speech recognition and face recognition) still have difficulties when handling unlimited targets. However, they are more suitable to be applied to processing personal videoconference archives since the number of contacts of one user is limited. Some distance learning systems [13,14] also includes multimedia-indexing tools; however, they involve many human interactions. Our system aims to release the user's work as much as possible.

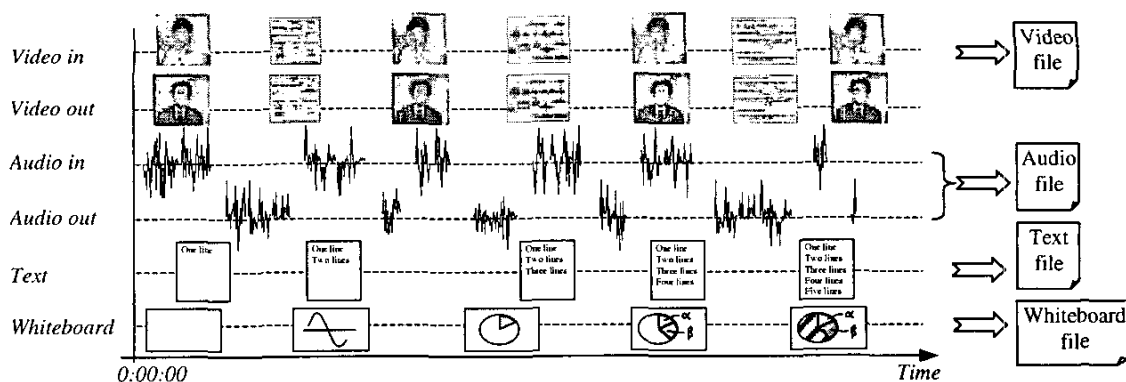


Figure 1. Illustration of the videoconferencing channels and the content storage

3 DESIGNS AND ARCHITECTURE

A typical videoconferencing client consists of six communication channels: *Video in*, *Video out*, *Audio in*, *Audio out*, *Text*, and *Whiteboard*. The *Video in* channel receives the image shown to the user from the server, while the *Video out* channel sends out the local image. The *Audio in* channel and the *Audio out* channel contain incoming and outgoing voices, respectively. The *Text* channel and the *Whiteboard* channel are shared to all participants for typed messages and handwritten contents, respectively. These channels are synchronized in time, as shown in Figure 1.

For the indexing purpose, we need to store the content of each channel including the timing information except the *Video out* channel since that channel contains the local image of the user side therefore it is not important. Thus, as illustrated in Figure 1, the *Video in* channel is stored in a video file. The *Audio in* channel and the *Audio out* channel are mixed and stored in an audio file since the participants talk alternately. The *Text* channel and the *Whiteboard* channel are stored in a text file and a whiteboard file, respectively.

Among the four channels, the audio channel is most important. Although we can obtain a script of the conversation by speech recognition, we still think it is necessary to keep the entire audio content in the file, since the speech recognition is not accurate enough.

On the other hand, the *Video* channel contains a lot of redundancies. The *Video* channel shows either a human image or a slide with a relatively long duration. Thus, it is unnecessary to keep the entire video content in the file. A set of key frames with the timestamp (relative time elapsed since the beginning of videoconference) can represent the video content well. The frame selection is based on the change of video content. Let $f(t)$ denote the function of video content feature which varies with time t . Thus, the changes of video content will be detected in $f(t)$ as peaks, as shown in Figure 2, and the valleys right after each peak can be 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 [15]. Upon selecting a key frame, the key frame is stored in JPEG format. The current timestamp and the JPEG file name are written into the video file as a pair. For example, if the key frame at the time 0:02:34 is saved as "v1.jpg", the line written into the video file will be "<0:02:34> FILE=

v1.jpg; FACE=True". The "FACE" flag indicates whether this frame contains a human face.

Since the *Text* channel and the *Whiteboard* are not updated frequently, we also use the timestamp-based format similar to the video file to store them, that is, storing the timestamp of update and the updated information as a pair into the file.

The *Text* channel is updated when a new message from any participant is sent. Therefore, the current timestamp and the new message are written into the text file. A fragment of a text file is shown in Figure 3.

The *Whiteboard* channel contains handwritten texts and graphics. The update of this channel happens not at a point, but in a period of time. To detect when the update, i.e., the writing action, begins and finishes, we monitor the whiteboard by sampling it as a binary image every two seconds. As illustrated in Figure 4, the current sample is compared with the last sample by the XOR operation. If they are different, the update begins. When they become same again, the update finishes and the current sample is the

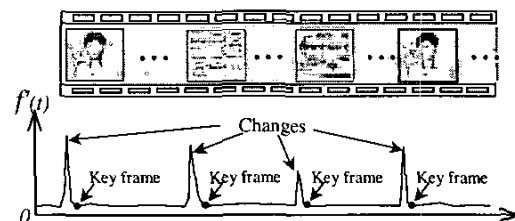


Figure 2. Key frame selection in the Video channel

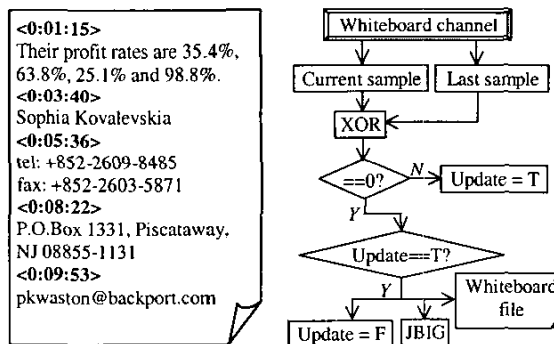


Figure 3. A text file Figure 4. Whiteboard file generation

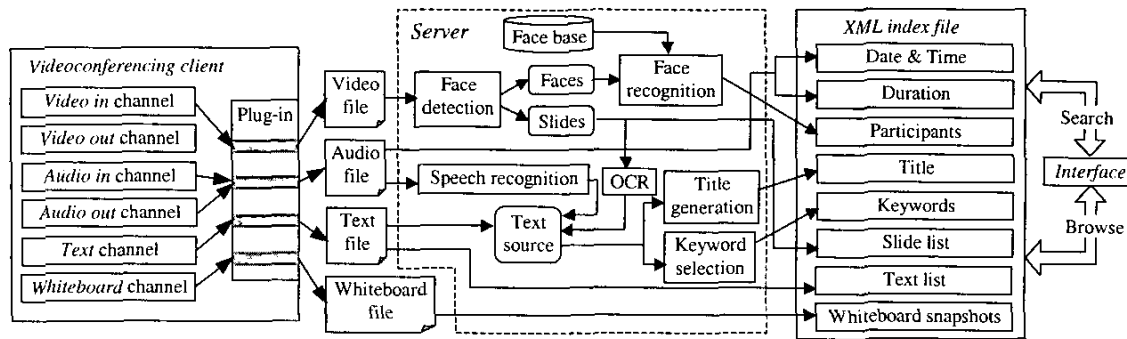


Figure 5. The architecture of PVCAIS

updated whiteboard image. Similar to the video file, the current timestamp and the filename saving the current sample are written into the whiteboard file as a pair. The binary sampling image should be saved in a compressed format that supports the lossless compression of a binary image, such as JBIG, to ensure the fidelity.

With the above techniques, we obtain four content files after a videoconference. The generator of these files should be implemented as a plug-in to the videoconferencing client. These files are managed by the indexing system, which stores the files of the same conference in one folder. The folder name indicates the date and time of the videoconference.

Although these four content files contain the outline of the videoconference with the timing information, they do not provide an index for the videoconference. Thus, the following techniques are applied to generate more indices, including the participants' names, a title and a group of keywords.

The participants are identified by performing the face recognition on the video file. Since the video file contains both human images and slides, the face detection [10] should be performed first to discriminate them and set the above-mentioned "FACE" flag. The indexing system maintains a face base of the user's contacts. Therefore, if a face is recognized from the video file as an existing contact, the contact's name will be added to the participants automatically. Initially, the face base is empty. It grows in the following way. If a face is detected from the video file, but it is not recognized according to the current face base, the indexing system will prompt the user to identify the face by specifying a name. The new face and its name will then be added to the face base. Thus, for each contact, the user only needs to specify his/her name once. After that, the indexing system will recognize them automatically.

The title and keywords for the videoconference are generated from the text source of the videoconference. The text source consists of three parts. The first part is the script of the audio file, which is obtained by the speech recognition. The second part comes from the text file. The third part is extracted by OCR [16] from the slides in the video file. With plentiful words in the text source, the automatic title generation technique [6] will compose a title to describe the content, and the text clustering technique [7] will find a group of keywords that feature the content best.

Finally, all indices and content files are integrated into an XML-based index file, which structurally summarizes the videoconference. The index file includes the date and time,

the duration, the participants' names, the title, the keywords, the slides list imported from the video file (FACE=false), the text list imported from the text file, the list of whiteboard snapshots imported from the whiteboard file, and a link to the audio file. Therefore, the index file is fully searchable. The indexing system also provides a graphical user interface for to search and browsing the indexed videoconference archives, as demonstrated in Section 4.

Based on the above designs, the entire architecture of the personal videoconference archive indexing system, i.e., PVCAIS, is shown in Figure 5.

Figure 6 shows the file system structure maintained by PVCAIS, whose home directory is "MyVideoConference". Since PVCAIS is a personal system, it only serves the same user in the same computer. After a videoconference finishes, the plug-in creates a new folder named by the current date and time under the home folder. All the content files, except the XML index file, are written into this folder. Then, the indexing server is invoked to further analyze the content files and generate the XML index file.

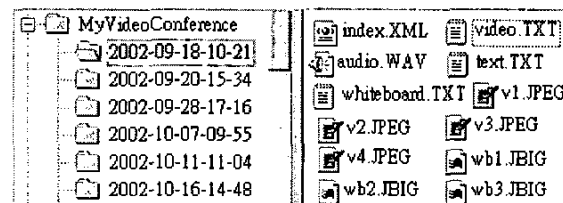


Figure 6. File system structure of PVCAIS

4 USER INTERFACE

Figure 7 shows an example user interface of PVCAIS. The interface has two main functions: searching and browsing. Besides, it also allows the user to print the videoconference memo, to edit some index items and to re-index the specified videoconference.

After starting the interface, the default function is searching, as shown in the left screen of Figure 7. PVCAIS will load all existing videoconferences and list them in the left list. The user can search by date, participant, title or keywords, or any combination of them. After clicking the "Search now!" button, the left list will be updated to show only those videoconferences satisfying the searching condition. The user may select one videoconference and click the "Browse archive" button to browse the summary of the conference, as shown in

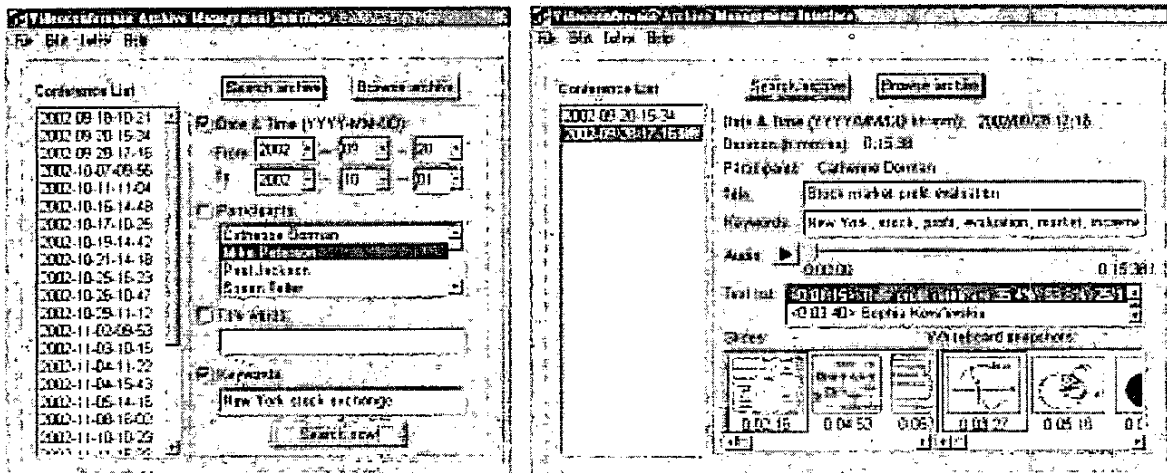



Figure 7. An example user interface of PVC AIS

the right screen copy of Figure 7. The user can modify the title and keywords if necessary. He/she also can click a slide or a whiteboard snapshot to view it in its actual size. When the user presses the  button to play the audio record, PVC AIS supports the synchronized presentation of text lists, slides and whiteboard snapshots by their timestamps, which facilitates the user to review the whole videoconference.

5 CONCLUSIONS

Videoconference becomes more and more popular, which demands the development of personal videoconference archive indexing tools. For this purpose, this paper proposes a practical personal videoconference archive indexing architecture based on many multimedia-indexing techniques, which integrates the video, audio, text and whiteboard information effectively to automatically compose the concise but comprehensive summaries of the videoconference archives. The indexing process is transparent to users, who will only be asked to identify each new contact once. Its graphical user interface also provides the convenient searching and browsing support for users.

ACKNOWLEDGEMENT

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 4360/02E).

REFERENCES

- [1] J.A. Sprey, "Videoconferencing as a communication tool," *IEEE Trans. on Professional Communication*, 40(1): 41-47, March 1997.
- [2] S.G. Deshpande and J.-N. Hwang, "A real-time interactive virtual classroom multimedia distance learning system," *IEEE Trans. on Multimedia*, 3(4): 432-444, 2001.
- [3] M. Christel, T. Kanade, M. Mauldin, R. Reddy, S. Stevens, and H. Wactlar, "Techniques for the Creation and Exploration of Digital Video Libraries." *Multimedia Tools and Applications (Vol. 2)*, Borko Furht, editor. Boston, MA: Kluwer Academic Publishers, 1996.

- [4] vic, vat, wb: Video, audio, whiteboard conferencing tools. Available online: <http://www-nrg.ee.lbl.gov/>.
- [5] rat, nte: Audio, text tools. Available online: <http://www-mice.cs.ucl.ac.uk/multimedia/software/>.
- [6] R. Jin and A. Hauptmann, "Learning to Select Good Title Words: A New Approach based on Reversed Information Retrieval," in *Proc. Intl. Conf. on Machine Learning (ICML'01)*, Jun 28-Jul 1, 2001.
- [7] K. Lagus and S. Kaski, "Keyword selection method for characterizing text document maps," In *Proc. Intl. Conf. on Artificial Neural Networks*, vol.1, pp. 371-376 1999.
- [8] M. Witbrock and A. Hauptmann, "Speech Recognition for a Digital Video Library," *Journal of the American Society for Information Science*, 49(7): 619-632, 1998.
- [9] M. Cai, J. Song and M.R. Lyu, "A new approach for video text detection," In *Proc. Intl. Conf. on Image Processing*, pp. 117-120. Rochester, New York, USA, September 22-25, 2002.
- [10] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1): 23-38, 1998.
- [11] Y. Gao and M.K.H. Leung, "Face recognition using line edge map," *IEEE Trans on Pattern Analysis and Machine Intelligence*, 24(6): 764-779, 2002.
- [12] S. Sato and T. Kanade, "NAME-IT: Association of Face and Name in Video," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 368-373, 1997.
- [13] G.D. Abowd, *et al.*, "Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project", In *Proc. ACM Multimedia'96 Conference*, pp. 187-198, November 1996.
- [14] A. Ginsberg, *et al.*, "'The little web schoolhouse' using virtual rooms to create a multimedia distance learning environment", In *Proc. ACM Multimedia'98 Conference*, pp. 89-98, 1998.
- [15] F. Dirfaux, "Key frame selection to represent a video," In *Proc. Intl. Conf. on Image Processing*, vol.2, pp. 275-278, Vancouver, BC, Canada, 2000.
- [16] I. Bazzi, R. Schwartz, and J. Makhoul, "An omnifont open-vocabulary OCR system for English and Arabic," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 21(6): 495-504, 1999.