XVIP: An XML-Based Video Information Processing System

C.H. Ngai, P.W. Chan, Edward Yau, Michael R. Lyu
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
{chngai, pwchan, edyau, lyu}@cs.cuhk.edu.hk

Abstract

We describe XVIP, an XML-based video information processing system, which extracts information from video and stores the information in a multimedia digital video library. XVIP encapsulates a number of extraction techniques, including scene change detection, video optical character detection and recognition, and geometric coding. It also provides a seamless approach to scale up the contents created in and delivered by the target multimedia digital video library. Furthermore, XVIP can handle multilingual contents. XVIP is based on a multi-modal concept, which treats each content extraction component as a modality, and users can easily add new modalities into XVIP. The information extracted from the video is then stored in a flexible, scalable and reusable way based on a generic XML structure, providing a convenient mechanism for data representation on Web browsers. Also, the content in the XML file can be used to perform knowledge enrichment on top of the primary information extracted from the video. This enriched data representation helps users search for multimedia video content more efficiently.

Keywords: *Multimedia systems, Multi-modal data representation, XML, Internet and Web-based systems.*

1. INTRODUCTION

1.1. Motivation

Recently, many digital video libraries are developed and a lot of attentions are paid to how to present multimedia information. However, little study has been performed on how to extract information from video and to store this information with flexibility for indexing and searching purpose. This paper will focus on the extraction of the pertinent, multi-modal information from videos and the integration of the extracted information into a flexible, XML-based digital library system.

Not only the accuracy of information extraction techniques is the key to the success of a digital library, but also the increasing number of different techniques affects the design of an "open" digital video library system to a great extent. How to scale up the digital library in terms of adding new extraction components also presents challenges. Whenever a new extraction method is developed, it imposes a series of new indexing and presentation functions to be included. Therefore, a generic framework for presentation and visualization of video information is curial to the deployment of the digital library. We attempt to implement a system that can meet this challenge.

1.2. Objective

In this paper, we aim at illustrating how different information in videos can be extracted and edited, how the information can be stored into an XML format, how the secondary information can be included and used, and how the XML documents can be presented in other structure, e.g. SMIL. Consequently, we designed and implemented an XML-based video information processing system, XVIP, to address these issues.

XVIP achieves the following targets:

- To provide an open architecture that can ease the overhead of integrating different video processing, searching, indexing and presentation of various digital video library functions.
- To increase the reusability of the information extracted from the videos, including information interchange between distributed video libraries and different publishing media.
- Video information is extracted once and delivered and presented multiple times to different computing platforms.

The above objectives are achieved via the following methods encapsulated in XVIP:

- Modality concept of the digital video library functions throughout the video information life-cycle.
- Collaboration of the video information processing modules.
- Generic framework for presentation and visualization of the video information.

1.3. Overview

Various information pieces can be extracted from a video processing sequence: generating multimedia abstractions, classifying topics, recognizing speeches, detecting human faces, selecting key frames, skimming videos, and performing video optical character recognition (VOCR) [1]. Digital Video Library combines

0730-3157/02 $17.00 © 2002 **IEEE**

all these techniques [2]. It is a new approach for automated video and audio indexing, navigation, visualization, searching and retrieval, and the application of these techniques for content creation in education, information and entertainment environments [3]. Storage of information extracted from video in a flexible, scalable and reusable way becomes very important. We chose XML in our system implementation as it satisfies all the above requirements. Figure 1 shows the overall processing of XVIP system in several phases.

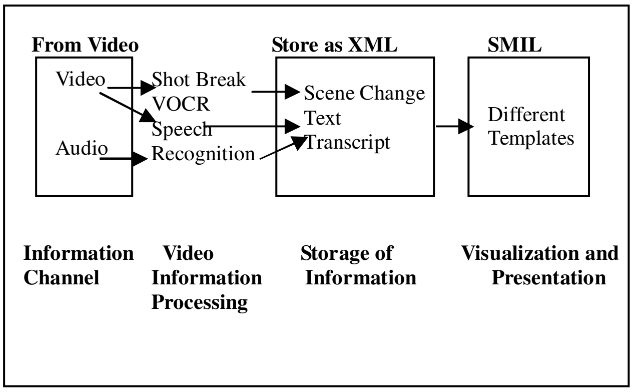


Figure 1. Overview of XVIP system

Multi-modal information processing is an important concept in XVIP design. We define a modality as a domain or type of information that can be extracted from the video. Examples are text extracted by speech recognition and human identity by face recognition. The set of functions that a modality supports the digital library applications is called the modality dimension. Typical processes are video information extraction, indexing and presentation. For different modalities, the requirements for the whole series of library functions can be totally different. Furthermore, at the front end of the presentation client, a text visualization tool is very different from a human face visualization tool in terms of layouts, contents and presentation styles. Although the implementation of individual modality will be different, the "life-cycle" of these modalities is similar and can be grouped as follows:

- Video Information Processing
- Information Repository
- Indexing and Searching
- Visualization and Presentation

By introducing the modality concept, the integration interface and information exchange of different modalities (dimensions) can be easily identified. This provides an efficient way to adding different new modalities into the system without losing the flexibility.

2. Related works

For an effective digital video library, users need to be able to find the video segments they want. Realizing this goal requires automatic content-based indexing of videos that significantly improves the users' ability in accessing specific segments of interest within the videos. Videos, soundtracks and transcripts will be digitized, and information from the soundtracks and transcripts will be used to automatically index videos in a frame-by-frame manner. This will allow users to quickly search indices for multiple videos, to locate segments of interest, and to view and manipulate these segments on their remote computers [1].

A number of information items can be extracted from a video sequence:

1. Integration of speech, natural language, and image processing.
2. Text processing.
3. Audio processing.
4. Image processing.
5. Video processing [3].

3. Extraction techniques

XVIP integrates video information extraction techniques in a flexible, and reusable framework. We describe a few extraction techniques in detail, including scene change detection, video optical character recognition (VOCR), and video optical character detection (VOCD). Scene change technique is chosen, as it is the most effective method for segmenting a video sequence into significant components. VOCR can help to extract the captions or important texts appearing in a video for better understanding, while VOCD is a technique that can locate topics of interest in a series of video frames.

3.1. Scene change

Scene change detection is an effective method for segmenting a video sequence into significant components, generally called shots [4]. A shot is defined as an unbroken sequence of frames recorded from a single camera, which forms the building block of a video. The purpose of shot boundary detection is to segment the video stream into multiple shots [5].

From the result of some research, statistical and structural properties of images are used to identify scene changes. These features are used in three steps to identify the scene changes sequentially, such as abrupt scene change detection, dissolve scene change detection, and wiping scene change detection. Since abrupt scene change is the most common in video sequence, we use this approach in our system.

Our experiments show that the accuracy of the histogram difference method is acceptable. We enhance it with a dynamic threshold to further improve its accuracy for scene change detection and to reduce over-detection and un-detection. We sample one scene from the video every 0.05 seconds for comparison with the pervious scene. The grasped scenes are coded in 24-bit image, eight bits for each color (red R, green G, blue B). Consequently, we can check each pixel and classify them into different classes.

For efficiency purpose, we only consider the most significant two bits for each color. Then, we calculate the total difference of the whole histogram (H), which is:

$$H = \Sigma_{(0,\,63)}(P_{a,}(i) - P_b(j))^2$$

If H > threshold (T), we consider there a scene change. The threshold can be adjusted to reduce over-detection and un-detection. Fixed threshold cannot perform equally well for all videos. Therefore, it must be dynamically assigned to tolerate variations in individual frames, while still ensuring a desired level of performance. Dynamic threshold can be determined by the minimum of the difference among color histograms. We apply one-dimensional entropic thresholding to histogram difference to find the optimal.

After shots are segmented, key frames can be extracted from each shot. Depending on the content complexity of the shot, one or more key frames can be extracted for a single shot [6].

Since we need to demonstrate how to extract and store the data from scene change detection into XML, we choose an approach that is less complex. Thus, we engage shot boundary based approach to get the key frame of each shot. A sample result is shown in Figure 2.

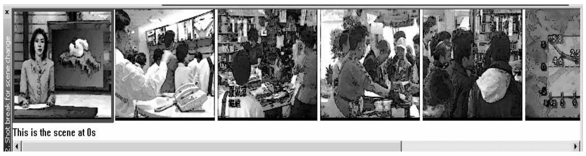


Figure 2. Shot breaks obtained in XVIP

3.2. Video optical character recognition

VOCR is a technique that can greatly help to locate topics of interest in a large digital news video archive via the automatic extraction and reading of captions and annotations [7]. Compared with VOCR from document images, caption extraction and recognition in news video presents new challenges:

- Complex background
- Low resolution of characters

It can be observed that the two frames shown in Figure 3 contain complex background and low resolution of characters.

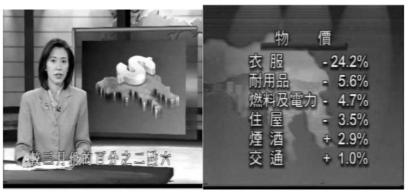


Figure 3. Two frames extracted from the digital news with characters on them

The main steps of VOCR for digital news library are as follows:

1. **Detection of Text Region:** for locating the caption from the video frame.
2. **Image Enhancement:** for increasing the resolution of each caption and reducing the variability in the background.
3. **Character Recognition and Segmentation:** For recognizing characters and performing segmentation between characters.
4. **Post-Processing:** for improving the recognition rate.

3.3. Detection of text region by VOCD

Before VOCR, VOCD has to be performed. A typical text region can be characterized as a horizontal rectangular structure of clustered sharp edges, because characters usually form regions of high contrast against the background. For our system, we have implemented a text region detection algorithm, which works well for both English and Chinese characters.

We select frames and extract regions that contain textual information from the selected frames. For extraction of vertical edge features, we apply a horizontal differential filter to the entire image with appropriate binary threshold. If a bounding region, which is detected by the horizontal differential filtering technique, satisfies size factor, fill factor and horizontal-vertical aspect ratio constraints, it is selected for recognition as a text region. Detection results are selected by their location to extract specific captions, which usually appear at the lower positions in frames. Figure 4 shows typical VOCD results by XVIP.

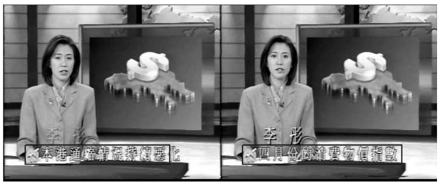


Figure 4. Rectangular boxes highlight news captions.

4. XML

Most existing digital video libraries use relational database system to store the data extracted. XVIP, on the other hand, engages XML for storing the extracted data. XML is a meta-language that permits a set of users to create their own mark-up language for describing the contents of Web documents. Since an XML file contains not only data but also metadata, i.e., structural and semantic information about that data, an XML document is very similar to a database.

4.1 Reasons for using XML

As XML has so many advantages, we choose XML for storing the data extracted from videos. The major factors that we think XML is suitable for our tool instead of other database systems are:

- XML's flexibility allows it to serve as a meta-language for defining other markup languages specialized for specific contexts. For each modality dimension, it manages its own XML DTD (Document Type Definition) or XML schema. The DTD would be extended, if knowledge enrichment or cross-referencing were applied.
- XML can describe the information internally, so contents prepared in XML can be searched easily and accurately, and only the related result is shown to the searcher. The extracted information is thus useful and meaningful.
- XML is scalable: We can add more information without affecting others. As XVIP is a multi-modal system, more components can be added to XVIP if a new extracting information technique is developed. In the following XML, it contains both the VOCR and scene change modalities.

```
<MODALITY NAME="VOCD">
  <TIME VALUE="1">
    <TEXT>
          Hong Kong is a beautiful place.
          <GEONAME>Hong Kong</GEONAME>
    </TEXT>
  </TIME>
  <TIME VALUE="5">
    <TEXT> the best clothing </TEXT>
  </TIME>
</MODALITY>
<MODALITY NAME="SCENE CHANGE">
  <TIME VALUE="0" SRC="a1.jpg"/>
  <TIME VALUE="5 SRC="a5jpg"/>
</MODALITY>
```

- XML can be displayed in different ways. As its style sheets are separate, XML allows users to download a document and then display it in different platforms. Extensible Style-sheet Language (XSL) expresses rules that indicate how to transform an XML document to a presentation format such as HTML or SMIL.

4.2. Schema

The process of designing an XML document for the information extracted from a video requires several element tags as schema representation of the document content and structure. This process starts with choosing a vocabulary to describe all the required aspects of the extracted video content. For example, "video" can be chosen for representing the video information, and "time" for the time that the information is extracted from the video. The next step is to show relationship between vocabulary entries, or in other words, to create ontology of the video information domain. A common way of doing this is using UML (Unfilled Markup Language) for its diagrams and notations.

The UML-based ontology is the initial native visualization of the content of the XML documents to be created. It is a thorough conceptual model that captures a human view of the domain, its object (elements), properties (attributes) and relations. The UML-based ontology of extracted video information by XVIP is shown in Figure 5.

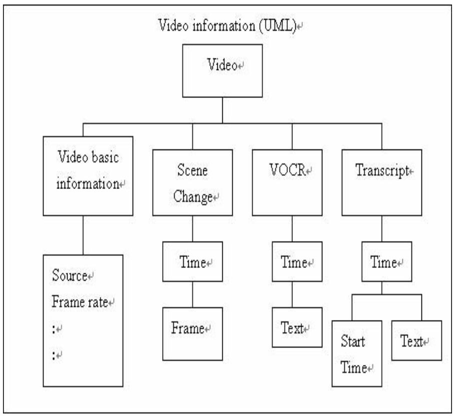
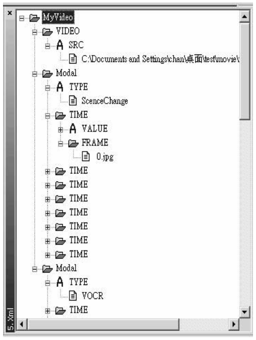


Figure 5. UML-based ontology

4.3. Presenting XML in XVIP

XVIP interface is based on the XML DOM (Document Object Model) visualization, generalized and enhanced to make it more suitable for effective visual interaction with metadata. Basically, it is a part of a DOM tree where all logical and structural components of an XML document are represented as nodes. Its main difference from a DOM tree is that XVIP borrows only the recurring metadata structure from a standard DOM.



```
MyVideo
  VIDEO
    SRC
      C:\Documents and Settings\chan\桌面\test\movie\
  Modal
    TYPE
      ScenceChange
    TIME
      VALUE
      FRAME
        0.jpg
    TIME
    TIME
    TIME
    TIME
    TIME
    TIME
    TIME
  Modal
    TYPE
      VOCR
    TIME
```

Figure 6. XML representation as a tree in XVIP

The DOM tree model becomes very similar to an XML schema. On the other hand, the DOM tree model inherits the structural organization of a DOM rather than of a schema. All types of metadata (elements, attributes and text) are represented as nodes that show element/attribute names or the text content and their relative places within the XML document structure. Each attribute has the status of a special child element of the parent element. The overall presentation of XML in XVIP is illustrated in Figure 6.

5. Knowledge enrichment

After the XML is built from XVIP, secondary information can be extracted for more video understanding. Typical digital video libraries contain a large number of videos. A good query engine is not sufficient because often the candidate result sets grow in number as the library grows. Interfaces to browse both the library and the defined library subsets such as the results from a query become increasingly important. The ability to extract names of organizations, people, locations, dates and times is essential for correlating occurrences of important facts, events, and other metadata in the video library [8], and is central to the production of information collages.

5.1. Extracting video geographic information

The transcript provides primary source of geographic information, though there are other sources, such as VOCR. Extracting geographic information from videos begins with the text metadata as the source material to be processed, and a geographic database. The text metadata, which associates text with video times, are then matched with the terms in the geographic database, which maps geographic text terms to the identified longitude and latitude.

5.2. Example scenario

XVIP extracts the names of major cities mentioned in the video. The names of major cities and their details are stored in an XML file. The following is a sample XML input file used in XVIP:

```
<MAJORCITY>
      <CITY ID=0>
            <NAME>上海 (Shanghai)</NAME>
            <COUNTRY>中國 (China)</COUNTRY>
            <LONGITUDE>31 15N</LONGITUDE>
              <LATITUDE>121 26E</LATITUDE>
        </CITY>
        <CITY ID=1>
              <NAME>北京(Beijing)</NAME>
              <COUNTRY>中國(China)</COUNTRY>
              <LONGITUDE>39 55N</LONGITUDE>
              <LATITUDE>116 20E</LATITUDE>
        </CITY>
        <CITY ID=2>
              <NAME>香港 (Hong Kong)</NAME>
              <COUNTRY>中國(China)</COUNTRY>
              <LONGITUDE>22 11N</LONGITUDE>
              <LATITUDE>114 14E</LATITUDE>
        </CITY>
</MAJORCITY>
```

The geographic information extraction process establishes a relationship between the video and the place names. As an example, we extract the names of major cities from the text obtained from VOCR, and show the following part of the XML output file generated from XVIP:

```
<MyVideo>
<VIDEOSRC="C:\movie\news2\atvnews02.mpg"/>
<MODALITY NAME="VOCR">
    <TIME VALUE="1">
        <TEXT>
        The reporter is speaking    ...
        </TEXT>
    </TIME>
    <TIME VALUE="34">
        <TEXT>
        A flight flying from 西雅圖 (Seattle) to 華盛頓
(Washington) ....
            <CITY>西雅圖 (Seattle) </CITY>
            <CITY>華盛頓 (Washington)</CITY>
        </TEXT>
    </TIME>
    <TIME VALUE="38">
        <TEXT>
        There is an accident    ...
        </TEXT>
      </TIME>
</MODALITY>
</MyVideo>
```

5.3. Enrichment of the video information

In our XML file, the whole content of VOCR, including the text, the time references, and the city names extracted from the text, are included inside the tag "MODALITY". This indicates that VOCR is only one of the modality names used in XVIP. There can be a number of different modality names. For example, we can have a modality name called "Speech Recognition" to recognize voices in the video. The result of speech recognition can be represented in a way similar to that of VOCR.

XVIP uses the text extracted from VOCR and Speech Recognition, and compares them with the names of major cities in its database. The database may come from another XML files or some other sources. Also, the knowledge that can be extracted is not limited to a particular type of information. Whenever a new database is imported, new information, e.g., names of political people, international organizations, etc., can also be extracted from the video.

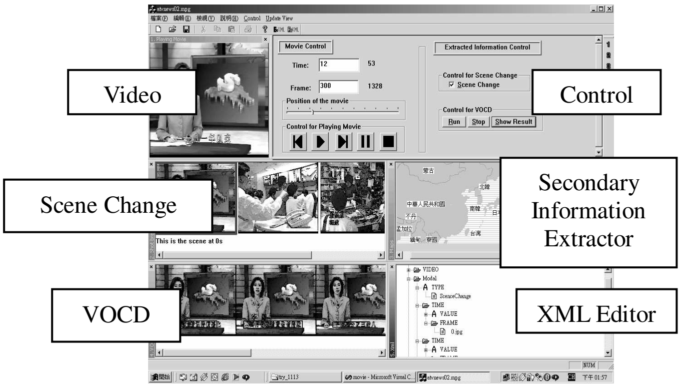
6. Implementation

The multi-modal XVIP system consists of several parts and each of them is implemented as a separate View interface in a docking window. View is a class in Visual C++ for displaying content, including text, picture, dialog box, ActiveX control and new software.

There are six main components in XVIP:

1. Modified Video Player from DirectShow
   It supports a wide variety of video formats, and provides control functions for playing video.
2. Control
   It provides functions for importing video and XML, exporting XML, and controlling docking windows.
3. Scene Change
   It extracts shot breaks from the video, and provides functions for inserting new shots, deleting current shots, or resizing the shot breaks to be displayed.
4. Video Optical Character Detection (VOCD)
   It locates the captions and characters on the videos and highlights them with rectangle boxes.
5. XML Editor
   It displays an XML file and allows user to edit.
6. Secondary Information Extractor
   It extracts specific information from the video.

The overall view of XVIP interface is illustrated in Figure 7.



Movie Control

Time: 12 53

Frame: 300 1328

Position of the movie

Control for Playing Movie

Extracted Information Control

Control for Scene Change

☑ Scene Change

Control for VOCD

Run Stop Show Result

This is the scene at 0s

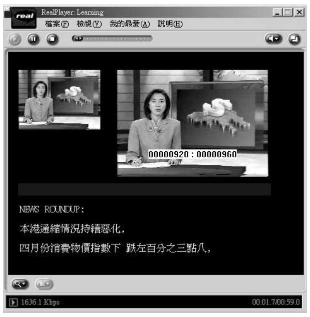
```
VIDEO
Modal
  TYPE
    ScenceChange
  TIME
    VALUE
    FRAME
      0.jpg
  TIME
    VALUE
```

Figure 7. Overall view of the interface of XVIP

The XML-enriched video data can then be presented in a Web browser using XSLT [9] or emerging presentation format like SMIL [10] at user's discretion.

SMIL is a recommendation from the W3C. It is an XML extension and it provides an author with the capability to synchronize the playback of all multimedia elements in a way to take into account the connection speed. These advantages make SMIL a good format for presenting multimedia. XVIP is able to transform the generated XML format output file into a SMIL format for presentation.

We implemented an XML to SMIL transformer in XVIP. It is able to generate different presentation templates of SMIL with results from knowledge enrichment process. Basically, XSLT will be used for the first phrase of the transformation. However, due to some special features of SMIL, a post-processor is added to complete the whole transformation. A snapshot of SMIL presentation for the digital library contents generated by XVIP is shown in Figure 8.



NEWS ROUNDUP:

本港通縮情況持續惡化,

四月份消費物價指數下 跌左百分之三點八,

Figure 8. Snapshot of SMIL presentation

7. Conclusion

We develop a system with multiple functions for processing and storing information extracted from video files. The system, XVIP, integrates different video information extraction techniques with a multi-modal concept, generates an XML file for storing the extracted information, and allows users to perform editing on the XML file exported from the system. The multi-modal concept and XML format give users a flexibility to integrate new techniques and to perform further enrichment or modification on the video data in the future. Finally, XVIP comprises an XML to SMIL transformer to generate different SMIL templates for a seamless multimedia presentation over the Internet environment.

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 (Project No. CUHK4222/01E), and the Hong Kong Innovation and Technology Fund (Project ID ITS/29/00).

References

[1] "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library." Wactlar, H., Christel, M., Gong, Y., Hauptmann, A., *IEEE Computer*, 32(2): 66-73,Feb 1999.

[2] "The Digital Video Library System: Vision and Design." S. Gauch, R. Aust, J. Evans, J. Gauch, G. Minden, D. Niehaus, J. Roberts. *Digital Libraries '94*, June 19-21, 1994, College Station, TX, ed. J. Schnase, J. Leggett, R. Furuta, T. Metcalfe, pp. 47-52.

[3] "New Directions in Video Information Extraction and Summarization." Wactlar, H., *10th DELOS Workshop*, Santorini, Greece, June 24-25, 1999.

[4] "Performance Evaluation of Scene Change Detection Algorithms", Jong Whan Jang and I1 Kyun Oh, Spring 1997.

[5] "Adaptive Key Frame Extraction Using Unsupervised Clustering", Yueting Zhuang, Yong Rui, Thomas S. Huang and Sharad Mehrotra, *IEEE International Conference on Image Processing, 1998 (ICIP98),* Oct. 1998.

[6] "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption,",T. Sato, T. Kanade, E. Hughes, M. Smith, and S. Satoh, ACM Multimedia Systems Special Issue on Video Libraries, February, 1998.

[7] "Video OCR for Digital News Archives,",T. Sato, T. Kanade, E. Hughes, and M. Smith, *IEEE International Workshop on Content-Based Access of Image and Video Databases*, January, 1998, pp. 52 - 60.

[8] "Multi-Document Summarization and Visualization in the Informedia Digital Video Library." Wactlar, H., *New Information Technology 2001 Conference (NIT 2001),* Tsinghua University, Beijing, May 29-31, 2001.

[9] "XSLT for Tailored Access to a Digital Vied Library", Micheal G. Christel, Bryan Maher, Andrew Begun, JCDL 2001, June 2001.

[10] "Towards Second and Third Generation Web-based Multimedia" Jacco van Ossenbruggen, Joost Geurts, Frank Cornelissen, Lynda Hardman and Lloyd Rutledge, *10th International World Wide Web Conference*. 1-5 May 2001.