A Wireless XML-Based Handheld Multimodal Digital Video Library Client System

Sam K. S. Sze, Henry K. P. Choi, Michael R. Lyu
{samsze, kpchoi, lyu}@cse.cuhk.edu.hk
Computer Science and Engineering Department
The Chinese University of Hong Kong
Shatin, Hong Kong

ABSTRACT

We have developed a client system for accessing a multimodal digital video library (DVL), namely, *iVIEW*. It provides a user interface that meets the challenge of rich multimodal information presentation on wireless handheld devices. An XML schema is employed to organize the multimodal metadata. Furthermore, we investigated a context awareness mechanism complementary to the XML schema to facilitate scalable degradation under restricted resources in wireless application environment.

1. INTRODUCTION

Video is a content-rich medium for perception of information. Modern DVL systems [1] extract and integrates multimodal information including transcripts, key-frames, on-screen text and geographical locations. With the evolution of Internet, [2] further proposes related techniques for DVL deployment over Internet. The challenge of an increasing demand of accessing video content anytime and anywhere has emerged recently. We developed an intelligent video over Internet and wireless (*iVIEW*) DVL to meet this challenge.

In general, the main issues come from the inherited physical limitation of wireless devices, including limitation of screen size, CPU processing power, battery and input devices. Wireless device is also characterized by limited bandwidth. *iVIEW* wireless client, the client we have implemented, aims to make a compromise under these constraints. On the one hand, it provides an intelligent interface to handle sophisticated multimodal presentation and result set visualization. On the other hand, it utilizes the knowledge of the context to make adjustment of resources usage.

2. *iVIEW* SYSTEM OVERVIEW

The *iVIEW* DVL system consists of three major subsystems: Video Information Processing (VIP) System, Searching and Indexing System, and Client System. The overall architecture is shown in Figure 1.

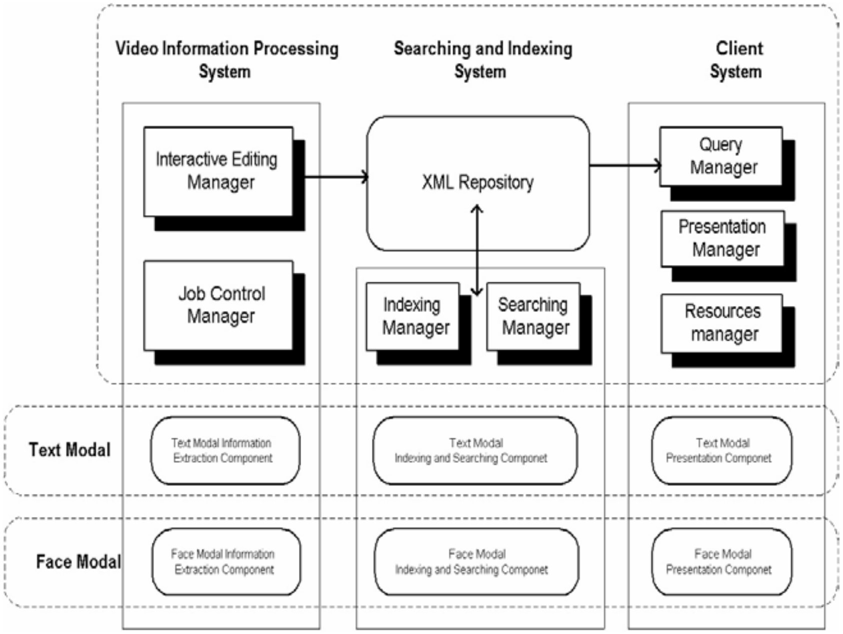
Video Information Processing System
Searching and Indexing System
Client System
Interactive Editing Manager
XML Repository
Query Manager
Presentation Manager
Job Control Manager
Indexing Manager
Searching Manager
Resources manager
Text Modal
Text Modal Information Extraction Component
Text Modal Indexing and Searching Componet
Text Modal Presentation Componet
Face Modal
Face Modal Information Extraction Component
Face Modal Indexing and Searching Componet
Face Modal Presentation Componet

Figure 1 - Overall System Architecture

We define modality as a domain or type of information that can be extracted from the video. The VIP system extracts multimodal information from video segment that includes transcript, on-screen text, human face, text abstract, key-frame, topic and geographical locations. An XML schema is designed to integrate and describe the extracted metadata of a video segment. And, the set of XML metadata descriptions are stored in an XML repository for queries by the client system. Currently, *iVIEW* system contains approximately 35 hours Chinese video and 100 hours English video and multimodal information extracted from the video.

3. *iVIEW* WIRELESS CLIENT ARCHITECTURE

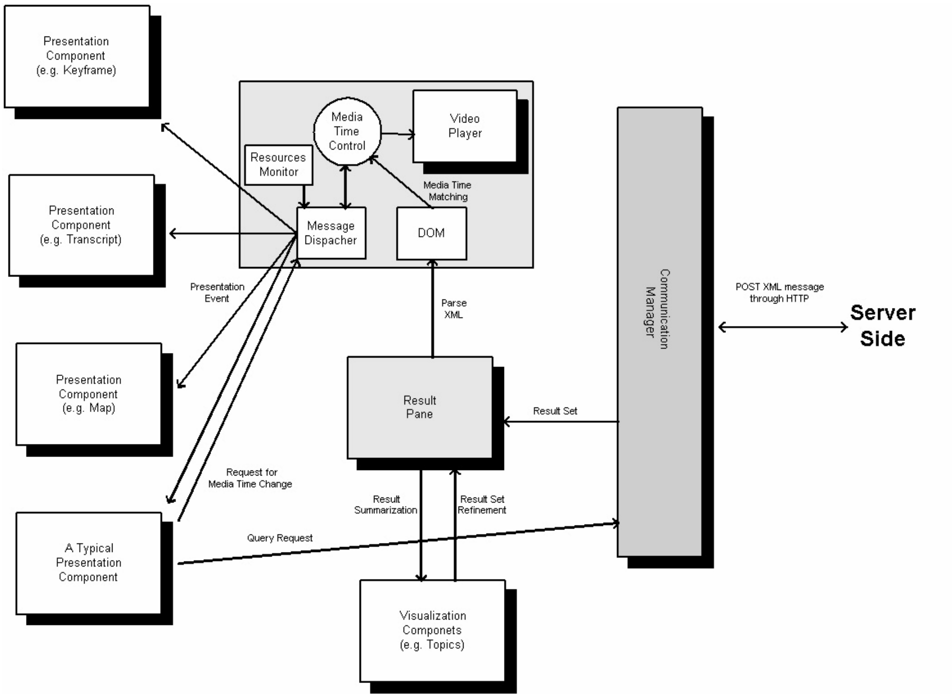
Presentation Component (e.g. Keyframe)
Media Time Control
Video Player
Resources Monitor
Media Time Matching
Presentation Component (e.g. Transcript)
Message Dispacher
DOM
Communication Manager
Presentation Event
Parse XML
POST XML message through HTTP
Server Side
Presentation Component (e.g. Map)
Result Pane
Result Set
Request for Media Time Change
Result Summarization
Result Set Refinement
Query Request
A Typical Presentation Component
Visualization Componets (e.g. Topics)

Figure 2 - iVIEW Client System Architecture

The *iVIEW* wireless client is a component-based system (Figure 2) composed of infrastructure and presentation components. The component-based approach makes the system scalable to support a potential expansion of additional modal dimensions.

This wireless client was developed on iPAQ H3630. The wireless connection permits the use of 802.11 PCMCIA, Bluetooth Compact Flash card, Nokia Card Phone 2.0 supporting GSM HSCSD or CDMA data modem cards.

The client-server communication message is coded in XML through HTTP POST to bypass firewalls blocking [3]. Although it does not conform to XML query standard, the message in XML is already self-explanatory. Once a search result is attained, the multimodal description in XML [4] is retrieved from the server. Results in XML are parsed using Document Object Model (DOM).

Media time, which is obtained as part of the result, provides a time cue for the message dispatcher to synchronize the multimodal information on the client. This media time is extracted with the time-related content to allow a unified presentation of video content at the client.

4. CLIENT USER INTERFACE

For content retrieval from a DVL, we define three phases of interactions: query, result set visualization and presentation. The client now supports two modes of query: text query (Figure 3a) and geographical location query (Figure 3b). For text query, keywords for text query are accepted. For the geographical query, user can drag a rectangular box to highlight the searching area. For each matched video segments returned by the search engine, a text abstract and a poster frame are then presented (Figure 3c). To facilitate result refinement, a visualization aid is developed. Each result element is pre-assigned with associated topics. The relation is displayed spatially in our user interface (Figure 3d). The topics are arranged as text tag in a circular shape. A point within the circle represents a video segment. The spatial displacement of a point is related to its semantic closeness to each topic. The users can then refine to their interested result points.

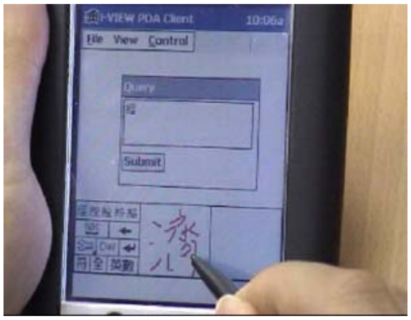

Figure 3a - Query by Text

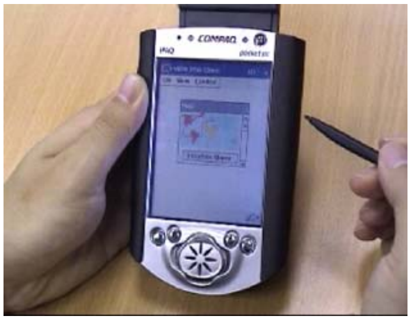

Figure 3b - Query by Map

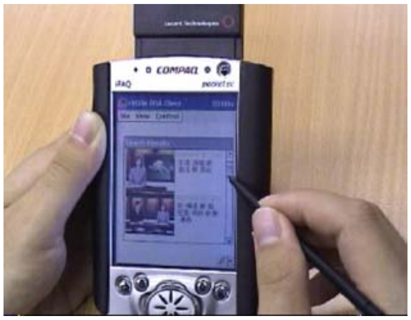

Figure 3c - Result Abstract

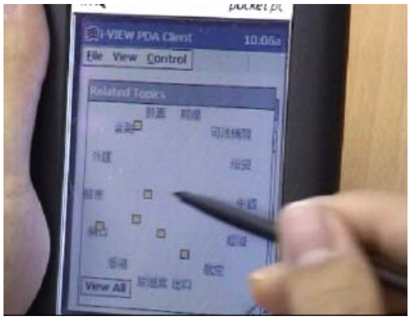

Figure 3d - Visualization by Topics

In the presentation phrase, the XML description of the selected video segment is retrieved. The user interface is presented in a multi-window manner (Figure 4). Each window presents information of a particular modality. The user can arrange the windows and select the interested modality in one's own fashion.

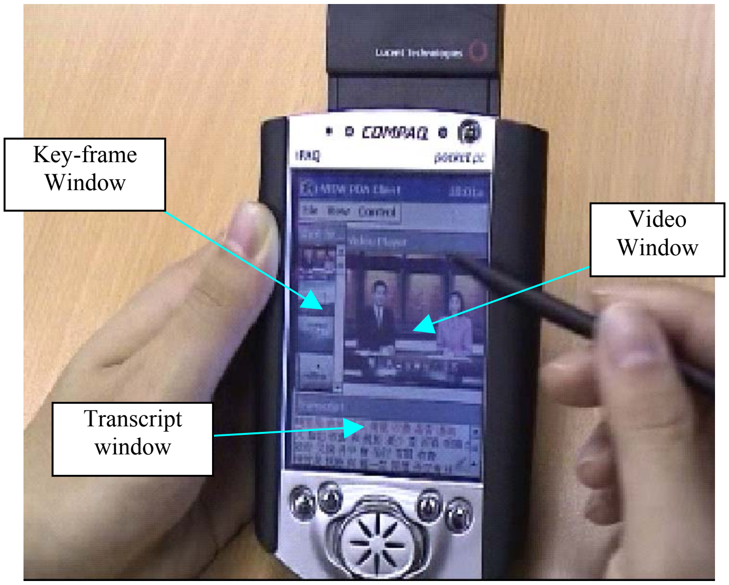

Figure 4 - Multimodal Presentation in Multi-window

5. CLIENT SCALABILITY MECHANISM

Figure 5 shows a multimodal presentation description in XML, where a particular video modal context type is assigned to each video. With a clearly specified context types, the client system can be aware of the overall context and individual modal context being presented.

In a multimodal presentation, different modalities involve different level of consumption of system resources. As the modalities are complementary information extracted from the video source. The reduction of a modality may cause degrade in overall content. Meanwhile, portion of the core content can still be preserved depends on the significance of the modality being removed.

```
<?xml version="1.0" encoding="big5" ?>
   <sequence path="/iview/video/">
     <time start="0">
        <script> AFTER A 15 YEAR SUSPENSION,  </script>
        <frame file="frame141_01.jpg" />
   </time>
     <time start="6">
        <script> MARGARET LOWRIE HAS THE DETAILS FROM LONDON
        </script>
        <map>LONDON</map>
   </time>
   . . . . .
 </sequence>
```

Figure 5 - XML for Multimodal Presentation Description

To achieve such presentation scalability, a set modal significance order chain (e.g. $S(m_1) > S(m_2) > S(m_3) > \ldots$) should also be defined for different video types. The *iVIEW* wireless client system embeds a performance monitor that keeps track of system resources including CPU, memory and bandwidth utilization. If system utilization saturates, the dispatcher will be signaled to make a decision to stop dispatching an active modality according to:

$$\max \sum_{i=1}^{n} S_k(m_i)x_i \mid \sum_{i=1}^{n} R(m_i)x_i = R_{available} \quad where\ x \in \{0, 1\},\ i=1\ldots n$$

m_i is a particular modality
$S_k(m_i)$ is the significance of a modal presentation in video type k
$R(m_i)$ is the resources utilization of a modal presentation
$R_{available}$ is the total available resources

As a result, the modality is inactivated. This process repeats until the system obtains enough resources. Inversely, an inactivated modality can be re-activated when the system gets enough resources.

6. CONCLUSION

A wireless DVL client on a handheld is implemented with the capability to handle searching, result visualization and presentation of multimodal video information. It equips a scalability mechanism to achieve graceful presentation degradation in wireless environment.

7. ACKNOWLEDGEMENT

The work described in this paper was fully supported by two grants from the Hong Kong Special Administrative Region: the Hong Kong Innovation and Technology Fund (under Project No. ITS/29/00) and the Research Grants Council (under Project No. CUHK4222/01E).

8. REFERENCES

[1] H.D. Wactlar, T. Kanade, M.A. Smith, S.M. Stevens. “Intelligent Access to Digital Video: Informedia Project,” *IEEE Computer*, volume 29, issue 5, May 1996, pp. 46-52.

[2] H.D. Wactlar, “Informedia – Search and Summarization in the Video Medium,” *Imagina 2000 Conference,* Monaco, January 31 - February 2, 2000.

[3] W. H. Cheung, M. R. Lyu, and K.W. Ng, “Integrating Digital Libraries by CORBA, XML and Servlet,” *Proceedings First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia, June 24-28 2001, pp. 472.

[4] M. Christel, B. Maher, and A. Begun, “XSLT for Tailored Access to a DVL,” *Joint Conference on Digital Libraries (JCDL '01)*, Roanoke, VA, June 24-28, 2001, pp. 290-299.