A Wireless Handheld Multi-modal Digital Video Library Client System

Michael R. Lyu
Computer Science Dept.
Chinese University of Hong Kong
lyu@cse.cuhk.edu.hk

Jerome Yen
System Eng. Dept.
Chinese University of Hong Kong
jyen@se.cuhk.edu.hk

Edward Yau
VIEW Laboratory
Chinese University of Hong Kong
edyau@cse.cuhk.edu.hk

Sam Sze
VIEW Laboratory
Chinese University of Hong Kong
sequence@netvigator.com

ABSTRACT
We developed technologies for transmitting video contents over wireless platforms, and encapsulated these video delivery and presentation technologies into a client system for accessing a multi-modal digital video library. The mobile access system, *iVIEW client*, provides a user interface that meets the challenge of rich multi-modal information presentation on wireless hand-held devices. An XML schema is employed to organize the multi-modal metadata for better data interoperability. Furthermore, we investigated a context awareness mechanism complementary to the XML schema to facilitate scalable degradation under restricted resources in wireless application environment. This paper presents the design, implementation, and evaluation of the *iVIEW* system and its associated technologies for video information management and delivery on pervasive devices over wireless networks.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *evaluation/methodology, hypertext navigation and maps, video.*

General Terms
Management, Measurement, Documentation, Performance, Design, Human.

Keywords: Mobile Applications, Multi-modal Content Retrieval, Browser and Interface on Mobile Devices, Multimedia Management and Support, Multimedia Information Retrieval, XML.

1. INTRODUCTION

Video represents rich media content. Evolution of digital video library (DVL) enables people to search and access rich video content. The techniques involved in composing videos into vast DVL for content-based retrieval are provided in the literature [1-3]. The development of DVL includes video information extraction [4]. In additional to the video itself, brilliant multi-modal information can be in company, making the presentation of the video delivery system much more prosperous. The information that can be extracted from a video includes video streams, scene changes, camera motions, text detection [5], face detection [6], object recognition, geo-coding [7], word relevance statistics, transcript generation, and audio level tracking. We integrated these technologies in an interactive Video over InternEt and Wireless (*iVIEW*) system [8].

The evolvement of Internet enables users to surf DVL through browsers or dedicated Internet applications. Related issues involving design and implementation of an Internet access digital video library system are discussed in [9, 10, 11]. With the evolving of wireless devices like mobile handsets and PDAs, the realization of a handheld wireless client for digital video library becomes a challenge [12]. Users demand to access digital video library anytime, anywhere, with any devices through the mobile networks.

The main issues of wireless applications, however, come from two major aspects: the resource constraints inherited from the hardware of handheld devices, and the bandwidth limitation and instability. In designing a client software system for rich video information access and presentation in the wireless environment, we are then facing two major challenges. First, we require an intelligent user interfaces in the source-limited handheld devices. The user interface should be able to support the following features: (1) Enable users to *select contents* to be presented in different scales, from coarse-grain (e.g., to view the whole scenario in full-screen) to fine-grain (e.g., to grasp a specific media). (2) A large digital video library usually returns lots of results per query. User-friendly interface and schemes for data organization *refinement* should be designed to support seamless navigations. (3) Integrates *hand manipulation techniques* into overall user operation habits. We will describe the design of our user interface that meets these challenges.

Second, in a presentation session that synchronizes multiple media under a fluctuating wireless environment, we need a control scheme that manages the CPU, memory resources, and bandwidth utilizations. We will present our proposed XML schema that facilitates a scalable degradation in multi-modal presentation. We also describe the related mechanism that leads to context awareness. The mechanism involves a mini system monitor embedded in our client system to keep track of CPU consumption, memory allocation, and bandwidth usages.

2. iVIEW SYSTEM OVERVIEW

2.1 Overview

MIR'03, November 7, 2003, Berkeley, California, USA.

The *iVIEW* system is a multi-modal and multilingual DVL. Figure 1 shows the overall architecture of the *iVIEW* system. The system is composed of three major subsystems: Video Information Processing (VIP) Subsystem, Searching and Indexing Subsystem and Client Subsystem. We define *modality* as a domain or type of information that can be extracted from the video. Examples shown in Figure 1 are the text generated by speech recognition and the human identity by face recognition.

There is no theoretical upper limit regarding the amount of contents that can be stored in the *iVIEW* system. Currently, the system stores Chinese news videos and English news videos and multi-modal information extracted from the video. The system manipulates text in both Chinese and English. The details of these three major subsystems and their associated modality processing techniques are discussed in the following sub-sessions.

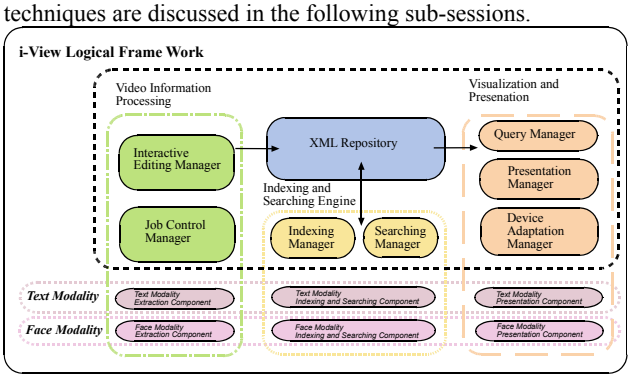

Figure 1. The *iVIEW* Logical Framework

2.2 Video Information Processing Engine

The Video Information Processing (VIP) Subsystem handles multi-modal information extracted from a video file.

The multi-modal information is organized in an XML format. The VIP Subsystem processes video in two modes. An offline mode coordinated by the Job Control Manager schedules video recording and launches jobs to process the video file offline. An online interactive mode provides a user interface for a content editor to monitor the process and view the results. Therefore, human intervention and correction of the data is available.

The VIP Subsystem currently extracts various multi-modal information of a video:

- transcript extraction by speech recognition
- keyframes extraction by scene change detection
- video text abstract from transcript
- topic assignment by transcript
- geographical locations extraction from the words of the transcript
- on-screen characters recognition

2.3 Multi-modal Indexing and Searching

The Indexing and Searching Subsystem is responsible for video information indexing and searching. A file containing an XML format structure that describes and associates multi-modal information is produced from each video file. This XML file is indexed for multi-modal searching through the Indexing Manager. For each query, the search engine will return a set of XML files that match the query.

If we view the information flow in Figure 1 from left to right, multi-modal information (text, image, face, etc) can be extracted, processed, stored and then indexed. Information can be searched by individual modal dimension or a composite of multiple modal dimensions in logical relations. After the searching process, multi-modal information is presented to the end users.

Each modal dimension is processed, preserved and correlated with other dimensions throughout the end-to-end video processing. The *iVIEW* system is designed to apply a unified scheme for processing different modal dimensions. Therefore, we can add a new modal dimension to the whole system seamlessly, including extraction, processing, indexing, searching and presentation of the new modal dimension. This unique feature facilitates the integration of newly obtained techniques on evolving modal information for video processing.

2.4 Client Access

The client subsystem handles queries, result sets visualization, and delivers multi-modal presentations in time-synchronized manner. Detailed architecture of the client subsystem is discussed in the next section.

3. iVIEW CLIENT ARCHITECTURE

3.1 General Architecture

The *iVIEW* client (Figure 2) is a component-based subsystem composed of a set of infrastructure components and presentation components. The infrastructure components (shaded-color areas in Figure 2) provide services for client-server communications and time synchronizations among different presentation components by message passing. The presentation components (white areas in Figure 2) accept messages passed from the infrastructure components and generate the required presentation result. This component-based approach makes the system scalable to support a potential expansion of additional modal dimensions.

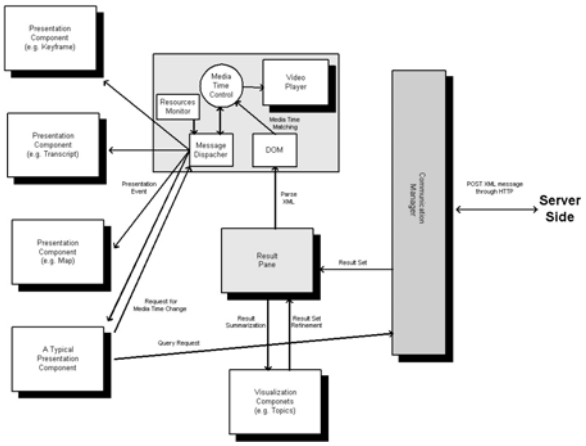

Figure 2. *iVIEW* Client System Architecture

The client-server communication message is coded in XML through HTTP. The XML is embedded into an HTTP POST message. Employing HTTP enjoys the advantage that the service is seldom blocked by firewalls [9]. It also facilitates the deployment of an application to content providers or data centers. Although it does not yet conform to XML Query standard [13], the message in XML is already self-explanatory. Once a search result is attained, the multi-modal description in XML is obtained from the server. The client parses the XML using Document Object Model (DOM). The infrastructure gets the media time through playing the video. The recorded media time is then matched with the media description to seek the event that a presentation component needs to perform at a particular time frame.

The message dispatcher sends out media events to different presentation components according to a component registry. The component registry records the presentation components that the client system runs. Consequently, video information is only processed once from VIP, while multiple deliveries of the video contents can be facilitated for different demands depending on various features of client devices and platforms.

Based on this architecture, we implemented our client system for different platforms in the following ways: (1) As a desktop client using Microsoft Windows native programming tools; (2) As a JAVA applet using Java XML Parser (JAXP), and Java Media Framework (JMF); (3) As a Web page using common Web development techniques and standards including HTML, CSS, DHTML and scripting languages; or (4) As a native PocketPC application on Compaq's iPAQ handheld device. We focus on elaborating the last approach as a solution on multimedia, multi-modal retrieval for pervasive devices.

3.2 Wireless Client Implementation

Taking software video decoding as a benchmark, it roughly requires a CPU of at least 150 MIPS processing power [14]. The processing power baseline constraints our device selection. We selected the Compaq's iPAQ H3630 [15] as our reference solution platform. The video is played in the streaming mode so the memory limitation does not cause a major bottleneck in our application. Different PC cards accomplish wireless connection of the device. The supported wireless cards include IEEE 802.11b wireless LAN PCMCIA card, Bluetooth Compact Flash card, Nokia Card Phone 2.0 supporting GSM HSCSD, and CDMA data modem card.

The wireless client is developed by Microsoft Embedded Visual Tools and with the DirectX Platform Adaptation Kit (DXPAK). We use DirectShow for streaming video files encoded in Windows Media Video Format (.wmv). It is a streaming format base on MPEG4. The encoded audio stream and video stream bandwidths are 10.2Kbps and 16Kbps, respectively.

4. CLIENT USER INTERFACE

4.1 User Interface Overview

The *iVIEW* wireless client user interface integrates various state-of-the-art techniques in multimedia information presentation and visualization. As shown in Figure 3, the *iVIEW* wireless client includes a set of mini-windows user interface. Each mini-window works as a user control for a modality presentation. User interface using window widgets have been widely accepted. The mini-windows can be selected and dragged by the pen input device inside the visible area or partly beyond the visible boundaries.

There is usually not enough space for tiling all the mini-windows in the visible area. However, the overlapping or out-of-boundary placement of these windows can facilitate the partial visualization of each modality. The windows moving capability also provides flexibility for the user to arrange the best viewing order according to the significance of a modality. There are pull-down menu options providing the open and close functions of a particular mini-window.

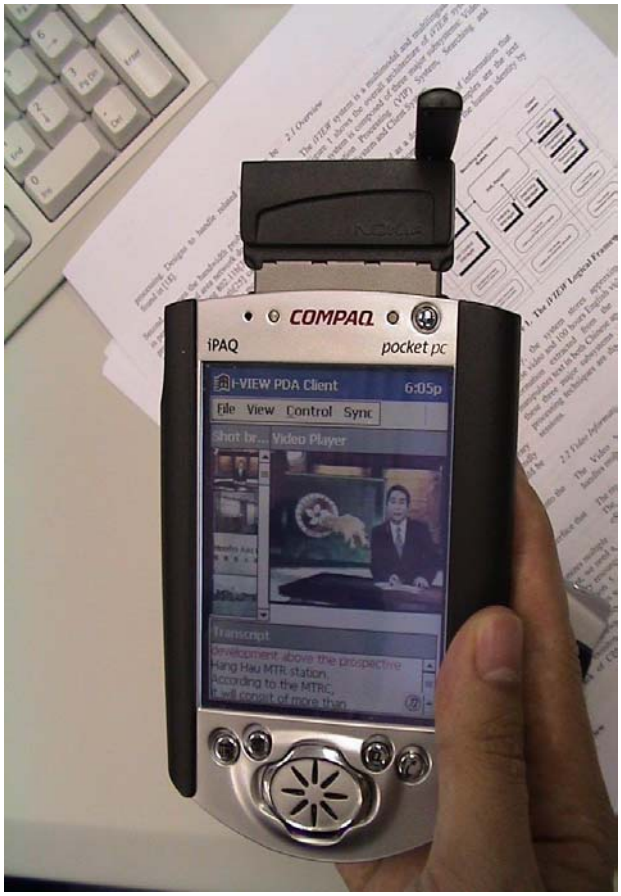

Figure 3. *iVIEW* Client Overview

The digital video library searching and presentation process is composed of three phrases, Query, Result Set Visualization and Presentation. We arrange each phrase into a particular interface with its own interaction behavior. In the following we illustrate each interface with screen captures.

4.2 Multi-modal Query Interface

The system supports two modalities of query. They are query by text and query by geographical location. In the text query window, a user can type in the keywords for text query. For English, the user

can input by using a small on-screen keyboard provided by the PocketPC environment or other input assisting software. For Chinese, we deploy the Gimsoft Chinese handwriting recognition embedded system [16].

For query by geographical location, a user can select the map modality to display a world map. Then, the user can drag a rectangular area on the map to indicate the area of interest, as shown in Figure 4. The query is then submitted by pressing the submit button. This kind of searching manner takes advantage of the pen interface with the handheld device.

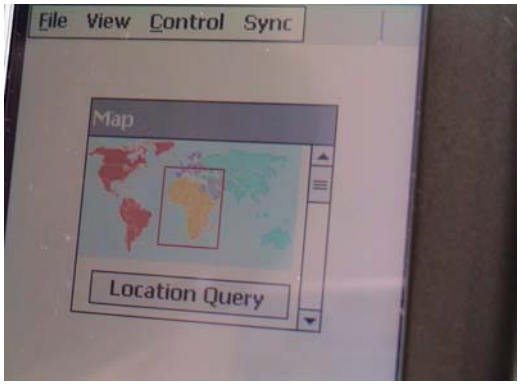

Figure 4. Query by Geographical Locations

4.3 Result Set Visualization and Manipulation

After the query is submitted, the search engine returns a set of matched results represented by the poster images and the text abstracts, as illustrated in Figure 5. Each result item represents a video segment that matches the query with the poster image of the closest keyframe that matches the highest hit query text. Text abstract is a digest of the video segment's transcript. It is done via the term frequencies and inverse document frequencies (TFIDF) techniques in the VIP process.

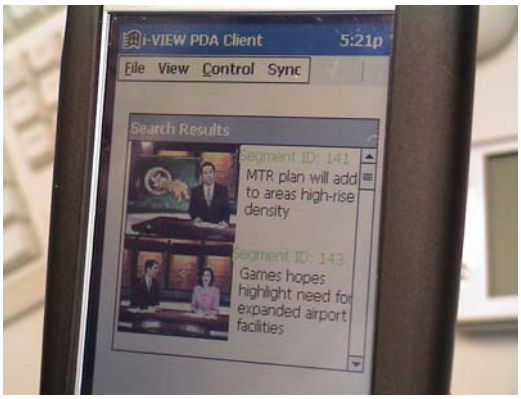

Figure 5. Query Result Set

Queries of such a large digital library usually returns with large result sets. We provide two interface techniques to facilitate users for refining the result set: visualization by tree and visualization by topics. The refinement is done by systematic classification of the existing result set into specific categories. In a thin client approach, the result set is resent to the search engine to collect category information.

iVIEW wireless client supports visual digest by tree for geographical locations. That is, the result set is classified into hierarchical structures of location names. User can refine their result set by digging into the interested geographical locations. Due to the hierarchical nature of geographical locations, the tree view fits the representation naturally.

The topics visualization is developed in a similar spirit as the Visualization by Example (VIBE) technique [17]. In topics visualization, each result element is assigned to one or multiple predefined single-level topics. Approaches on topics labeling similar to [18] are applied at VIP. The topics are the text tags arranged in a circular shape (Figure 6). Topics can be famous names, location names or general category keywords like "economic", "politic", "education" or "sports". A point within the circle represents a result. The spatial displacement of a point is related to its closeness to each topic. When the mouse is over a point, a floating tool-tip will appear to indicate the related topic of this point. A user can drag the mouse to highlight a rectangular area that contains the results that the user is interest in. The result set will then confine to the selected results.

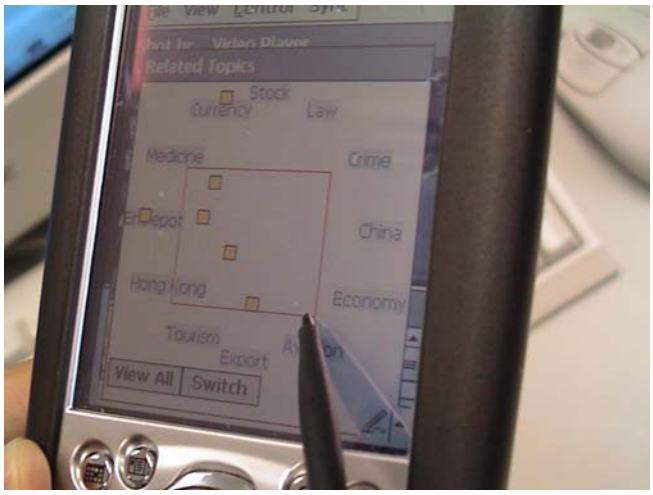

Figure 6. Result Set Visualization by Topics

4.4 Media Presentation

After a final result video segment selected by the user who clicks the poster frame, the media XML is fetched from the server. A sample multi-modal presentation description XML is listed in Figure 7. The media XML is the skeleton of the whole multi-modal presentation. It only contains text content and therefore the required bandwidth and download time for this XML under existing mobile network is relatively small.

```
<?xml version="1.0" encoding="utf-16" ?>
 <sequence path="/iview/video/">
    <time start="0">
       <script> GOVERNMENT HAS RESTORED FULL
             </script>
       <frame file="frame141_00.jpg" />
    </time>
```

```
<time start="2">
    <script> DIPLOMATIC RELATIONS WITH LIBYA
    </script>
</time>
<time start="4">
    <script> AFTER A 15 YEAR SUSPENSION. CNN'S
        </script>
    <frame file="frame141_01.jpg" />
</time>
<time start="6">
    <script> MARGARET LOWRIE HAS THE DETAILS
        FROM LONDON
    </script>
    <map>LONDON</map>
</time>
<time start="8">
    <script> BRITISH FOREIGN SECRETARY ROBIN
        COOK
    </script>
</time>
<time start="11">
    <script> ANNOUNCED DIPLOMATIC TIES WITH
        LIBYA WOULD BE RESTORED BECAUSE
        IT
    </script>
</time>
<time start="14">
    <script>IT NOW ACCEPTS RESPONSIBILITY
    </script>
    <frame file="frame141_02.jpg" />
</time>
. . . . .
</sequence>
```

Figure 7. Sample Multi-modal Presentation Description XML

The dispatcher reads the media XML stored in DOM and refers different media activities to different presentation components. They are various mini-windows as seen by users in Figure 8. The video window keeps displaying the video and sends audio output. Meanwhile, the keyframe and transcript frame automatically scroll to the corresponding presentation point. Keyframe has been shown to be an effective mean for video abstract [19]. Psychological practices indicate that such a captioned synchronization display enhances children's learning and improves accessibility [20].

Figure 8. Multi-modal Presentation in Action

All the modalities are presented in a synchronized manner. Direct skipping to a particular transcript line or keyframe is allowed. An user can scroll to certain point of a mini-windows, like the transcript window, click on a line, and the whole presentation session will automatically aligned to this point of the timeline.

In general all *iVIEW* presentation components support three general programming interface functions behind the scene:

- Initialization
  iVIEW client highlights all matched items when the video media initializes. For example, all matched text is highlighted.

- Passive Synchronization
  A particular piece of information will be highlighted when the media time is matched. For example, a specific transcript line is scrolled to a visible area when it is read.

- Active Synchronization
  Through a user action, a presentation component can change the media time to a particular point. For example, when a user clicks on a specific filmstrip, the video and other presentation components are re-synchronized to the time when that filmstrip occurs.

5. CLIENT SCALABILITY MECHANISM

5.1 Mechanism Overview

In a typical mobile network environment where data rate is unpredictable and the power of devices is limited, degradation in presentation quality often happens. Recent research like [21] proposed a scalable summarization scheme for text. We describe a mechanism inside the *iVIEW* wireless client that enables the degradation of multi-modal presentation according to the available resources.

We target to preserve optimal multi-modal content in scale with the available system resources. This scalability feature is accomplished by the XML multi-modal presentation description that leads to content awareness, and the component-based architecture of the client design.

We first review the XML schema for multi-modal presentation description. Then, we look into how the client system makes use of the XML to achieve scalability in degrading quality. In additional, a formal comparison of our proposed XML schema and the well-known Synchronized Multimedia Integration Language (SMIL) standard is discussed. We illustrate the design concerns and advantages of the *iVIEW* client through this contrast.

5.2 XML Schema for Multi-modal Presentation Description

Figure 9 describes the formal XML schema definition on the multi-modal presentation description XML shown in Figure 7. The XML format can facilitate scalable multi-modal presentations. In this XML schema, a particular video context type is assigned to each video. Moreover, the context type of each modality is specified. Therefore, the client system can be aware of the overall context and individual modality context being presented.

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
    <xsd:annotation>
        <xsd:documentation xml:lang="en">
```

```
        Schema for Multimodal Presentation escription
      </xsd:documentation>
    </xsd:annotation>
  <xsd:element name="sequence" type="SequenceType" />
  <xsd:complexType name="SequenceType">
      <xsd:element name="time" type="TimeType"
          maxOccurs="unbounded" />
      <xsd:attribute name="path" type="xsd:anyURI" use="required"
          />
      <xsd:attribute name="type" type="xsd:string" />
    </xsd:complexType>
  <xsd:complexType name="TimeType">
      <xsd:element name="script" type="xsd:string" maxOccurs="1" />
      <xsd:element name="frame" type="FrameType" maxOccurs="1"
          />
      <xsd:element name="map" type="xsd:string" maxOccurs="1" />
      <xsd:attribute name="start" type="xsd:unsignedInt" />
  </xsd:complexType>
  <xsd:simpleType name="FrameType">
      <xsd:attribute name="file" type="xsd:anyURI" use="required" />
  </xsd:simpleType>
</xsd:schema>
```

Figure 9. Multi-modal Presentation Description Schema

5.3 Scalability Mechanism Through Context Awareness

In a multi-modal presentation, different modalities involve different level of consumption of system resources. As different modal information offers complementary information to other modal information extracted from the video source, the reduction of a modality may cause degradation in the overall content. Meanwhile, portion of the core content can still be kept, depending on the significance of the modality being removed.

For example, in a documentary video, the core content can still be preserved if the audio is removed while the system continues to provide a transcript for the ordinary people. For a lecture video, the core content can be preserved if the video is removed while keeping the audio and pictures of slides running. A multi-modal presentation system can degrade gracefully based on its awareness of the media context being presented.

To achieve such multi-modal scalability, a presentation system should be responsive to the media context being presented. In our system, the XML tags indicate the context type explicitly. A set of modality significance order chain (i.e. $S(m_1) > S(m_2) > S(m_3) > \ldots$) can also be defined for different video types.

The *iVIEW* wireless client system embeds a performance monitor for the context awareness purpose. The performance monitor keeps track of system resources including CPU load, memory usage, and bandwidth consumption. Figure 10 shows the visible interface of the performance monitor as provided in our development version.

Via the component-based design, the message dispatcher passes data for a particular modal presentation to a corresponding presentation component (refer to Figure 7). The performance monitor couples with the dispatcher. If the system utilization saturates, the dispatcher will be signaled to make a decision to stop dispatching a set of active modalities of least significance. Mathematically,

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

Figure 10. Resources Monitor Interface

where
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, and
$R_{available}$ is the total available resources.

The dispatcher only sends data to the set of modalities m_i satisfying the above equation. As a result, the set modalities that cannot satisfy the equation will be inactivated. This process repeats until the system obtains enough resources. Inversely, an inactivated modality can be re-activated when the system gets enough resources.

In additional to supporting degradation in circumstances with unstable resources, this design can also deliver the same content over different wireless devices. A device has the capability to present one modality but may lack the other. Using this scalability mechanism, a device can present the supporting modalities selectively.

5.4 Comparison with SMIL

SMIL is a World Wide Web Consortium (W3C) recommendation that enables simple authoring of interactive audiovisual presentations [22]. Our XML schema for multi-modal presentation is similar to SMIL but with several advantages over the current SMIL standard and player implementations. Meanwhile, our XML schema enjoys simplicity due to its specific presentation nature.

At presentation, we classify each activity in a particular modality as a media object with specified duration time interval. An activity can be showing of a picture, a text string, a video clip, a song, etc. The corresponding objects are named, correspondingly, picture, text, video, song, etc.

In mathematical notation, we can define a media object as:

$$O_{m,t,s}.$$

where m is the modality type, t is the time interval of the presentation duration, and s is the spatial displacement and the area on the presentation device.

A presentation P during T is a set of media objects.

$$P_T = \{ O_{m,t,s}, \ldots \}$$

SMIL is a generic system that supports heterogeneous media object presentation in different timing combination. The *iVIEW* client, on the other hand, is a video-centered presentation system, where the video is the basic media object and its duration is the whole presentation duration.

Revisiting the media description XML in Figure 7, we employ context tags to represent media, such that the player is furnished with the knowledge of the media context. This capability facilitates:

- Intelligent object spatial displacement by the player

  Using SMIL, the author has to specify (or the language implicitly describes) the spatial displacement of a media object. A SMIL player cannot destruct the spatial relationship, as it does not possess this knowledge on the context.

  However, if the player knows the context of each media object then the users can set preference on their client side. In our case, it is not necessary for the author of the presentation to determine the spatial relationship. (i.e. *s* is not necessary). This flexibility makes the multi-window implementation possible.

- Scalable presentation for mobile device

  If the context of a particular media is known, the player can determine the priority of each media for presentation as discussed before.

- Generic Application Support

  In SMIL, if we want to present a map, say, Africa, we can make it through a picture. However, the client side may have installed with a map application. Without the knowledge on the context, bandwidth is wasted to transfer the picture file. With the knowledge on the context, we can simply send text "Africa" with context attribute "map" to the player and the player can launch the corresponding application to deal with the context. This can establish a standard for opening context support and in many cases, save bandwidth.

We can formulate support for each media modality as a function called Associate Media Application, *A(m)*

$$A(m) = \{Application\ that\ supports\ a\ particular\ modality\}$$

For example,

$$A(text) = \{\text{"MS-Word"}, \text{"WordPerfect"}, \text{"vi"} \ldots \}$$

Our client system therefore can handle the media object flexibly. If there is no associated application, the message dispatcher can just skip it. The system hence can be easily expanded to support additional modalities when they are added to the system.
Due to the benefit from the component-based system, our player not only can be synchronized through dragging the timeline, but also can be driven by context for fast forward or backward. As described in the previous section, we can locate a media object in a particular modality and force all other modalities to synchronize with it.

In the usual practice of using SMIL, presentation authors specify a spatial area for captions or images. Media elements are presented one by one in a slide-show manner. Since our presentation components contain advanced user interface controls like text box with scrollbar, the whole text content or series of images can be included in the control before the presentation time. This allows users to browse content before or after the presentation time *t*.

We may consider this as a super object $O_{m'}$ that includes all the same modality objects within the presentation. Namely,

$$O_{m'} = \{O_{mt}\} \ \ where\ t \in T$$

We can then combine our concept into the SMIL standard as an additional XML module.

5.5 System Evaluation

We evaluate the current *iVIEW* implementation under different networking environments through wireless connections including 802.11 access point, bluetooth access point, GSM HSCSD, and 2.5G CDMA. Through experiments with the resources monitor, we can obtain resources utilization profile for each modality. The collected parameter can then be fed to constraints for degradation determination.

We tested our degradation mechanism for news clips under scarce bandwidth, shown in Figure 11 At the time when network bandwidth consumption is saturated, the client system starts degradation by suspending the video. The degradation process starts at 25sec and the system resources consumption drops at 39sec. Under this degraded situation, the user can still listen to the audio with visual abstracts by keyframes.

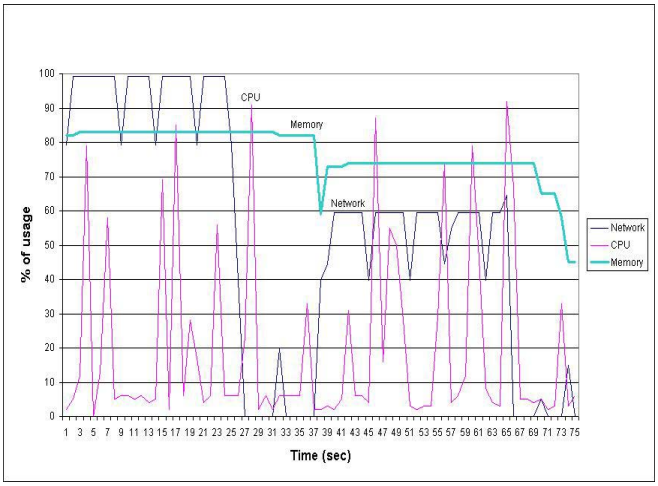

Figure 11. Transition Scenario at Video Suspension

Other experimental results indicate that the most resources demanding modality is video as predicted. The streaming video and audio provide acceptable visual quality in a 176 × 144 pixels mini-window. The *iVIEW* client is stable both in wireless local area networks and in public mobile networks. At stationary, the system runs smoothly at the full multi-modal scale with the 56kbps streaming video and keyframe file size of 1K bytes. In transport,

the scalable degradation process is triggered frequently due to inefficient bandwidth caused by fading, but transmission of the main modality remains uninterrupted. The *iVIEW* wireless client viewing system has been demonstrated with local news clips to 100 persons from different backgrounds. Only three persons responded negatively on the video visibility. Others were satisfied with the results.

6. CONCLUSION

A wireless DVL client system, which meets the challenges of handling complex multi-modal presentations in a handheld device, has been developed. We show a user interface design that allows flexible content selection, result set refinement, and multiple window views that fit the pen computing operation profile. The system is equipped with the capability of multi-modal presentation. We also define an advanced XML structure with its corresponding context-aware scalable degradation mechanism that optimizes the presentation content under limited resources and fluctuating environments. Experiments have been performed in real local-area and wide-area wireless networks to demonstrate the efficiency and applicability of the system.

7. ACKNOWLEDGEMENT

The work described in this paper is fully supported by a grant from the Research Grants Council (Project No. CUHK4182/03E) and by a grant from the Innovation and Technology Commission (Project ITS/00/029), the Hong Kong Special Administrative Region, China.

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, pp. 46-52, May 1996.

[2] M. Christel, A. Warmack, A. Hauptmann, and S. Crosby, "Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library," *IEEE Advances in Digital Libraries Conference 1999*, Baltimore, MD. pp. 98-104, May 19-21, 1999.

[3] Huiping Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video" *IEEE Transactions on Image Processing*, Volume: 9 Issue: 1 , Jan. 2000, Page(s): 147 –156.

[4] C.H Ngai, P.W. Chan, E. Yau, and M.R. Lyu, "XVIP: An XML-Based Video Information System," *Proceedings 26th Annual International Computer Software and Applications Conference (COMPSAC2002)*, Oxford, England, pp. 173-178, August 26-29 2002.

[5] M.Cai, J.Q. Song and M.R. Lyu, "A New Approach for Video Text Detection," *Proceedings International Conference on Image Processing (ICIP2002)*, vol. 1, Rochester, New York, pp. 117-120, Sept. 22-25 2002.

[6] K.F. Jang, H.M. Tang, M.R. Lyu, and I. King, "A Face Processing System Based on Committee Machine: The Approach and Experimental Results," *Proceedings 10th International Conference on Computer Analysis of Images and Patterns (CAIP2003)*, Groningen, The Netherlands, August 25-27 2003.

[7] M. Christel, A. Olligschlaeger, and C. Hung, "Interactive Maps for a Digital Video Library," *IEEE Multimedia* 7(1), pp. 60-67, 2000.

[8] M.R. Lyu, E. Yau, and K.S. Sze, "A Multilingual, Multi-modal Digital Video Library System," *Proceedings ACM/IEEE Joint Conferences on Digital Libraries*, Portland, Oregon, pp. 145-153, July 14-18 2002.

[9] 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, pp.472, June 24-28 2001.

[10] Rune Hjelsvold, Subu Vdaygiri, and Yvew Leaute, "Web-based Personalization and Management of Interactive Video", *Proceedings the Tenth International World Wide Web Conference*, Hong Kong, May 1-5, 2001.

[11] J. Son, M.R. Lyu, J.-N. Hwang, and M. Cai, "PVCAIS: A Personal Videoconference Archive Indexing System," *Proceedings 2003 International Conference on Multimedia & Expo (ICME2003)*, Baltimore, Maryland, July 6-9 2003.

[12] W. Wang and M.R. Lyu, "Automatic Generation of Dubbing Video Slides for Mobile Wireless Environment," *Proceedings 2003 International Conference on Multimedia & Expo (ICME2003)*, Baltimore, Maryland, July 6-9 2003.

[13] World Wide Web Consortium, "XML Query Requirement", W3C Working Draft, February 15 2001.

[14] Ruby B. Lee, "Realtime MPEG video via software decompression on a PA-RISC processor", *Proceedings Compcon '95*, pp 186-192, 1995.

[15] Compaq iPAQ Pocket PC,

http://www.compaq.com/products/handhelds/pocketpc

[16] Gismosoft Pocket Knife Chinese package

http://www.gismosoft.com/eng/products

[17] Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B. and Williams, J.G., "Visualization of a Document Collection: the VIBE System", *Information Processing & Management*, Vol. 29, No. 1, pp 69-82. 1993.

[18] Schwartz, R., Imai, T., Kubala, F., Nguyen, L. and Makhoul, J., "A Maximum Likehood Model for Topic Classification in Broadcast News, Eurospeech 97", *5th European Conference on Speech Communication and Technologies*, Rhodes, Greece, September 1997.

[19] Andreas Girgensohm, John Boreczky and Lynn Wilcox "Keyframe-Based User Interfaces for Digital Video", *IEEE Computer* , volume 34, number 9, pp. 61-67, 2001.

[20] NCAM: CPB/WGBH National Center for Accessible Media http://www.boston.com/wgbh/pages/ncam/

[21] Orkut Buyukkokten, Hector Garcia-Monlina and Andreas Paepcke, "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices", *Proceedings the Tenth International World Wide Web Conference*, Hong Kong, May 1-5, 2001.

[22] World Wide Web Consortium, "Synchronized Multimedia Integration Language (SMIL 2.0) Recommendation", August, 2001. http://www.w3.org/TR/smil20/.