iVIEW: An Intelligent Video over InternEt and Wireless Access System

Michael R. Lyu
Sam Sze
Edward Yau
Computer Science and Engineering Department
The Chinese University of Hong Kong
Shatin, Hong Kong
{lyu,samsze,edyau}@cse.cuhk.edu.hk

Abstract

We describe the design and implementation of a digital video content management system, *iVIEW*, for intelligent searching and access of video contents over Internet and wireless devices. The *iVIEW* system allows full content indexing, searching and retrieval of multilingual text, audio and video material. *iVIEW* integrates image processing techniques for scenes and scene changes analyses, speech processing techniques for audio signal transcriptions, and multilingual natural language processing techniques for word relevance determination. *iVIEW* is composed of three subsystems: Video Information Processing (VIP) Subsystem for multimodal information processing of rich video media, Searching and Indexing Subsystem for constructing XML-based multimedia representation in enhancing multimodal indexing and searching capabilities, and Visualization and Presentation Subsystem for flexible and seamless delivery of multimedia contents in various browsing tools and devices. We present overall and detailed infrastructure of *iVIEW*, describe its system characteristics, and evaluate the customer view on its performance and functionality. Integrating image, speech, and natural language processing techniques into Web-based environments for friendly user interface and seamless browsing, *iVIEW* provides a unified, end-to-end management of video-based media contents, from their creation to delivery, over WWW and mobile Web designed for today and tomorrow.

Keywords: Multi-Modal Interactions, Middleware and Browser Interactions, Browser on Mobile Devices, Multimedia Management and Support, Applications.

Approximate Word Counts: 6800

1. Introduction

Videos represent rich media over World Wide Web. Video contents can be explored and engaged in historical documents, museum artifacts, tourism information, scientific and entertainment films, news clips, courseware presentations, edutainment material, and virtual reality applications. They enrich the Web not only by the enhancement of its knowledge base but by an effective delivery and exchange of knowledge with dynamic user interfaces. Video information processing requires rigorous schemes for appropriate delivery of its rich contents over Web and mobile Web. The information that can be extracted from a video includes video streams, scene changes, camera motions, text detection, face detection, object recognition, word relevance statistics, transcript generation, and audio level tracking. The techniques involved in composing videos into vast digital video libraries for content-based retrieval are provided in the literature [Christel96].

Video information processing includes, among others, speech recognition, optical character recognition (OCR), text detection, and face recognition. The basic theory for speech recognition is hidden Markov models (HMMs) [Rabiner89]. In developing practical systems, speech recognition can be specially tailored to broadcast news transcription [Woodland98], spoken document retrieval [Meng01], and speaker identification [Lovekin01].

Optical character recognition (OCR) is a subject for many years' research [Mori92, Nagy92]. More recently, [Jaehwa00] describes hierarchical OCR, a character recognition methodology that achieves high speed and accuracy by using a multi-resolution and hierarchical feature space. [Xu99] proposes a prototype extraction method for document-specific OCR systems. The method automatically generates training samples from un-segmented text images and the corresponding transcripts. [Ganis98] presents a neural network classification scheme based on an enhanced multi-layer perception (MLP) and describe an end-to-end system for form-based handprint OCR applications designed by the National Institute of Standards and Technology (NIST) Visual Image Processing Group.

Before OCR can be applied to videos for text generation, text has to be detected in the videos first. Some research work has been performed in text detection on videos. [Li00] implements a scale-space feature extractor that feeds an artificial neural processor to detect text blocks. The text-tracking scheme consists of two modules: a sum of squared difference (SSD) based module to find the initial position and a contour-based module to refine the position. In [Wu99] text is first detected using multi-scale texture segmentation and spatial cohesion constraints, then cleaned up and extracted using a histogram-based binarization algorithm. [Gargi99] describes a system for detecting, tracking, and extracting artificial and scene text in MPEG-1 video. [Garcia00] proposes a text detection and segmentation

algorithm that is designed for application to color images with complicated background. [Agnihotri99] describes a method for detection and representation of text in video segments. The method consists of seven steps: channel separation, image enhancement, edge detection, edge filtering, character detection, text box detection, and text line detection.

Object detection and recognition are essential to content-based searching on videos, among which face recognition is of the major interest. As one of the earliest work, [Turk91] projects face images onto a feature space ("face space") that best encodes the variation among known face images. The face space is defined by the "eigenfaces", which are the eigenvectors of the set of faces; they do not necessarily correspond to isolated features such as eyes, ears, and noses. The framework provides the ability to learn to recognize new faces in an unsupervised manner. [Lee96] computes the frame bounds for the particular case of 2D Gabor wavelets and derives the conditions under which a set of continuous 2D Gabor wavelets will provide a complete representation of any image, particularly face images. [Wiskott97] presents a system for recognizing human faces from single images out of a large database containing one image per person. Faces are represented by labeled graphs, based on a Gabor wavelet transforms. Image graphs of new faces are extracted by an elastic graph matching process and can be compared by a simple similarity function. [Belhumeur97] develops a face recognition algorithm that is insensitive to large variation in lighting direction and facial expression. The projection method is based on Fisher's Linear Discriminant and produces well separated classes in a low-dimensional subspace, even under severe variation in lighting and facial expressions. Finally, a comparative study of three recently proposed algorithms for face recognition, eigenface, auto-association and classification neural nets, and elastic matching, can be found in [Zhang97].

Integration of these techniques is also an important direction on video-based research. [Wallick01] presents algorithms that improve the interactivity of on-line lecture presentation by the use of optical character recognition and speech recognition technologies. [Viswanathan00] describes a scheme to combine the results of audio and face identification for multimedia indexing and retrieval. Audio analysis consists of speech and speaker recognition derived from a broadcast news video clip. [Houghton99] develops Named Faces, a fully functional automated system that builds a large database of name-face association pairs from broadcast news. Faces found in the video where superimposed names were recognized are tracked, extracted, and associated with the superimposed text. With Named Faces, users can submit queries to find names for faces in video images.

The major integration effort for digital video research and system development, however, is the Informedia project [Wactlar96]. The Informedia Digital Video Library project provides a technological foundation for full content indexing and retrieval of video and audio media. It describes three technologies involved in creating a digital video library. Image processing analyzes scenes, speech

processing transcribes the audio signal, and natural language processing determines word relevance. The integration of these technologies enables us to include vast amounts of video data in the library. Surrogates, summaries and visualizations have been developed and evaluated for accessing a digital video library containing thousands of documents and terabytes of data. Although Informedia represents a key milestone in digital video library archival and retrieval integration technique, it was not originally designed for Web-readiness, interoperability, and wireless accessibility.

In this paper we describe the design and development of *iVIEW*, an Intelligent Video over InternEt and Wireless access system. Similar to the Informedia project, we integrate a variety of video analyzing and managing techniques, including speech recognition, face recognition, object identification, video caption extraction, and geographic information representation, to provide content-based indexing and retrieval of videos to users using various browsing devices. The major advantages of *iVIEW* over Informedia are three fold: (1) *iVIEW* is designed on Web-based architecture with flexible client-server interactions, scalable multimodal media information processing and retrieval, and multilingual capabilities. (2) *iVIEW* facilitates an XML-based media data exchange format for interoperability among heterogeneous video data representations. (3) *iVIEW* constitutes dynamic and flexible user interfaces sensitive to client devices capacity and access schemes over wired and wireless networks. The details are described in the following sections.

2. Research Motivation, Objectives and System Requirements

In the previous section a number of techniques for the video information extractions are described. Not only is the accuracy of these techniques a key to the success of a digital library, but the increasing number of different techniques affect the design of an "open" digital video library system to a great extend. The scalability issue of the digital library in terms of adding new extraction components is also very challenging. When a new extraction method is developed and included in the video library, it implies a series of newly added indexing and presentation functions. Particularly, a generic framework for presentation and visualization of video information is curial to the deployment of the digital library over the Web. Before getting into the details of the client browser for *iVIEW*, we first provide a short overview on the design methodology of the system.

The *iVIEW* system is based on an open architecture methodology and Web application models to achieve 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 form the videos, including information interchange between distributed video libraries and different publishing media.
- To allow single processing and extraction of video information and multiple delivery and presentation of the video contents to different computing platforms and devices.

The above objectives are achieved via the following methods: (1) modality concepts of the digital video library functions and video information life-cycles; (2) the collaboration of the video information processing modules; and (3) a generic framework for presentation and visualization of video information.

Modal Concept

We define the modal domain as a domain or type of information that can be extracted from the video. Examples are the text generated by speech recognition and the human identity by face recognition. The set of functions that a modal domain supports in the digital library applications is called the modal dimension. Typical processes are video information extraction, indexing and presentation. For different modal dimensions, the requirements for the whole series of library functions may be totally different. Compare, for example, the query input for a text search and a face search, where the former is a string while the latter is picture. The text indexing method is quite different from the face indexing method. Furthermore, at the front-end interface of the presentation client, a text visualization tool is also very different from the human face visualization tool in their layouts, content displays, and presentation styles. Although the implementation of individual modal dimension will be different, their "life-cycle" is similar to each other 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 modal dimensions can be easily identified. This provides an efficient way to adding different new modal dimensions into the system without losing the flexibilities.

Collaboration of the Video Information Processing

Video information processing is the content creation step for the digital video library. The collaboration of different video information extraction techniques is at the information exchange level, mainly including knowledge cross-referencing and knowledge enrichment.

For some video processing techniques, the accuracy of the recognition process can be increased by cross-referencing information generated by other modal dimensions. For example, to identify a human face in the video, the face recognition technique can be the primarily extracted modal information. On the other hand, the on-screen title of the person's name, when available, can be recognized and served as the cross-reference knowledge for identification of the person. An example of knowledge enrichment is the geographical naming process. The geographical naming database represents a knowledge repository of geographical names of countries, states, cities, and other identities. By applying this information to the text recognized from the speech recognition, the knowledge encapsulated in the text can be enriched.

However, difficulties still arise, as no single developer can provide all the video processing modules. There should be a standard way for different modules or developers to understand each other. At the same time, this standard should be flexible enough to maintain the openness required in the Web.

Generic Presentation and Visualization Framework

Individual modal dimension has its own interface requirements and the presentation method may be unique for each dimension. A generic presentation and visualization framework is required to collaborating different presentation modules. Different from the video information processing, the collaboration between the presentation and visualization modules are at the event levels, such as time synchronization and user events.

Another requirement of the framework is a transparent encapsulation of the delivery platforms. Desktop PCs, personal digital assistants and mobile devices pose different restriction for various needs in presentation and visualization. The proposed framework should be able to adapt to different configurations without additional installations or considerable overheads.

3. *iVIEW* Overall Architecture

The *iVIEW* system is a solution that attempts to achieve the objectives and system requirements posted in the previous section. Figure 1 shows the overall architecture of *iVIEW* system. The system is composed of three major subsystems: Video Information Processing (VIP) Subsystem, Searching and Indexing Subsystem, and Visualization and Presentation Subsystem.

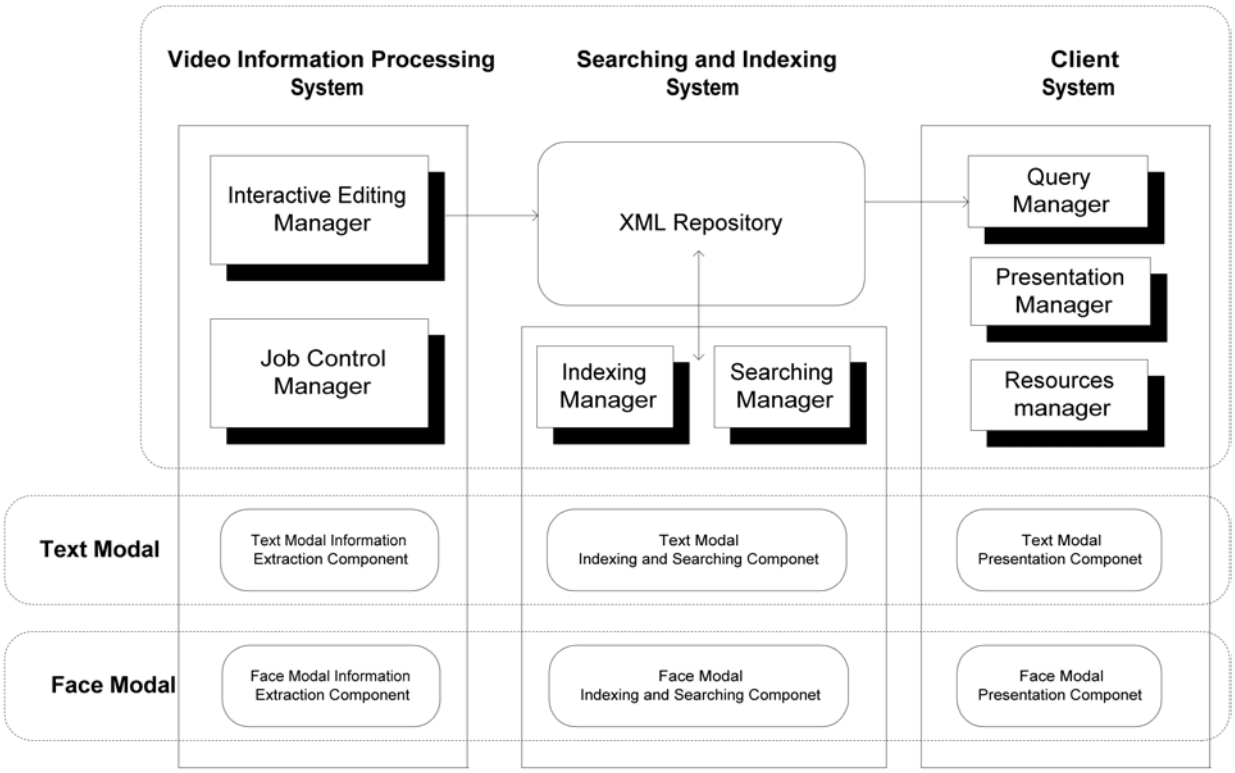

Figure 1: The *iVIEW* Logical Framework

The VIP Subsystem handles multimodal information extracted from a video file. The multimodal 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 file is available.

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

The Visualization and Presentation Subsystem handles query results, sets visualization mechanisms, and delivers multimodal presentations in time-synchronized manner. The Resources Manager automatically keeps track of the client device resources and bandwidth capacities for appropriate and efficient content delivery.

If we view the information flow in Figure 1 from left to right, multimodal information (text modality, image modality, face modality, 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, multimodal 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 or presentation of the new modal dimension. It facilitates the integration of newly obtained techniques on evolving modal information.

The details of these three major subsystems and their associate modality processing techniques are discussed in the following sessions.

4. Video Information Processing (VIP)

The Video Information Processing is the first processing step in the *iVIEW* system. The information contained in the video is extracted and recognized by a series of processes. The VIP is implemented on the Microsoft Windows 2000 platform. Figure 2 shows the overview of the VIP. Starting from the information channel contained in the digital video, going through the analysis and recognition processes, the primary information is generated. Followed by the knowledge enrichment and knowledge cross-referencing processes, the secondary information or the modal information is produced using the primary information as input.

Using the face modal dimension as an example, the VIP consists of two processes, the face recognition from the video information channel and the knowledge cross-referencing with text modal information to generate named face information. The text modal information is obtained by the Video OCR processes where the on-screen words are being recognized. Together with the language processor to identify the name (a proper noun) inside the text modal information, the name information is then cross-referenced with face modal information to generate the named face.

The number of analysis or recognition processes is different in each modal dimension. For example, the text modal dimension has two major recognition processes, the video OCR and the speech recognition. But the face modality has one recognition process only. The primary text modal information is stored in XML format similar to the following structure:

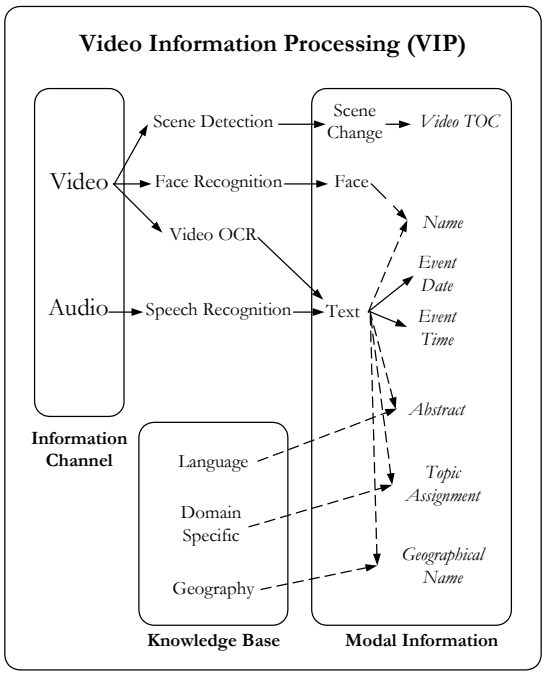

Figure 2: Overview of the video information processing.

```
<modal name="text">
      <text start="0s" end="5s">Hong Kong is a beautiful place.</text>
      <text start="5s" end="7.5s">You can find</text>
      <text start="7.5s" end="9s">the best clothing here.</text>
</modal>
```

Each modal dimension manages its own XML DTD or XML schema. The DTD would be extended, if knowledge enrichment or cross-referencing is applied. In the above example, the geographical name process will extend the XML as follows:

```
<modal name="text">
      <text start="0s" end="5s">
            <geoname>Hong Kong</geoname>
            is a beautiful place.</text>
      <text start="5s" end="7.5s">You can find</text>
      <text start="7.5s" end="9s">the best clothing here.</text>
</modal>
```

By extending the XML DTD, the new tag that represents new modal information is added to the original XML file. The new modal dimension, in this example the geoname, can be treated as an extended-modal dimension of the parent text modal dimension. As defined in the previous section, a

modal dimension consists of the presentation and visualization process. The geoname modal dimension uses a visual map as the presentation media. The corresponding geographical names in the video can also be visualized via the same map. The rendering process of geoname in the XML part, therefore, is governed by the geoname XML DTD. In other words the rendering process of geoname only needs to "understand" the geoname XML DTD, not the whole XML DTD.

For two independent modal dimensions, text and scene change, the XML part will be combined as follows to form the final XML document. Figure 3 shows the snapshot of VIP after scene change processing.

```
<modality name="text">
    <text start="0s" end="5s">
        <geoname>Hong Kong</geoname>
        is a beautiful place.</text>
    <text start="5s" end="7.5s">You can find</text>
    <text start="7.5s" end="9s">the best clothing here.</text>
</modality>
<modality name="scene change">
    <scene start="0s" end="20s" src="/101.jpg"/>
    <scene start="20s" end="33s" src="/102.jpg"/>
</modality>
```

Figure 3: Snapshot of the VIP for Scene Changes

5. XML Search Engine

Resembling the Web search engine, we have developed an XML search engine for the digital video library. The search engine is divided into two parts, the XML repository and the searching and indexing subsystem. The XML documents that represent the information extracted from the VIP are in general very static. They are not frequently modified. Thus an XML repository server is a natural way to manage XML documents. Using the HTTP protocol, clients can obtain the XML documents by sending the HTTP GET request with specified URLs. Updates of the XML documents are through the HTTP PUT mechanism. Enhanced access control is implemented to avoid un-legitimate updates.

Differ from a database solution of the XML search engine, the indexing part of the XML documents are handled by individual modality process as illustrated in Figure 4. Each modal dimension would implement the Searching and Indexing Subsystem according to the specification of the module that governs the interface and function requirements. The basic functions include module management, query management of specific modality domains, indexing of the modal information and resource management. For example, the face modal dimension will use a high dimension tree data structure for indexing, but the text modal dimension may only need an inverted tree implementation. Beside the performance consideration of the search engine, this loadable module design also makes the subsystem more easily scaleable in terms of functionality.

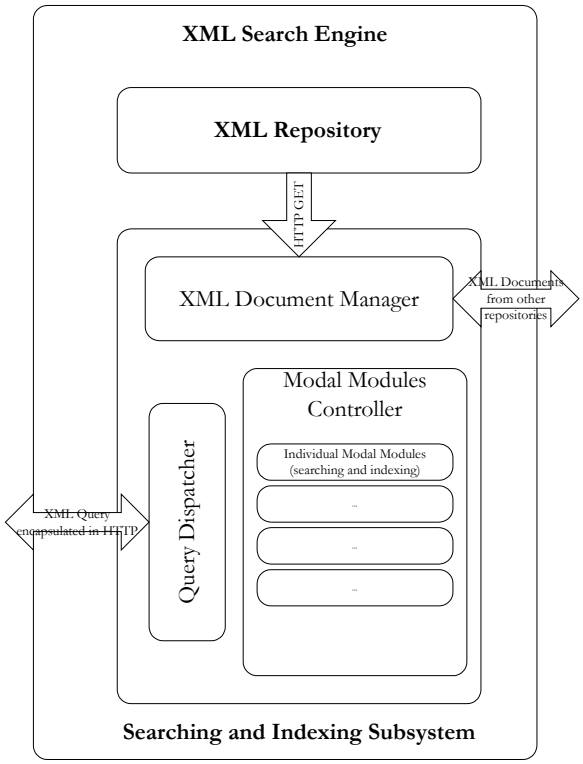

Figure 4: Overview of the Searching and Indexing Subsystem

To support multimodal search, the multimodal query from the client is passed to the query dispatcher. The dispatcher decodes the request and passes the sub-query to individual modality modules. After the individual results are returned, the dispatcher combines the search results into a summary reply and sends it back to the client.

6. Visualization and Presentation Subsystem

6.1 Background

The *iVIEW* Visualization and Presentation Subsystem, or simply the client, handles multimodal query, visualization of the result set and presentation of multimedia contents dynamically over the Web. It is also customized for visualization and presentation over wireless devices, encapsulating the capability to synchronize among various modal dimensions at the content presentation stage.

Client design can directly affect user experience; therefore, it plays a significant role in digital video library deployment. Previous research and implementation work in this area includes [Christel99, Christel00, Christel01]. Our goal is to develop a client with multilingual, multi-windows, and multi-device (including wireless) support. We have implemented the client subsystem based on the following approaches:

- A Java applet using our proposed architecture supporting multilingual, multimodal and Web-ready information processing and presentation.
- A Web-based client for Internet and wireless access.
- A Windows CE native application for wireless access.

Due to the cross-platform nature of Java, our Java applet client is browser interoperable. XML is employed for the client-server communication. Schemas of our XML messages focus on context-aware presentation. We recognize that the context-aware presentation has several advantages over the existing media presentation standards like SMIL and SAMI. The content awareness capability of the client leads to the scalable handling of media presentation, which is a key for designing generic browser architecture to suit different platform capabilities from desktops to mobile devices.

6.2 Java Applet Implementation for Internet Deployment

Our chief reference implementation is programmed as a Java applet (See Figure 5). The Java applet's nature makes the system accessible through any Web browser. We have verified the system's

compatibility with Microsoft Internet Explorer 5 and Netscape 4. We make use of JAXP 1.0 as the XML Parser. Java Media Framework (JMF) 2.1 is employed for playing streaming video. With the usage of plug-ins in JMF, we currently support two popular streaming video formats including Quicktime and Real. The Microsoft Media format is not supported in our Java client as there does not yet exists any JMF plug-in that supports Microsoft Media format.

Figure 5: A Screen Shot of the *iVIEW* Java Client

Basically, the *iVIEW* client is a component-based subsystem composed of a set of infrastructure components and presentation components. The infrastructure components provide services for client-server message communication and time synchronization among different presentation components by message passing. The presentation components 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 6 shows the architecture of the Visualization and Presentation subsystem. Infrastructure components are in shaded color.

The client-server communication message is coded in XML through HTTP. The XML is embedded into an HTTP POST message. Using HTTP enjoys the advantage that the service is seldom blocked by firewalls [Cheung01]. It also facilitates the deployment of an application to content providers or data centers. Although it does not conform to XML query standard, the message in XML is already self-

explanatory. Once a search result is attained, the media 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.

The message dispatcher dispatches media events to different presentation components according to the performance measuring inputs from resources monitor. 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 client devices and platforms.

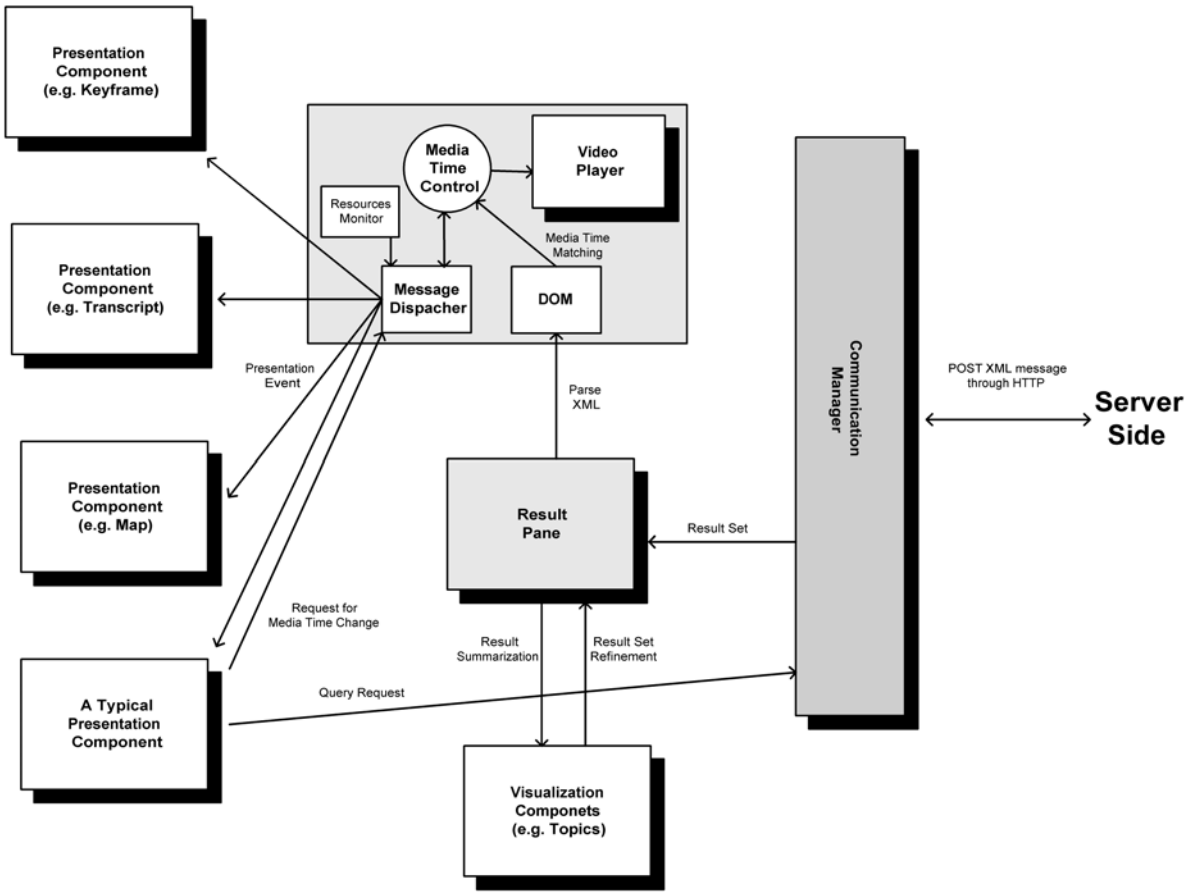

Figure 6: *iVIEW* Visualization and Presentation Subsystem Architecture

The *iVIEW* Java client has a multi-windows user interface. In the query window, a user can type in the keywords for text query, and a set of matched results represented by the poster images is shown in Figure 7. A tool-tip floating box shows the abstracts of a video clip when the mouse points to its poster image. A sample query and result is listed on Figures 8 and 9, respectively.

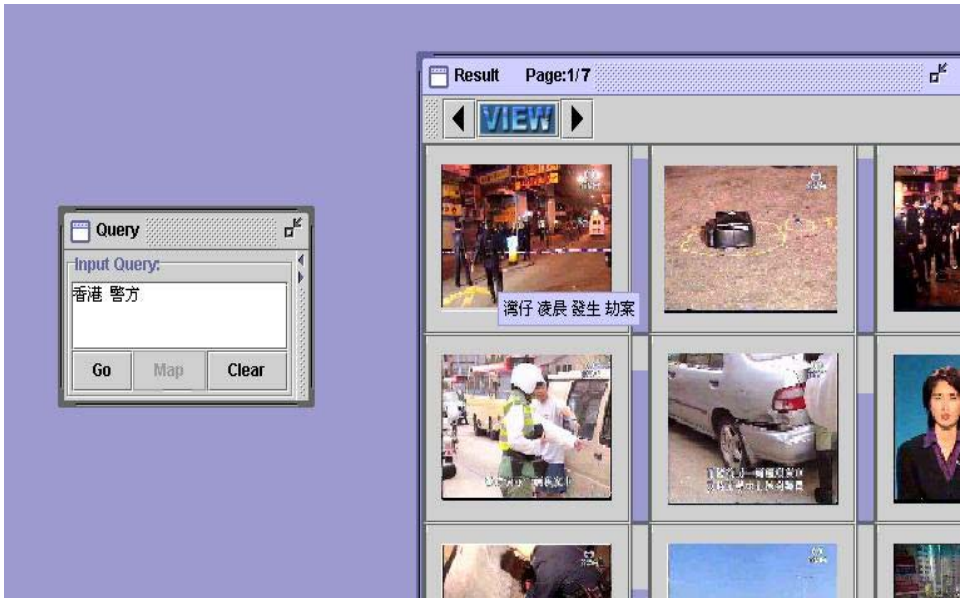

Figure 7: Java Applet Client Query and Result Set

```
<query>
  <relation logic="and">
    <item id="1" modality="text">
      <text>government</text>
    </item>
    <item id="2" modality="text">
      <text>policy</text>
    </item>
    <item id="3" modality="map">
      <mapstart longitude="100" latitude="60" />
      <mapend longitude="120" latitude="90" />
    </item>
  </relation>
</query>
```

Figure 8: A Sample Multi-modality Query in XML

```
<result path="/iview/data/">
  <item id="1" video="236" score="100">
    <abstract>.....</abstract>
  </item>
  <item id="2" video="206" score="95">
    <abstract>.....</abstract>
  </item>
  <item id="3" video="102" score="88">
    <abstract>.....</abstract>
  </item>
  .....
</result>
```

Figure 9: A Sample Result set in XML

The result set may be large in many cases; it is a difficult task for a user to select her desired result out of the many. Therefore, visible categorization of each element of the result set can aid the user to refine the large result set. The result set can be visualized in different views by summarizing them in different aspects. [Wactlar00] has contributed to various summarization techniques. Those techniques include classification by geographical locations, timeline, visualization by example (VIBE) and topics assignment.

Figure 10 illustrates a screen shot of *iVIEW* Java Applet Client result set categorized in topics. Each result element is assigned to one or multiple predefined topics. The topics are the text tag arranged in a circular shape. 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 she is interest in. The result set will then confine to the selected results.

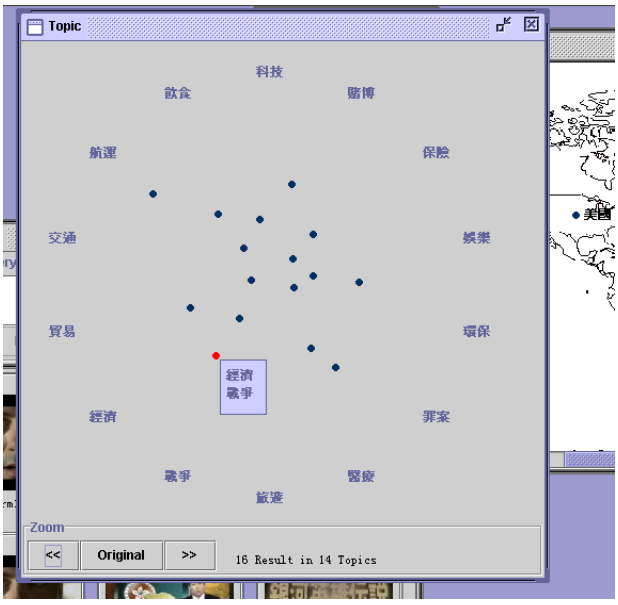

Figure 10: Result Set Visualization in Topics

After the user selected her target video clip, the user can right-click to play, view filmstrips or transcript of the video clip. The matched items are highlighted with different colors, as seen in Figure 11. The video clip and all other presentation components are presented in a synchronized manner.

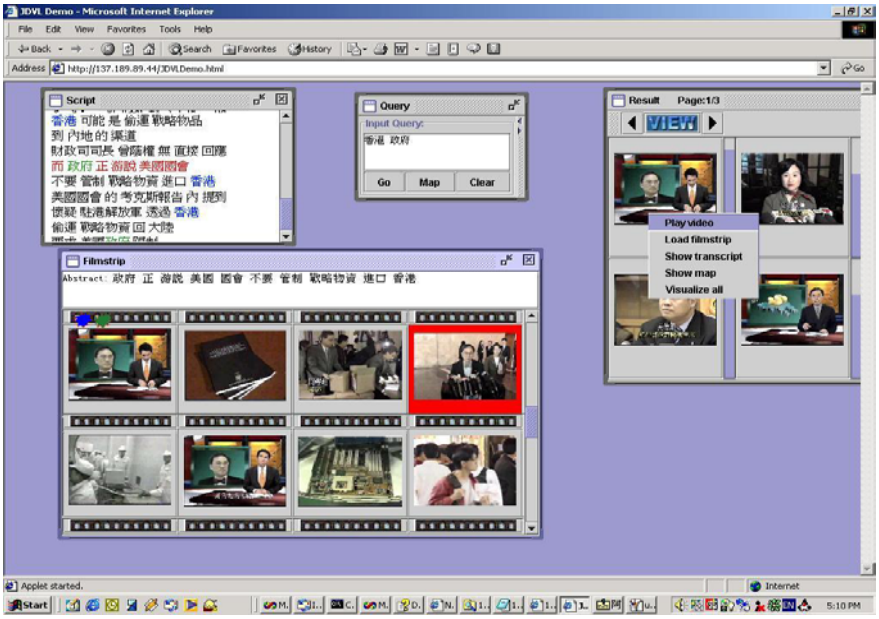

Figure 11: All modals are presented in synchronized manner and with matched items highlighted

In general all *iVIEW* presentation components support four general interface functions:

1. Initialization

   iVIEW client shows all matched items when the video media initializes. For example, a map shows all matched geographical locations when a video starts to play.

2. Passive Synchronization

   A particular piece of information will be highlighted when the media time is matched. For example, a geographical name is further highlighted when the location is mentioned in the video clip.

3. 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 particular filmstrip, the video, and also other presentation components, are re-synchronized to the time when that filmstrip occurs.

4. Search Input

   A presentation component can be inversely used as an input domain for searching. For example, a user can search video by highlighting certain geographical locations (as shown in Figure 12).

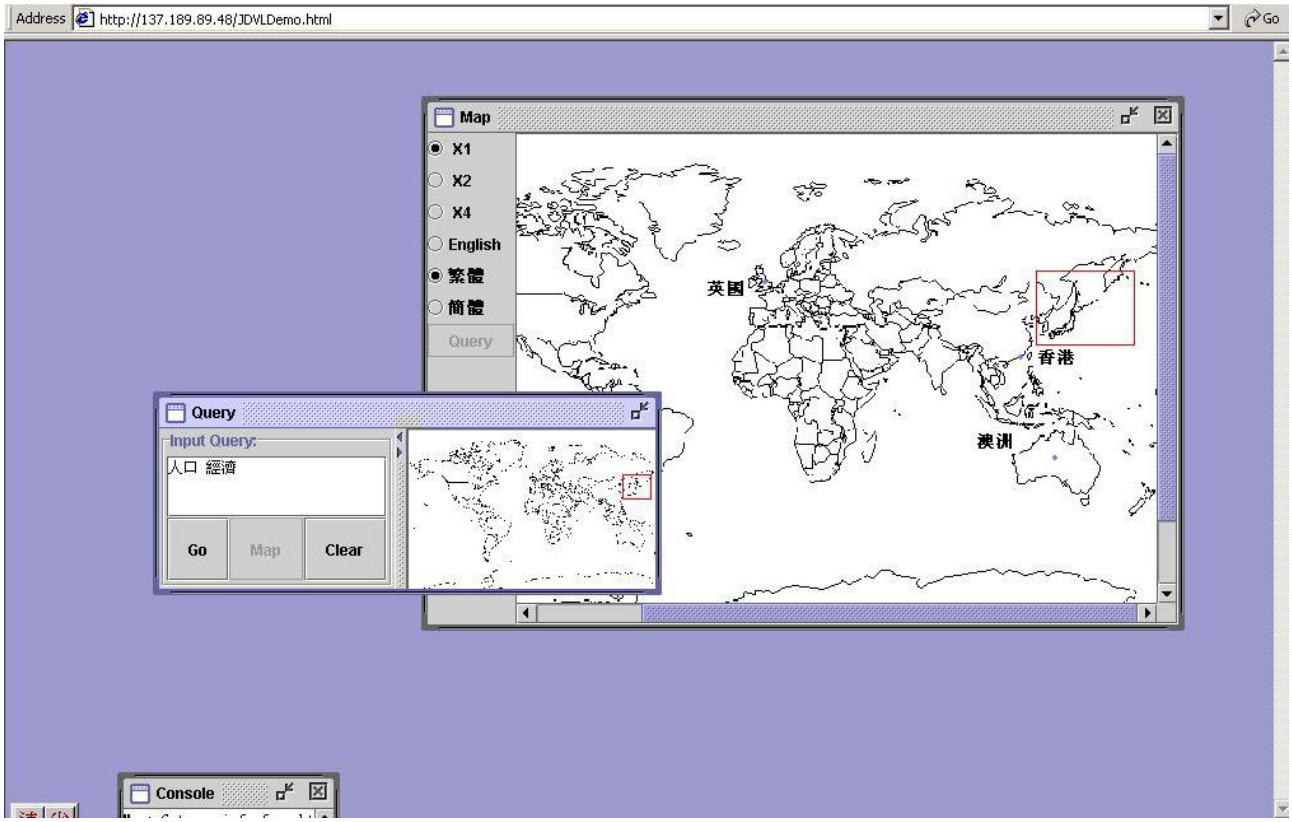

Figure 12: The Map Component Can Serve for Presentation as Well as Query Input

6.3 A Comparison with SMIL

SMIL is a W3C recommendation that enables simple authoring of interactive audiovisual presentations. The latest version is SMIL 2.0 [SMIL 2.0]. Players that supports SMIL 2.0 includes the RealOne Platform by RealNetworks [Real web] and GriNS [Oratrix web] by Oratrix. SAMI is a simplified format developed by Microsoft to support video captioning [Microsoft SAMI]. Our XML schema for multimodal presentation is similar to SMIL but with several advantages over the current SMIL standard and player implementations. Meanwhile, our XML schema enjoys simplicity like SAMI due to the specific presentation nature.

At the presentation stage, we classify each activity in a particular modal dimension as a media object with a specified duration time interval. An activity can show a picture, a text string, a video clip, a song, etc. Consequently, the corresponding objects are namely 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

s is the spatial displacement and the display area on the presentation device

A presentation P during T is a set of media objects. $P_T = \{ O_{m,t,s}, ... \}$

SMIL is a generic system that supports heterogeneous media object presentation in different timing combination. SAMI only supports two types of media objects: video and caption/transcript (O_{video} and $O_{caption}$). As the *iVIEW* client is a video-centered presentation system, the video is the basic media object and its duration is the whole presentation duration. (i.e. T of P_T = t of $O_{video,t,s}$)

Figure 13 is a sample XML content that the *iVIEW* client employs. It sequentially indicates the media objects.

```
<?xml version="1.0" encoding="big5" ?>
 <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" />
      <map>香港</map>
    </time>
   <time start="6">
      <script> MARGARET LOWRIE HAS THE DETAILS FROM
               LONDON
     </script>
    </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 13: *iVIEW* XML Media Presentation Description

We apply context tags to represent the media, such that the player can obtain the knowledge about the media context. This approach can facilitate the following features:

1. Intelligent object spatial displacement by the player

   Using SMIL, the author has to specify (or the language has to implicitly describe) the spatial displacement of a media object. In our case, it is not necessary for the author of the presentation to determine the spatial relationship (i.e. *s* is not necessary). The users can set preference on their client side, and the player knows the context of each media object for best delivery according to user preference.

2. 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. For example, if the user selects music and pictures together, when the bandwidth or CPU of a mobile device is not enough to cope with them both, the player, through default or user preference, can degrade the presentation gracefully by eliminating the low priority media. In this case, it can give way of transmission of pictures to music.

3. Generic application support

   In SMIL, if we want to present a map of, say, Hong Kong, we can make it through a picture. In fact, the client side may have installed a map application. Without knowledge of the context, bandwidth is wasted to transfer the redundant picture file. If the context is known, we only need to send text "Hong Kong" with context attribute "map" to the player and the player can launch an corresponding application to deal with the context. This can establish a standard for open context support, and in many cases it saves the precious bandwidth in, say, wireless connection.

   We formulate the client modality support as a function called Associate Media Application, *A(m)*

   $$A(m) = \{\text{Applications that support a particular modal dimension } m\}$$

   For example, *A(text) = {"MS-Word", "WordPerfect", "vi" ... }*

   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 easily be expanded to support additional modal dimensions.

As a benefit from the component-based approach, *iVIEW* client player can be synchronized not only through dragging the timeline, but also fast forward or backward driven by the context. As described in

the previous section, users can locate a media object in a particular modal dimension and force all other modal dimensions to synchronize with it.

In a usual practice of using SMIL, presentation authors specify a spatial area for captions or images and present the media element one by one in a slow, slide-showing manner. On the other hand, since *iVIEW* client presentation components may be advanced user interface controls like text box with scrollbar, the whole text content or the whole series of images can be included in the control before the presentation time. This capability facilitates flexible user browsing of selected content before or after the presentation time *t*.

We many consider this capacity as a super modal dimension *m'* that includes all modal dimensions *m* during the presentation. Mathematically we can denote it as:

$$m' = \{m_t\} \quad where\ t \in T$$

Consequently, we can combine this concept into the SMIL standard as an additional XML module for the *iVIEW* client.

6.4 PDA Implementation for Wireless Deployment

The Compaq PDA iPAQ H3630 equipping with Intel StrongARM 206MHz is considered to posses enough processing power for handling multimedia information. This device running Windows Pocket PC is selected as the target device for our development. The *iVIEW* client has been implemented successfully in two approaches for PDA deployment over wireless. Both versions can be applied through 802.11b Access Point, Bluetooth Access Point, and GSM HSCSD (High Speed Circuit Switching Data) public network. HSCSD is a technology that a data communication session can occupy three GSM data channels concurrently, which achieves the bit rate of 43.2kbps (i.e., three times of 14.4kbps which is the GSM data channel bandwidth).

The first approach is to implement a web-based application using techniques of Hyper-text Modeling Language (HTML), Cascaded Style Sheet (CSS), Dynamic HTML (DHTML) and JavaScript. This implementation is derived from the Java client architecture described in an early section. A translation layer is applied to convert the XML messages to HTML. It is viewed in Pocket Internet Explorer 3.0. This version of Internet Explorer only supports a limited subset of JavaScript 2.0 and lacks the support for DHTML. This makes the implementation of multi-windows presentation impossible. The operation series and screen demo of this approach is shown in Figure 14, from (a) to (f).

Figure 14a: **PDA with GSM HSCSD Phone Card Setup**

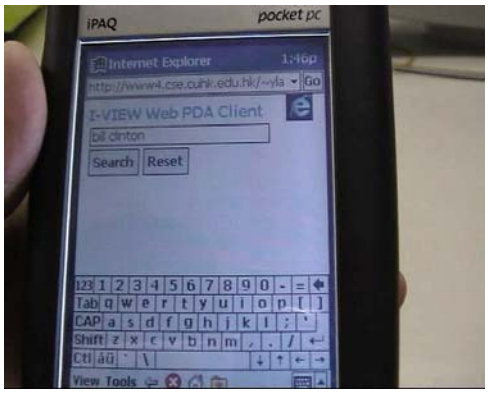

Figure 14b: **Query Page**

Figure 14c: **Result Page**

Figure 14d: **Abstract and Key Frames**

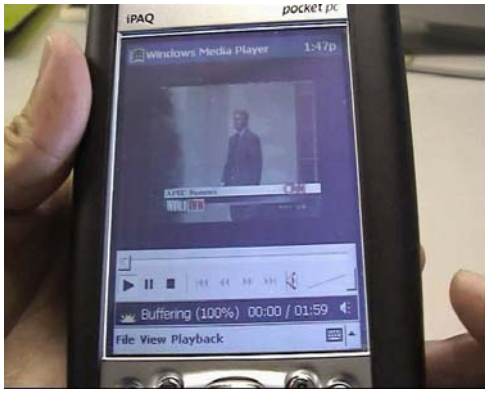

Figure 14e: **Streaming Video**

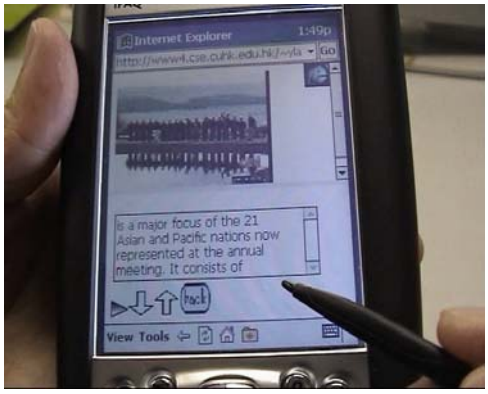

Figure 14f: **E-book Mode**
(A low bandwidth consumption configuration, which is a key frame slide-show in synchronization with scrolling transcript)

The second approach is to implement a native windows application for Pocket PC using Microsoft Embedded Visual Tools 3.0. Its design conforms to our Java client implementation reference. We applied the Direct X Platform Adaptation Kit for the video handling. A picture of our second approach implementation is shown in Figure 15. It utilizes the Windows API to build up a multi-windows environment like a desktop environment in the PDA device.

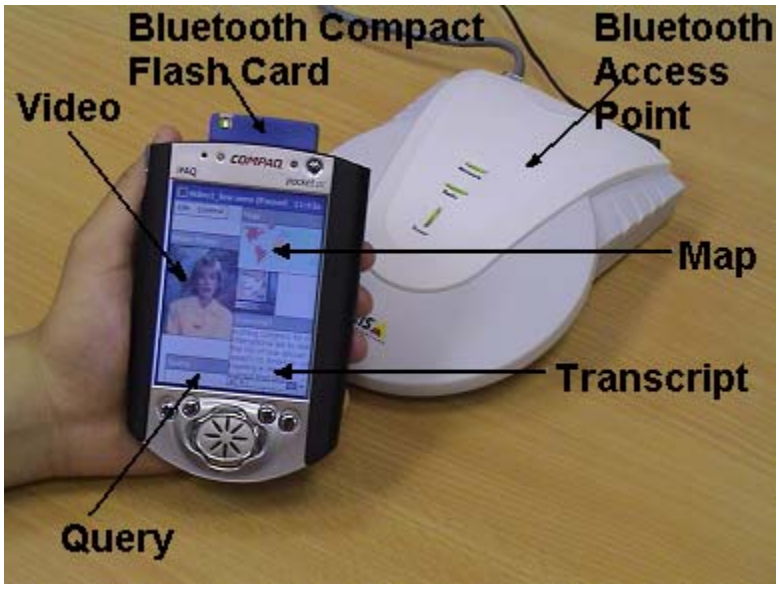

Figure 15: Wireless PDA Client as a Native Windows CE Application (configured for Bluetooth Access)

The video clips are encoded in Microsoft Media format in 64kbps. For local area wireless access via 802.11b or Bluetooth access point, the video data streaming is very smooth. For public access using a Nokia Phone Card 2.0 that supports GSM HSCSD, it provides more than acceptable delivery of quality video and audio. We have made more than 20 public presentations of this solution to over 100 industry people. The feedback has been overwhelmingly satisfactory.

7. Conclusions

In this paper we present the design and implementation of *iVIEW* as an intelligent digital video content management system for video searching and access over Internet and wireless devices. The *iVIEW* system allows full content indexing, searching and retrieval of multilingual text, audio and video material. The major components of *iVIEW* includes Video Information Processing, Searching and Indexing, and Visualization and Presentation. We present the detailed infrastructure of *iVIEW*, describe

its system characteristics, and perform customer evaluation on its user interface and performance. We emphasize the *iVIEW* design on its device-adaptable client implementation and flexible, dynamic user interfaces. We also distinguish our approach from the complementary SMIL standard, and highlight the advantages of *iVIEW*. Integrating multimodal video information extraction techniques into Web-based environments, *iVIEW* provides an XML-based, end-to-end processing of video-based media contents that can be seamlessly delivered over WWW and mobile Web. *i-VIEW* is a working prototype which has been widely demonstrated with satisfactory results and rave reviews.

Acknowledgement

The work described in this paper is 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).

References:

[Agnihotri99] L. Agnihotri, and N. Dimitrova, "Text detection for video analysis," *Proceedings 1999 IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '99),* Page(s): 109 – 113.

[Belhumeur97] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transactions on pattern analysis and machine intelligence,* VOL. 19, NO. 7, JULY 1997.

[Cheung01] 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.

[Christel96] M. Christel, H.D. Wactlar, S. Stevens, R. Reddy, M. Mauldin, and T. Kanade, "Techniques for the Creation and Exploration of Digital Video Libraries" *Multimedia Tools and Applications* (Volume 2), Borko Furht, editor. Boston, MA: Kluwer Academic Publishers, 1996.

[Christel99] 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.

[Christel00] M. Christel, A. Olligschlaeger, and C. Hung, "Interactive Maps for a Digital Video Library," *IEEE Multimedia* 7(1), pp. 60-67, 2000.

[Christel01] M. Christel, B. Maher, and A. Begun, "XSLT for Tailored Access to a Digital Video Library," *Joint Conference on Digital Libraries (JCDL '01)*, Roanoke, VA, pp.290-299, June 24-28, 2001.

[Ganis98] M.D. Ganis, C.L. Wilson, and J.L. Blue, "Neural network-based systems for handprint OCR applications," *IEEE Transactions on Image Processing*, Volume: 7 Issue: 8 , Aug. 1998, Page(s): 1097 –1112.

[Garcia00] C. Garcia, and X. Apostolidis, "Text detection and segmentation in complex color images," *Proceedings 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00*, Volume: 4 , 2000, Page(s): 2326 –2329.

[Gargi99] U. Gargi, D. Crandall, S. Antani, T. Gandhi, R. Keener, and R. Kasturi, "A system for automatic text detection in video," *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR '99)* , 1999, Page(s): 29 –32.

[Houghton99] R. Houghton, "Named Faces: putting names to faces," IEEE Intelligent Systems, Volume: 14 Issue: 5 , Sept.-Oct. 1999, Page(s): 45 –50.

[Jaehwa00] Jaehwa Park, V. Govindaraju, and S.N. Srihari, "OCR in a hierarchical feature space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 22 Issue: 4, April 2000, Page(s): 400 –407.

[Lee96] Tai Sing Lee, "Image Representation Using 2D Gabor Wavelets" *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 18, No. 10, October 1996.

[Li00] 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.

[Lovekin01] D.M. Lovekin, R.E. Yantorno, K.R. Krishnamachari, "Developing usable speech criteria for speaker identification technology", *Proc. IEEE*, pp. 421-424, vol. 1, May 2001.

[Meng01] H.M. Meng, P.Y. Hui, "Spoken document retrieval for the languages of Hong Kong", *Proceedings of 2001 International Symposium on Multimedia Processing*, pp. 201-204, May 2001.

[Microsoft SAMI] Microsoft Corporation, Understanding SAMI 1.0, October 2001 http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnacc/html/atg_samiarticle.asp

[Mori92] S. Mori, C.Y. Suen, and K. Yamamoto, "Historical review of OCR research and development," *Proceedings of the IEEE* , Volume: 80, Issue: 7, July 1992, Page(s): 1029 –1058.

[Nagy92] G. Nagy, "At the frontiers of OCR," *Proceedings of the IEEE*, Volume: 80, Issue: 7 , July 1992, Page(s): 1093 –1100.

[Oratrix Web] Oratrix Development's GriNS Player http://www.oratrix.com/GRiNS/

[Rabiner89] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, pp. 257-286, February 1989.

[Real Web] RealNetworks http://www.realnetworks.com

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

[Turk91] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces," *Proceedings IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Maui, Hawaii, 1991, pp. 586–591.

[Viswanathan00] M. Viswanathan, H.S.M. Beigi, A. Tritschler, F. Maali,"Information access using speech, speaker and face recognition", *IEEE International Conference on Multimedia and Expo*, pp. 493 - 496, vol. 1, July-August 2000, New York.

[Wactlar96] 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.

[Wactlar00] H.D. Wactlar, "Informedia – Search and Summarization in the Video Medium," *Imagina 2000 Conference,* Monaco, January 31 - February 2, 2000.

[Wallick01] M.N. Wallick, N. da Vitoria Lobo, M. Shah, "A system for placing videotaped and digital lectures on-line", *Proceedings of 2001 International Symposium on Multimedia Processing*, pp. 461-464, May 2001.

[Wiskott97] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Transactions on pattern analysis and machine intelligence,* Vol. 19, No.7 July 1997.

[Woodland98] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, and S.J. Young, "Experiments in broadcast news transcription", *Proc. IEEE*, pp. 909-912, vol. 2, May 1998.

[Wu99] V. Wu, R. Manmatha, and E.M. Riseman, “Textfinder: an automatic system to detect and recognize text in images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 21 Issue: 11 , Nov. 1999, Page(s): 1224 –1229.

[Xu99] Yihong Xu and G. Nagy, “Prototype extraction and adaptive OCR,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 21 Issue: 12 , Dec. 1999, Page(s): 1280 –1296.

[Zhang97] Jun Zhang, Yong Yan, and Martin Lades, “Face Recognition: Eigenface, Elastic Matching, and Neural Nets,” Proceedings of the IEEE, Vol. 85, No. 9, September 1997.