Decoupling the Information Application from the Information Creation: Video as Learning Objects in Three-Tier Architecture

Xiangming Mu
University of Wisconsin—Milwaukee, Milwaukee, WI, USA

mux@uwm.edu

Abstract

In this paper a new three-tier video application infrastructure is proposed. Video learning object is introduced as the intermedia tier and it connects the video metadata and video application tier. In addition to the traditional text metadata such as title and description, a wide range of visual metadata including key frames, video storyboards, and fast-forward video are integrated into the video learning object. The visual metadata is organized in hierarchical structure. Video timestamps are used to "mark" different video channels (i.e., visual channel, audio channel, text channel) for the purpose of synchronization. A MPEG-7 compatible XML metadata schema was developed to encode the video learning objects to achieve the reusability and interoperability—different video applications can use the same video learning object without having to reinvent the wheel. A semi-automatic video metadata authorization system called Video Annotation and Summary Tool (VAST) was developed to facilitate the creation of video metadata. The automatic process generates a series of key frame "candidates" (primitive frames) but human beings have to manually select the key frames. The number and the image size of the primitive frames are reconfigureable. The fast-forward version of the original video can also be created from the primitive frames. Finally, we demonstrate how the new three-tier infrastructure is used to guide the development of two independent video application projects: the Open-video digital library project and distance learning using the Interactive Shared Educational Environment (ISEE) project. Both projects utilize the same set of video learning objects.

Keywords: infrastructure, video, metadata, user interface, learning object

Introduction

Video, if used properly, can be effective for learning because: (1) it commonly contains scenes and action sequences that cannot be as well presented by other media (e.g., text, audio, and images), (2) it allows intensive examination of a particular process or event through replay, slow motion, and pausing to discuss or clarify a point, (3) it can contain computer graphics or diagrams in action or 3-D sequences that accurately illustrate complex processes, concepts, or events, (4) high quality, expensive-to-produce videos are usually available to learners at low cost or free, and (5) video lectures recorded from live classes enable learners to hear the speaker's tone/pitch and view gestures

providing additional clues to know what was emphasized.

The advance of network and computer technology increases the availability of digital videos over the Internet. Public video libraries, public/private institutes, special collections in libraries, and even personal home-made video collections are the common places to find digital videos for education and learning. Virtual digital video libraries, such as the Internet Archive Moving Image section (http://www.archive.org/movies/movies.php), CMU Informedia Project (http://www.informedia.cs.cmu.edu/), and UNC Open-Video Project (http://www.open-video.org), offer a large number of dedicated digital videos for education. In Open-Video project, for example, 294 videos from NASA's Distance Learning Center aiming at K-16 students, teachers, and parents are provided in multiple formats (i.e., MPEG-4, MPEG-2, MPEG-1, RealMedia™, and QuickTime™) (http://www.open-video.org/collection_detail.php?cid=10); in Informedia Project, 321 high quality videos are available for public education (http://www.open-video.org/collection_detail.php?cid=2); and in the Internet Archive digital library, 555 video episodes related to "Computer Chronicles" program are available for public access (http://www.archive.org/movies/computerchronicles.php).

Many companies and educational institutes also provide videos for education. For instance, Microsoft Multi University Research Laboratory serves as a digital repository for universities' lectures and seminars (http://murl.microsoft.com/default.asp). The video collection is widely used by both employees from Microsoft and public users for training and education. Stanford Center for Professional Development (SCPD) program at Stanford University (http://scpd.stanford.edu) provides online distance education using digital videos recorded from live classes. There are more than 450 companies and organizations and more than 5,000 students registered to receive the distance delivery programs. Public libraries may host some specific video collections. For example, approximately 1 million feet of unprocessed 16 mm film for 58 years (1922-1980) donated by a local television (WTMJ-TV) are digitalized by the library of University of Wisconsin--Milwaukee (http://www.uwm.edu/Libraries/arch/findaids/mss203.htm) to support education and research.

Accordingly, video-based application systems have also been developed to take advantage of these "free" online video resources for quality learning. The Interactive Shared Educational Environment (ISEE) (Mu & Marchionini, 2002) provides a distance learning environment that supports video synchronization between remote users based on "smartlinks"—a link from the interactive text chat room to the video player. eClass (Brotherton 2001) from Georgia Institute of Technology (formerly called class 2000) arranges class videos, associated web pages, and slides along a timeline to allow learners cross-reference. Hyperlinks used in each slide to build the connections to the relevant video segments. The BMRC Lecture Browser (www.bmrc.berkeley.edu/frame/projects), divides video segments based on the flipping of slides—each new slide indicates a new video segment, which is represented as a small thumbnail along a visual bar beneath the video player. Learners can click a thumbnail to go to the related video segment. A collaborative video viewing system developed by Microsoft allows distributed learners to collaboratively watch video using shared VCR controls (Cadiz et al. 2000).

A two-tier structure (user interface tier and video data tier) is usually the infrastructure for most of these video applications. On the user end (user interface tier), a specific user interface is designed to support particular functions (i.e., video browsing, video edit, and video retrieval). On the data end (video metadata tier), video and video metadata are stored in database(s) or as file(s) and are directly connected to a particular video application (i.e., the BMRC lecture browser).

The limitations for this two-tier infrastructure are two fold:

- The lacking of standards restrains the interoperability of the video metadata. With heterogeneous video metadata standards, it is hard for video metadata to be used in different

applications, even though the creation of high quality video metadata is very expensive (The creation of video abstraction and video key frame selection are usually manual process). For instance, the video key frames without timestamps meet the requirements for a displaying-oriented application such as the Internet Archive Moving Image web site, but are inappropriate for a sophisticated application such as the eClass distance learning system, which requires the video timestamps to build links between video segments and slides.

- The direct connection between video metadata and the video application limits the reusability of the video metadata. A specific video metadata format defined for a specific video application makes it hard for different applications to access and use the data. As a result, an extra Input/Output (IO) module must be developed for each of these applications as "adapters" in order to convert the format of that particular video metadata into the format the applications can "understand" and use.

Towards these issues, a concept of hierarchical video learning object, which decouples the direct connection between video/video metadata and video applications, is introduced. As a consequence, the conventional two-tier structure will be replaced by a new three-tier information infrastructure---namely, *video metadata tier, video learning object tier, and video application tier* (see Figure 1).

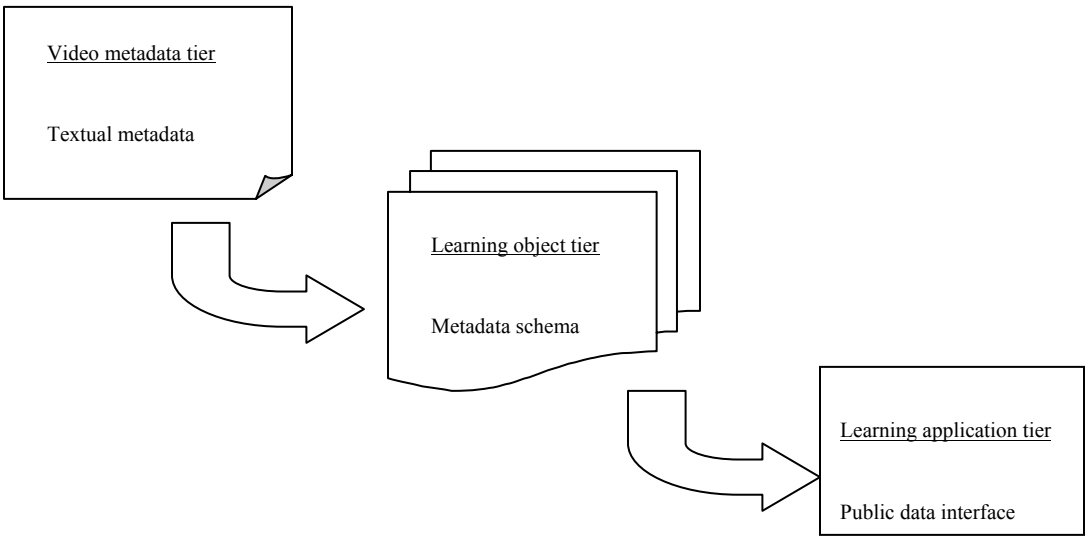


Figure 1: Three-tier video learing object infrastructure

Three-tier Architecture

Video learning object creates an independent tier called learning object tier between the video metadata tier and video learning application tier. Video and video metadata are re-constructed into video learning objects before they are utilized by heterogeneous video applications. In the conventional two-tier architecture, video and video metadata are treated as "data". No relationships and links are created between or within the data. Only after the data has been loaded into a particular application, the connections and relationships among different parts of the data begin to be established (i.e., the link between a frame and a video shot). In the new three-tier architecture,

however, by integrating video and video metadata into video learning object, both the data (video and video metadata) and the data structure (links and relationships of pieces of data) are included.

For each video learning object, a set of entities, attributes, and links are defined. The definition of these entities, attributes, actions, and links are based on a video metadata schema we developed which is compatible with MPEG-7 (http://www.chiariglione.org/mpeg/) and Dublin Core (http://dublincore.org/) standard. An entity describes one particular feature of the video learning object and one object may have multiple entities (i.e., a video key frame entity and a video description entity). Each entity has its own attributes (i.e., a key frame has the name attributes, sequence number attributes, etc.), links (i.e., a frame has a link to a video segment through the timestamp), and actions (i.e., a fast-forward can be played, pause, replay, et.). Borrowed from the concept of Object-Oriented Programming (OOP), the entities within an object are organized into a hierarchical structure due to the hierarchical nature of video---namely, attributes, links, and actions on an upper level entity are able to be inherited by the low level entities. Accordingly, the upper level entities are referred to as "parent" and the low level "child" entities. For instance, a shot is a collection of similar frames and is defined as the "parent entity" of a particular frame entity. In consequence, each frame inherits the attributes and links owned by its "parent entity", or shot entity. In addition, in the hierarchical structure each child entity can also "override" the attributes, actions, and links inherited from the parent entity. The "playback" action from the shot entity is overridden by the child frame entity to "display" action. Finally, XML is selected as the language to vehicle to the video learning objects. Indeed, a video learning object can be also viewed as a collection of video and video metadata wrapped by a set of XML tags defined by a particular schema.

On the video metadata tier, a semi-automatic Video Annotation and Summary Tool (VAST) was developed to help to create metadata in multiple granularity levels (i.e., frame level, shot level, and scene level) for learning objects. On the video application tier, a standard Input/Output (I/O) module was implemented to connect the learning objects with different video applications. In practice, over the past three years, two independent video application projects have adopted this new three-tier video learning object infrastructure--Interactive Shared Educational Environment (ISEE) project and the Open-video digital library project. Indeed, for both project, the same set of video learning objects were used.

Video Metadata Tier

Video metadata, also called the "data about the video data", refers to any data that helps for video information searching, personalization, and management. Video metadata can be classified into three major categories. *Content discovery and identification metadata* describes the video at the top level (i.e., title, publish date, abstract); *video content navigation and description metadata* get into sub-parts of the video content at various levels (segment level, scene level, shot level, and frame level); and *video signal processing and administrative metadata* relates to video information management such as copy right issue and video physical level description, such as the variation of color, shape, texture, and motion.

Video content identifier metadata provides general information for the entire video clip (title, author, date, type, subject, etc.), while video content description metadata not only provide the information about the video creation, usage, and media, but also describe the structural aspects of the video content. These aspects may include signal-based features (color, texture, shape, and motion) and semantic descriptions on various levels of video content, including segment, scene, shot, or even frame of the video. (A *frame* is a still image or picture of a video. A *shot* is a continuous sequence of frames captured from one camera. A *scene* is composed of one or more shots, which presents different views of the same event, related in time or space. A *segment* is composed of one or more related scenes.)

A number of standards are available for describing metadata. Dublin Core and IEEE LOM (Learning Object Metadata) are two import video content identification metadata schemas in the education domain, while MPEG-7 and SMPTE (The Society of Motion Picture and Television Engineers) are widely accepted as video content description metadata schemas. In our model we designed a new metadata schema based on Dublin Core and MPEG-7. Both content identification and description metadata are included in the new schema. A detailed discussion about the schema will be presented in a different paper. An example segment of the schema is presented in the later part of the paper.

Dublin Core is a simple content identification metadata model used for electronic resources. The original purpose of Dublin Core was for helping electronic information retrieval on the web. Now it has been extended to describe almost any metadata and now widely used in librarianship, computer science, text encoding, the museum community, and other related fields of scholarship across the world. There are fifteen elements defined in Dublin Core and most of them are semantically understandable and interoperable. With international consensus and strong flexibility, Dublin Core can be well integrated with other metadata schemas such as MPEG-7.

MPEG-7, an ISO standard developed by the MPEG (Moving Picture Experts Group) and officially named the "multimedia Content Description Interface," is a standard for describing features of audio and video content to facilitate video searching, browsing, and retrieving. MPEG-7 is a media independent international standard that gets supports from both the scientific field and the industrial field (MPEG 2001). MPEG-7 provides an open standard for describing the video content without being limited to a specific application. A set of Descriptors, Description Schemas, and relationships are defined as the basic components of MPEG-7. A Descriptor (D) is a representation of the feature and characteristic of the data. The description schema (DS) specifies the structure and semantics of the relationships between its components, which may be both Descriptors and Description Schemas. Both the DSs and Ds are encoded using XML. The syntax of the DSs and Ds are defined by the *Description Definition Language* (DDL), which is an extension of the XML Schema language.

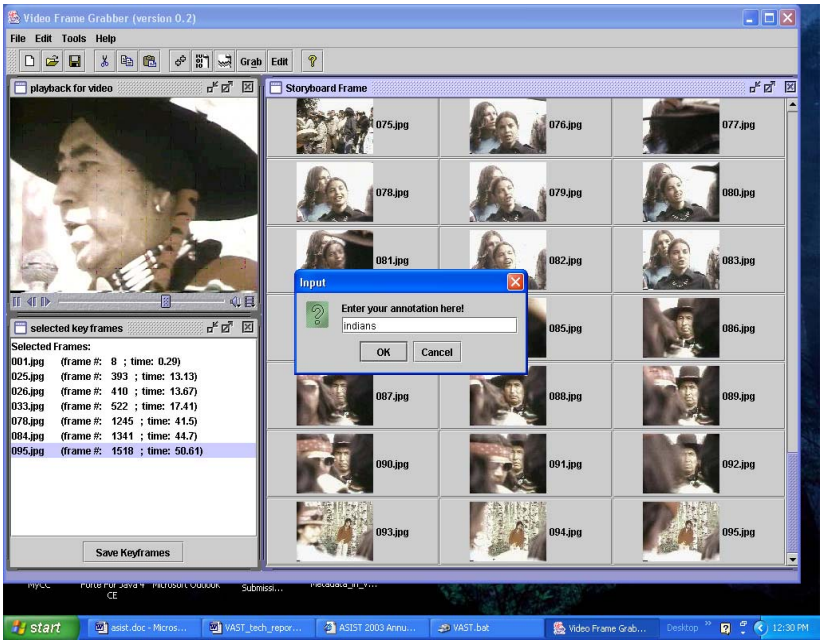


Figure 2: The user interface of VAST

To facilitate the creation of video metadata for the new three-tier infrastructure, a Video Annotation and Summary Tool (VAST) was developed. VAST is a semi-automatic video metadata creation tool--the automatic process generates a series of key frame "candidates" (primitive frames) for users to manually select the key frames. VAST supports the creation of both content identification and description metadata. Figure 2 is a snapshot of the user interface of the VAST. At the top of the desktop is the control panel. Beneath it on the right is the storyboard panel where frames generated from the video can be displayed. To the top left is a video player which enables users to monitor the playback of the video. A slide bar beneath it provides VCR video navigation support. Under the video player on the bottom left is the video annotation panel. A popup window appears for the user to add text annotation for a selected frame. Flexibility is provided for users of the interface by allowing them to freely rearrange the layout of the Graphic User Interface (GUI) in accordance to their personal preferences. The positions of various components can be dragged and dropped on an arbitrary position within the desktop. The size of each component is also changeable. For example, the user can enlarge the video player and place it in the center of the interface while watching the video, and minimize or close it while editing the metadata.

The video player is used to view and monitor the video playback. The default position for the video player is on the top left. Implemented in Java and the Application Program Interfaces (APIs) from Java Media Framework (JMF), this video player supports most of the current video formats including AVI, MPEG, and QuickTime. A visual VCR control component is located just beneath the video player to support manipulation functions such as pausing/resuming. A slider is also provided for users to jump to an arbitrary position along the video timeline. A tooltip, which indicates the current timestamp of the video, appears automatically when the user's mouse is over the slider and stay for several seconds.

The fast-forward version of the original video can be created by connecting primary frames with the normal display speed (30 f/s) using VAST. Many commercial tools (i.e., QuickTime™) can be utilized to do this task. The time compression ratio of the fast-forward is dependent on the parameter of the *Image grabbing ratio*, provided by the VAST configuration function. For example, choosing the default value of sixteen, the fast-forward will play sixteen times faster than the original one. Finally, all the frames as well as the associated metadata including the timestamp, frame number, name, and annotations are integrated into video metadata schema we developed in XML format.

Video Learning Object Tier

Video learning object tier is the intermedia tier that connects video, video metadata and video-based learning systems. In this tier the video and video metadata are re-constructed into video learning object.

Video Learning Objects

The Learning Technology Standards Committee (LTSC) describes learning objects in a fairly broad way:

> *Learning Objects are defined here as any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning. Examples of technology supported learning include computer-based training systems, interactive learning environments, intelligent computer-aided instruction systems, distance learning systems, and collaborative learning environments. Examples of Learning Objects include multimedia content, instructional content, learning objectives, instructional software and software tools, and per-*

sons, organizations, or events referenced during technology supported learning (LOM, 2004)

Wiley (2000) narrowed down the definition of learning objects to "elements of a new type of computer-based instruction grounded in the object-oriented paradigm of computer science and has the potential for reusability, generativity, adaptability, and scalability". However, this definition is still fairly broad to use for video learning object because limited specifications on the content and structures of video are specified. A new description about video related learning object is needed. In addition to the general characteristics described above, we find that video learning objects have two unique features: hierarchical structure and video surrogates.

Hierarchical Structure

Video, a medium with complex spatial and temporal structure, can be analyzed from two perspectives: synthetic structure perspective and semantic perspective (Srinivasan, Gu, Tsui & Simpson-Young, 1997). The synthetic perspective focuses on the visual structure and layout of the video. Video is usually indexed in a hierarchical structure: frame, shot, and cluster of shots (Lienhart, Pfeiffer, & Effeisberg, 1997). A video frame, which is a static image that captures a particular scene in a particular time, is the basic unit of the video. Video is consisted of a sequence of individual frames which are presented at a specific speed to convey meaningful stories. A shot can be considered as a set of consecutive frames recorded using a single camera without a cut. A shot is the basic visual unit that conveys the temporal information of the video. Key frames are usually used to represent as abstracts of shots. The clustering is usually based on similarity of shots which can be evaluated by comparing the similarity of the key frames that represent each shot (Girgensohn & Borecky, 2000). Sometimes the cluster of shots is also referred to as a scene (Lienhart et al., 1997). A video is composed of a serious of scenes or stories which can be viewed as a consequence of shots that convey some particular semantic meanings, which is also referred as a story.

Semantic perspective pays more attention on the "meanings" of the video and the "content" relationships between video segments. In a class lecture video, for example, most of the scenes are about the "talking head". But the content of the talking will continuously change. Defined by Lindley (Lindley, 1997), there were four levels of video semantics: the diegetic level, which designates the sum of the video's denotation (i.e., title, video description); the connotative level, which designates the metaphorical, analogical and associative meaning of the video (i.e., frame or shot annotations); the subtextual level, which corresponds to the hidden and suppressed meanings; and the cinematic level, which is concerned with the expressive artifacts.

The hierarchical layout of the video described in terms of frames, shots, and cluster of shots is utilized, however, eventually to help represent author's semantic themes. Snoek, and Worring (2005) proposed that video reflects authors' semantic ideas from multiple modalities: visual modality, audio modality, and text modality. Visual modality refers to what can be seen in the video and is associated with chrominance, luminance, and camera movement. Audio modality refers to what can be heard in the video and is associated with loudness, rhythm, and other musical properties. Text modality refers to text information such as closed-caption and text in images. Snoek and Worring (2005) suggested that the semantic structure of a video should take account of these multiple modalities. As a consequence, five hierarchical levels of semantic representation of video are given: purpose level, genre level, sub-genre level, logical unit level, and named events level. On the purpose level, video is classified into entertainment, information, communication, or data analysis. For example, different video summary strategies should be adopted in accordance with the purpose of the video. For example, the default fast-forward speed for entertainment video should be different from the speed for information video (Wildemuth et al. 2003). All these five level of semantics were implemented in the design of video learning objects but only the first

two levels of video semantics are supported by VAST and our metadata schema. Among these entities in our video learning object, key frame and shot are two primary entities that integrate both synthetic and semantic video metadata.

Key frame

Key frames can be selected either automatically by a computer system or manually by human beings. For automatic key frame generation, many researchers suggest using one frame associated with each shot (Aoki, Shimotsuji, & Hori, 1996; Zhang, Low, & Wu, 1995) or each clustering/segment of video (Girgensohn & Borecky, 2000), while other researches propose that frames can be simply extracted at an even interval across the video document (Taniguchi, Akutsu, Tonomura, & Hamada, 1995). A straight-forward way to extract the key frame is just to select the first or middle or last frame of a shot/segment, while more complicated approaches are also proposed that can be roughly classified into two categories: one is based on the statistically significant differences of changes and the other is based on the modeling of motion or image content characteristics such as color, texture, and shape (Lindley, 1997). To illustrate the temporal changes within a shot, two or three key frames are also frequently adopted (Ng et al 2003). The disadvantages for the automatic key frame selection are also obvious: The quality may be not guaranteed and the key frame is not the really "key" frame. The same number of key frames is gained for long shots and for short shots. And significant numbers of key frames may be extracted, which may not be helpful for display on storyboards or filmstrips for quick browsing. For example, if news videos have an average length of 3.36 seconds (Christel & Warmack, 2001), one hour of news might consist of nearly five hundreds shots, which means the same number of key frames at one key frame for each shot. To reduce the number of key frames, clustering techniques are introduced to merge similar shots (Uchihashi, Foote, Girgensohn, & Boreczky ,1999). Further studies also consider temporal constraints to prevent remote similar shots from being merged (Girgensohn & Borecky, 2000). However, as learning video objects are widely used by multiple users in multiple applications, it deserves the time and efforts for high quality. As a balance of efficiency and effectiveness, a semi-automatic key frame generation method is implemented in the VAST.

Shot

Shot can be treated as a higher level entity of frames in the hierarchical learning object structure. The key issue involved is how to effectively detect the shot boundaries. Usually shot boundaries are decided by the shot transitions, which can be cut transition or gradual transition. Cut transition refers to a complete change from one shot to another. At the frame level, the last frame of the previous shot is totally different from the first frame of next shot in terms of a particular threshold value. Gradual transition, however, means the transition process is gradual due to editing effects between two shots such as fading in/out, dissolving, and wiping. The accuracy of shot detection depends on the algorithms adopted and can also be affected by camera object movements. In general, all the shot detection algorithms are based on visual features extracted from frames and how much difference exists for these features between frames. Using a predefined threshold or on an adaptive threshold, these frame differences are utilized to determine the shot boundaries. Some basic algorithms include pixel difference algorithms (Nagasaka & Tanaka, 1992), pixel level statistical difference algorithms, and histogram algorithms (Zhang et al., 1995). For compressed video, the DCT-based (discrete cosine transform) algorithms are widely used (Kuo, Lin, Chen, Chen, & Ni, 1996). In our video metadata creation, the pixel difference algorithm was adopted and implemented in the VAST.

Editing effects between shots can bring noise to the shot detection algorithms and reduce the accuracy due to differences between two consequent frames in such conditions that are smaller than the threshold values and thus fail to detect a shot boundary. However, one important feature for

gradual transition that makes it stand out from gradual changes within a shot is that the accumulated changes for a consequent frames is larger than the threshold value. Based on this observation, a so-called Twin-Comparison algorithm is proposed (Zhang et al., 1995). Two threshold values Tb and Ts are defined in the Twin-Comparison algorithm. Tb is for shot break while Ts is used for gradual transition detection. Potential transition frames with value between Ts and Tb that have a mono-increase and accumulation of difference which is eventually greater than Tb are regarded as a gradual transition Object motion can also affect shot boundary detection. However, for simplicity, we do not consider the editing effects in the VAST.

Video Surrogates

Video surrogates provide concise representations of the video while preserving the essential messages (Ding, Marchionini, & Soergel, 1999). Video surrogates are also referred to as video abstracts (Komlodi & Marchionini, 1998), video summaries (Houten, Oltmans, & Setten, 2001; Yeo & Yeung, 1997), and video abstractions (Lienhart et al., 1997). As abstractions of documents, video surrogates provide users an effective and efficient way to determine the key video content without having to view the entire video. Video surrogate includes both textual surrogate and visual surrogate. Both of them are supported in the schema of the video learning object.

Text Surrogates

Text surrogate can be video title, key words, a brief description, closed-caption, and related video identification metadata such as author, date, producer, genre, and duration. Text surrogates are the most simple and widely used video abstractions in various video applications. As compared to visual surrogates, text surrogates are strong in implying the story or semantic theme of the video (Ding et al., 1999). Even though some automatic approaches have been explored (Christel & Warmack, 2001; Jin & Hauptmann, 2000) the majority of text surrogates are still created manually due to the consideration of quality. VAST provides means that allows users to make "annotations" on different hierarchical level, including frame, shot, and video.

Visual Surrogates

Visual surrogates usually refer to video frames or a 'skimmed' video (Christel, Hauptmann, Warmack, & Crosby, 1999). Poster frame, filmstrip, storyboard, fast-forward, slide-show, and video skim (Christel, Winkler, & Taylor, 1997) are some examples of visual surrogate types that are used in current digital video libraries such as Imformedia and Open-Video to facilitate video retrieval, browsing (Li, Gupta, Sanocki, He, & Rui, 2000), and understanding. Indeed, we include three types of visual surrogates in the video learning objects: poster frame, key frames (as storyboard) and fast-forwards.

Poster frames are selected manually from the key frames that represent the shots of the video. Combined with title or key words, poster frames provide visual cues of the video such as color, texture, resolution, size, and quality of the video. As a thumbnail, a frame only consumes limited network bandwidth and thus is ideal for slow network users to do quick visual browsing.

A storyboard (or filmstrip) is a static canvas that contains a set of graphics which are usually referred to as key frames. All the key frames are arranged in one or multiple array(s) and each represents a shot or a scene. This way, the temporal integrity of the video is stretched into a static comic book. As the temporal information and structure of video is stretched into the static 2D interface, one issue for storyboards is the high cost of screen real estate (Ding et al., 1999) (Tse, Marchionini, Ding, Slaughter, & Komlodi, 1998). To solve this problem, dynamic video summaries such as fast-forwards, slideshows, some video skims are introduced as alternative video summaries.

Dynamic video summaries require less screen space as each subsequent frames is presented on the same area. Fast-forward is a common operation for VCR or DVD player which allows users watch the video N times faster than the normal speed. For digital video, a straightforward method to create the effect of fast-forward video is just increasing the frame rate, such as the JMStudio (http://java.sun.com/products/java-media/jmf/index.jsp). The drawback is, however, the limitation of the maximum frame rate and the loss of the audio track. For example, JMStudio only support eight times normal speed. Another approach is by sampling frames from the original video and then concatenating these sample frames into a new video summary.

The frames for a new video summary may be sampled by either extracting at a fix interval or extracting based on content. Fixed interval sampling approach is based on segments and is referred to as video skims (Christel, Smith, Taylor, & Winkler, 1998). For example, 10 second video segments are extracted in each 100 second segment and are concatenated to create a 10 times shorter video skim. Christel et al. (1998) indicated that fixed interval sampling video skims might delete important video content and thus they proposed content-based sampling. In fact-finding and video gisting tasks, content-based video skims using heuristic rules that considered both video and audio information achieved better user preferences Fix interval sampling (Omoigui, He, Gupta, Grudin, & Sanocki, 1999) preserves the temporal structure of the original video via evenly distributed sampling while content-based sampling emphasizes the most important shots or scenes.

The frame-based fix interval video sampling technique, which can easily generate N times faster video summaries, was used in the VAST to automatically create "candidate" key frames for the manual selection and also as the fast-forward creation technology. For example, to create a 64 times faster video summary, the frames each 64th frame is sampled and these samples are concatenated into a new 64 times shorter video summary. This approach created the exact effect of fast-forward and was tested with up to 256 times faster video summaries (Wildemuth et al., 2002, 2003), which reduce a half-hour video into only seven seconds. In fact, based on our observations, with the increase in fast-forward speed, the visual effect of fast-forward becomes more like a slideshow. .

Learning Application Tier

Learning application tier refers to the application of video learning objects in heterogeneous learning systems. The reusability of the video learning objects makes video metadata easily to be applied to different systems. In other words, once the objects are created, they can be applied to different video applications. In our study, the video learning objects created by using the VAST system have been applied in the Open-Video digital library, user studies, and distance learning (using ISEE).

The Open-Video Project has more than eighteen thousand digitalized video segments with lengths from several seconds to nearly an hour. Key frames are extracted and the VAST is used to create storyboards. At this time, fast forward surrogates have been created for about half of the whole video collections.

The VAST was also utilized to create video objects for two user studies conducted in the Interactive Design Lab (IDL), University of North Carolina at Chapel Hill. In the first study, five types of surrogates including fast-forwards were evaluated in terms of their usefulness and usability in accomplishing specific information seeking tasks (Wildemuth et al., 2002). The compression ratio of the fast-forward was sixteen times faster. In the second study five fast-forward speeds (16x, 32x, 64x, 128x, and 256x) were studied in attempt to answer a simple question: how fast is too fast? (Wildemuth et al., 2003)

Public Data Interface

The reusability of the video learning objects means that they can be applied to different environments. It also means these heterogeneous applications need a standard Public Data Interface (PDI) in order to convert the video learning objects into the format that these applications can directly use. An initial effort has been made towards the creation of the PDI module based on the video metadata schema we developed. A sample piece of this video learning object schema is given as follows.

```
<program xmlns = http://www.ils.unc.edu/VAST/ExampleSchema>
<generalInfo>
    <title>...</title>
    <annotation>...</annotation>
    ...
</generalInfo>
<visualInfo>
<keyFrames datatype = "IntegerVector" size = "16" >
<kFrame >
    <sequenceNo datatype = "unsigned8"> 1 </sequenceNo>
<annotation> ... </annotation>
    <timestamp datatype = "RelTimePoint"  minValue = "0.0">
...
</timestamp>
    <duration datatype = "FractionalDuration" > ... </duration>
            ...
</kFrame>
...
</keyFrames>
<visualFF >
<location datatype = "String"> ... </location>
<startPoint datatype = "RelTimePoint"> 0.0 </startPoint>
<endPoint datatype = "RelTimePoint"> ...< /endPoint>
<ffDuration datatype = "RelTimePoint"> ..</ffDuration>
<compressionRatio datatype="unsigned8" > ...</compressionRation>
<generalInfo>
    <author> ..</author>
    ...
</generalInfo>
<format datatype= "String"> QuickTime </format>
...
</visualInfo>
...
</program>
```

Multi-channel Synchronization

One application of video learning object is to use the timestamp attributes for multi-channel synchronization. A video has multiple information channels (or tracks): visual channel, audio channels, and text channels. The audio channels may include background music channel, narrative channel, and/or conversation channels. The text channels may include title, author, text abstraction, closed caption, and texts that embedded in video. These channels are not presented inde-

pendent. They are correlated with each other along the temporal dimension. In other words, all these information channels can be coordinated or synchronized using the video timestamps.

In learning application, particular channel information with a particular timestamp is frequently requested. For example, in real-time video-based online education, the instructor's audio needs to be synchronized with the relative PowerPoint slide. When a specific slide is referred by learners, the associated audio channel will probable be also needed. Such a multi-channel synchronization function can be fully demonstrated in the Interactive Shared Educational Environment (ISEE) we developed for collaborative video-based distance learning.

Interactive Shared Educational Environment (ISEE)

ISEE is an advanced video application system that supports highly interactive collaboration distance learning (Mu & Marchionini, 2002). The data used in ISEE is the same video learning objects for the Open-video project and is created from the VAST. Figure 3 illustrates the Graphic User Interface (GUI) of the ISEE. A Video Player, an Interactive Chat Room (ICR), a video Storyboard, and a Shared Web Browser (SWB) are supported in the ISEE system.

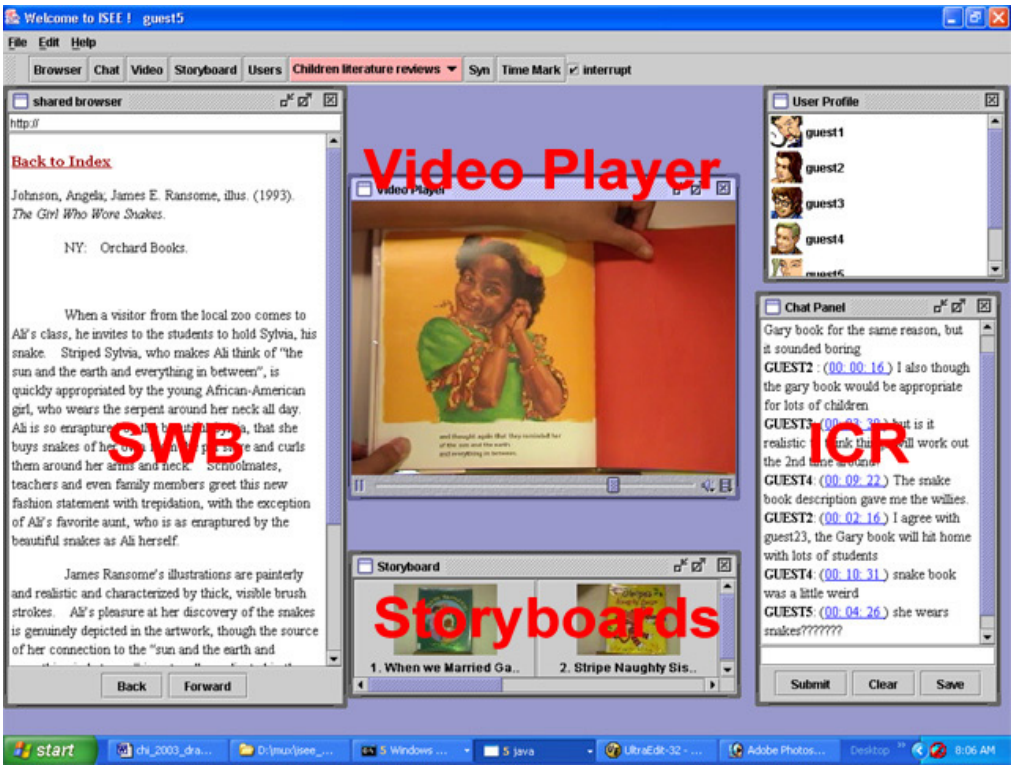


Figure 3: The User Interface of ISEE

The four learning and communication components are not only juxtaposed in appearance, but also coherently connected with each other through the "smartlink" mechanism. For example, a single click on a specific frame on the storyboard will update the video player to the same time stamp and begin to play there. Timestamp is integrated in the video learning objects and was loaded by ISEE into the storyboard. The video player accepts the signals passed from the storyboard and

triggers actions to start, stop, jump back, or jump forward. Another example is the link between the video player and the Interactive Chat Room (ICR).

In ISEE, the Interactive Chat Room (ICR) is different from the generic online text chat tools in terms of supporting interaction between users via messages. Each message sent by a user contains not only the sender's username and the message content, but also the sender's video timestamp representing the point when the message was sent (see Figure 4). A single click on the timestamp by a particular user immediately updates that user's local video player to the point of that time-stamp. In consequence, both the text messages and the associated video contexts are shared across users.

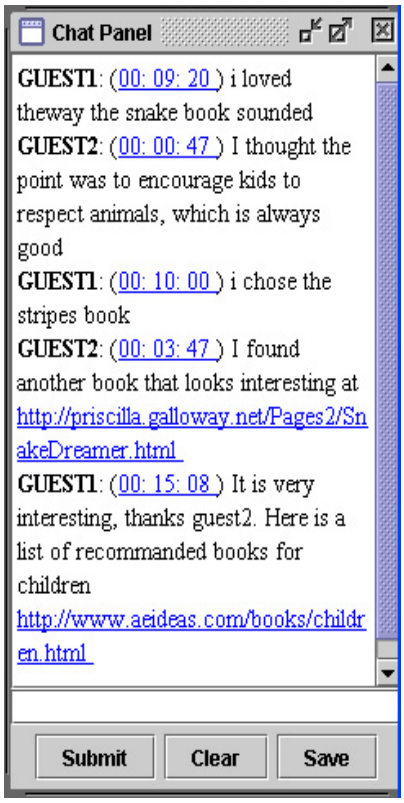


Figure 4: Interactive Chat Room with Smartlinks

Table 1 lists the currently available smartlinks. All these smartlinks are supported by using the video timestamps in the video learning objects. Any operations on these smartlinks are logged automatically into the database for later use (i.e., usability study, or access by later video viewers). The log data is also organized based on the same metadata schema and can be referred as new video learning objects.

Table 1 Smartlinks across ISEE components

	ICR	SWB	Storyboard	Video Player
Interactive Chat Room		√		√
Shared Web Browser	√			√
Storyboard				√
Video Player	√	√	√	

Conclusions and Future Work

An increasing number of video-based applications such as online digital video libraries, video-based distance learning, and technical training, have been developed to take advantage of the enormous collections of online digital videos. For a particular video application, not only the video, but also the video metadata information such as video description and video surrogates, are usually needed. The challenge, however, is that the video and video metadata are in most case tightly connected together, thus excluded another video application to directly utilize the same video metadata, which is usually expensive to create due to the requirements of human process. For the purpose of reusability, video metadata and video segments need to be reconstructed in a standard format. The primary concept of the three-tier infrastructure is to decouple the tight connection between video metadata and the associated application by introducing an intermedia tier—video learning object tier, to achieve the reusability and interoperability of video metadata.

A standard XML video metadata schema was developed based on Dublin Core and MPEG-7 standards to represent video learning objects. To facilitate the video metadata creation, a semi-automatic tool, VAST, was also developed. VAST supports both the video identification and video description metadata at two dimensions: synthetic and semantic video metadata. Video surrogates, particularly video frames and video shot, are generated automatically by VAST, followed by manually selection of the key frames and the creation of the fast-forward. A hierarchical infrastructure is adopted for the entities of video learning objects. Our initial application of this infrastructure and the video learning objects in two video applications, Open-Video project and ISEE project, proved to be successful.

However, some new challenges are also coming to the horizon while we implement the "three-tier" infrastructure and apply it in real applications. One challenge is the compatibility. Organizations such as ADL (http://www.adlnet.org/), CanCore (http://www.cancore.ca/indexen.html), Dublin Core, IEEE LTSC LOM (http://ltsc.ieee.org/), IMS (http://www.imsproject.org/), SMPTE (The Society of Motion Picture and Television Engineers), INDECS (www.indecs.org) ,TV-Anytime(http://www.tv-anytime.org/) and MPEG are involved in developing video metadata schemas. But none of these schemas are exclusively for video learning objects. In our effort to develop our own video learning object schema, we found it was fundamentally difficult to make it compatible with all the major video metadata schemas. Eventually we decide to choose the Dublin Core and MPEG-7 standards as the bases due to their popularity, scalability and their strength on both video content identification and video content description. We will continue to work on this issue to make our schema to be compatible with other major metadata standards.

Another challenge is related to the work of the so-called "self-learning" function of the video learning object. With this feature, feedback information about the video learning object usages in different applications will be automatically re-organized based on our video metadata schema and integrated into that particular learning object. For example, if two seemingly unrelated shots are frequently selected simultaneously by different applications (or users) for viewing or discussing, it might imply that there exist some kinds of semantic link between them, even though synthetically there is no relationship being found. As a consequence, these two shots be clustered together and a link will be automatically created in the learning object. This feature can be achieved by adding application-related attributes (i.e., application links) to the video learning objects. Given the huge amount of feedback information we collected (through application logs), the difficult is to decide the most important patterns of usage and present them in the video learning object, without leading the object to unacceptable size.

References

Aoki, H., Shimotsuji, S., & Hori, O. (1996). A shot classification method of selection effective key-frames for video browsing. In *Proceedings of Multimedia,*1-10.

Brotherton, J. A. (2001). Enriching everyday experiences through the automated capture and access of live experiences: eClass: Building, observing and understanding the impact of capture and access in an educational domain (Doctor thesis, College of Computing, Georgia Institute of Technology, 2001).

Cadiz, J., Balachandran, A., Sanocki, E., Gupta, A., Grudin, J., & Jancke, G. (2000). Distance learning through distributed collaborative video viewing. *Proceedings of the 2000 ACM conference on computer supported cooperative work,* 135-144.

Christel, M. G., Hauptmann, A. G., Warmack, A. S. & Crosby, S. A. (1999). Adjustable filmstrips and skims as abstractions for a digital video library. *Proceedings of IEEE forum on Research and Technology Advances in Digital Libraries* 19-21 May 1999, Baltimore, Maryland.

Christel, M. G., Smith, M. A., Taylor, C. R., & Winkler, D. B. (1998). Evolving video skims into useful multimedia abstractions. *Proceedings of the SIGCHI conference on Human factors in computing systems,*171-178.

Christel, M., & Warmack, A. S. (2001), The effect of text in storyboards for video navigation. *Proceeding of.IEEE international conference on acoustics, speech, and signal processing*. Salt Lake City, UT, May 2001. Vol. III. pp.1409-1412.

Christel, M., Winkler, D., & Taylor, R. (1997). Multimedia abstractions for a digital video library**.** *Proceedings of the second ACM international conference on Digital libraries,* 21-29.

Ding, W., Marchionini, G., & Soergel, D. (1999). Multimodal surrogates for video browsing. *Proceedings of the Fourth ACM conference on Digital Libraries*, Berkeley, CA, August 11-14.

Dublin Core, (2001), Dublin Core Metadata Glossary, Final draft, Feb. 24, 2001. Retrieved November 23, 2004 from http://library.csun.edu/mwoodley/dublincoreglossary.html

Girgensohn, A., & Borecky, J. (2000). Time-constrained keyframe selection technique. *Multimedia Tools and Applications, 11*, 347-358,

Houten, Y. V., Oltmans, E., & Setten, M. V. (2001). Video browsing & summarization. Telematica Instituut, TI/RS/2000/163

Jin, R., & Hauptmann, A. G. (2000). Title generation for spoken broadcast news using a training corpus. ICSLP 2000, October 16-20, Beijing, China.

Komlodi, A, & Marchionini, G. (1998). Key frame preview techniques for video browsing. *Proceedings for the third ACM conference on Digital Libraries,* 67-75.

Kuo, T. C. T., Lin, Y. B., Chen, A. L. P., Chen, S.-C., & Ni, C. Y. (1996) .Efficient shot change detection on compressed video data. *Proceeding of International Workshop on Multimedia Database Management Systems*, 101–108.

Lienhart, R., Pfeiffer, S., & Effeisberg, W. (1997). Video abstracting. *Communications of the ACM, 40* (12), 55-62.

Li, F. C., Gupta, A., Sanocki, E., He, L., & Rui, Y (2000). Browsing digital video. *Proceedings of the SIGCHI conference on Human factors in computing systems,* 2(1), 169-176.

Lindley, C. A. (1997). Multiple-interpretation framework for modeling video semantics. *Workshop on Conceptual Modeling in Multimedia Information Seeking*, Los Angeles, 6-7 November 1997.

LOM (2004). Learning Object Metadata WG12. Retrieved November 23, 2004 from http://ltsc.ieee.org/wg12/

Marchionini, G. & Geisler, G. (2002). The Open Video Digital Library. *D-Lib Magazine*, *8* (12). Retrieved November 24, 2004 from http://www.dlib.org/dlib/december02/marchionini/12marchionini.html

MPEG: 2001, Overview of the MPEG-7 Standard, ISO/IEC JTC1/SC29/WG11 N4031, Singapore, March 2001, Retrieved November 23, 2004 from http://ipsi.fhg.de/delite/Projects/MPEG7/Documents/w4031mpeg7overview.htm

Mu, X. & Marchionini, G. (2002). Interactive shared educational environment (ISEE): Design, architecture, and user interface. Technical Report, School of Information and Library Science, University of North Carolina at Chapel Hill, TR-2002-01.

Nagasaka, A., & Tanaka, Y. (1992). Automatic video indexing and full-video search for object appearances. *Visual Database Systems*, 113-127.

Ng, T., Christel, M., Hauptmann, A., & Wactlar, H. (2003). Collages as dynamic summaries of mined video content for intelligent multimedia knowledge management. *Spring Symposium Series on Intelligent Multimedia Knowledge Management*, Palo Alto, CA, March 24-26.

Olsen, S. (2004). Striking up digital video search. From *News.com* . Retrieved November 27, 2004 from http://news.com.com/Striking+up+digital+video+search/2100-1032_3-5466491.html

Omoigui, N., He, L., Gupta, A., Grudin, J., & Sanocki, E. (1999). Time-compression: systems concerns, usage, and benefits. *Proceedings of the SIGCHI conference on Human factors in computing systems,*136-143.

Snoek, C. & Worring, M., (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications, 25* (1), 5-35.

Srinivasan, U., Gu, L., Tsui, K. & Simpson-Young, W. G. (1997). A data model to support content-based search in digital videos. *Australina Computer Journal*, *29* (4),141-147.

Taniguchi, Y., Akutsu, A., Tonomura, Y., & Hamada, H. (1995). An intuitive and efficient access interface to real-time incoming video based on automatic indexing. *Proceeding of the third ACM international conference on Multimedia*, 25-33.

Tse, T. , Marchionini, G., Ding, W., Slaughter, L., & Komlodi, A.(1998). Dynamic key frame presentation techniques for augmenting video browsing. *Proceedings of the Advanced Visual Interfaces International Working Conference*. May 24-27, 1998, L'Aquila, Italy.

Uchihashi, S., Foote, J., Girgensohn, A., & Boreczky, J. (1999). Video mange: generating semantically meaningful video summaries. *Multimedia'99*, Oct. Orlando.

Wildemuth, B. M., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B., Hughes, A., & Mu, X. (2002). Alternative surrogates for video objects in a digital library: users' perspectives on their relative usability. *European Conference on Digital Libraries (ECDL) 2002*, June.

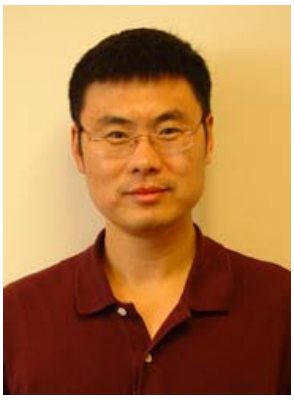
Wildemuth, B., M., Marchionini, G., Yang, M., Geisler,G., Wilkens. T., Hughes, A., & Gruss, R. (2003). How fast is too fast? Evaluating fast forward surrogates for digital video. *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 221-230.

Wiley, D. A. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D. A. Wiley (Ed.), *The instructional use of learning objects: Online Version*. Retrieved November 23, 2004 from: http://reusability.org/read/chapters/wiley.doc

Yeo,B. & Yeung, M. (1997). Retrieving and visualizing video. *Communicaitons of the ACM, 40* (12), 43-52.

Zhang, H. J., Low, C. W., & Wu, J. H.(1995). Video parsing, retrieval and browsing: an integrated and content-based solution. *Proceedings of the third ACM international conference on Multimedia*, 15-24.

Biography



Xiangming Mu, an assistant professor from School of Information Studies, University of Wisconsin—Milwaukee, earned his Ph.D in Information and Library Science from University of North Carolina. He worked in multiple research projects supported by National Science Foundation (NSF) and developed a suite of new information systems: Interactive Shared Educational Environment (ISEE) to support the Video-based Collaborative Distance Learning, Browser for Enriched Statistical Tables (BEST) to help non-specialists access and use statistical data in tabular form, Video Annotation and Summarization Tool (VAST) to facilitate the creation of both semantic and visual metadata, and Table-Hunter to support context + focus view of a complex web site. His current research interests include: learning technology and learning objects, digital library, multimedia information retrieval, human-computer interaction, and data mining.