RC 22463 (W0205-197) May 24, 2002
Computer Science

IBM Research Report

Video Personalization System for Usage Environment

Belle L. Tseng, Ching-Yung Lin, John R. Smith
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 703
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Video Personalization System for Usage Environment

Belle L. Tseng, Ching-Yung Lin, and John R. Smith

IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532
{belle, cylin, jrsmith}@watson.ibm.com

ABSTRACT

A video personalization and summarization system is designed and implemented incorporating usage environment to dynamically generate a personalized video summary. The personalization system adopts the three-tier server-middleware-client architecture in order to select, adapt, and deliver rich media content to the user. The server stores the content sources along with their corresponding MPEG-7 metadata descriptions. Our semantic metadata is provided through the use of the *VideoAnnEx* MPEG-7 Video Annotation Tool. When the user initiates a request for content, the client communicates the MPEG-21 usage environment description along with the user query to the middleware. The middleware is powered by the personalization engine and the content adaptation engine. Our personalization engine includes the *VideoSue* Summarization on Usage Environment engine that selects the optimal set of desired contents according to user preferences. Afterwards, the adaptation engine performs the required transformations and compositions of the selected contents for the specific usage environment using our *VideoEd* Editing and Composition Tool. Finally, two personalization and summarization systems are demonstrated for the IBM Websphere Portal Server and for the pervasive PDA devices.

Keywords: video personalization, summarization, adaptation, usage environment, MPEG-7, MPEG-21, annotation, rich media description, digital item adaptation, rights expression.

1. INTRODUCTION

With the growing amount of multimedia content, people become more willing to view personalized multimedia based on their usage environments. When people use their pervasive devices, they generally restrict their viewing time on the limited displays and minimize the amount of interaction and navigation to get to the content. When they browse video on the Internet, they may want to get only the videos that match their preferences. Because of the existence of heterogeneous user clients and data sources, it is a real challenge to implement a universally compliant system that fits various usage environments. Most existing video summarization tools address their applications on proprietary environments. Some systems display summarized videos using a number of key frames for each detected scene shot to generate a storyboard [16]. Merialdo *et.al.* generate personalized TV news programs based on user preference and time constraint [9]. Gong and Liu use Singular Value Decomposition (SVD) of attribute matrix to reduce the redundancy of video segments and thus generate video summaries [4]. In addition to the systems based on audio-visual information, some researchers have proposed methods to detect semantic important events based on other resources. For instance, Aizawa *et. al.* use brain waves to detect exciting moments of the subjects [1]. In industry, companies such as NTT DoCoMo [13] and Virage [15] have implemented preliminary video summarization systems for cellular phones with sports highlights.

This paper addresses issues of designing a video personalization and summarization system in heterogeneous usage environments and provides a tutorial introduction on their associated issues in MPEG-7 and MPEG-21. The server maintains the content sources, the MPEG-7 metadata descriptions, the MPEG-21 rights expressions, and content adaptability declarations. The client communicates the MPEG-7 user preference, MPEG-21 usage environments, and user query so as to retrieve and display the personalized content. The middleware is powered by the personalization engine and the adaptation engine. The personalization system adopts the three-tier server-middleware-client architecture in order to select, adapt, and deliver rich media content to the user.

This paper is organized as follows. In Section 2, we describe the personalization and summarization framework that is currently evolving in MPEG-21 and what had been included in MPEG-7. In Section 3, we describe our video personalization and summarization system, which includes the user client, the personalization and adaptation media middleware, and the database server. Details of these three tiers are described in Sections 4, 5, and 6.

2. DIGITAL ITEM ADAPTATION USING MPEG-7 AND MPEG-21

In order to personalize delivery and consumption of multimedia content for the individual users, a phrase coined by the ISO/IEC MPEG-21 group called *digital item adaptation* has been identified as the essential framework for personalization. Adaptation aims at providing the appropriate content types to the different terminals via the preferred networks. *MPEG-21 Digital Item Adaptation* provides the tools to support resource adaptation of content, descriptor adaptation through metadata, and quality of service management [11].

Figure 1 illustrates the fundamental components of digital item adaptation. The *content sources* may include text, audio, images, video, and other multimedia. Associated with each source is the corresponding *digital item declaration* that identifies and describes the content. Digital item declaration also includes components for resource adaptation, rights expression and adaptation rights of the content. *MPEG-21 Digital Item Declaration* provides a standard description for digital items. Multimedia content sources can also be described using the *MPEG-7 Media Descriptions*, which complement the components in the *MPEG-21 Digital Item Declaration.* MPEG-7 Media Descriptions describe audio-visual content using features, semantics, structures, and models, then represent them in XML. MPEG-7 also includes adaptation hints and that are covered under the *MPEG-7 Media Resource Requirements.* In addition, MPEG-21 has a similar requirement for content adaptability called *MPEG-21 Media Resource Adaptability,* and thus overlaps with MPEG-7 as illustrated in the figure [10].

Rights expression is another subcomponent of *MPEG-21 Digital Item Declaration* and describes the rights associated with multimedia content. They include the rights and permissions for an individual or designated group to view, modify, copy or distribute the content. Among these expressions that are relevant to personalization is *adaptation rights.* For example, the owner of an image may allow cropping of the image to fit the desired display size, but not scaling.

From the perspective of the user, the *usage environment* defines the user profiles, terminal properties, network characteristics, and other user environments. *MPEG-21 Usage Environment* is in the process of defining specific

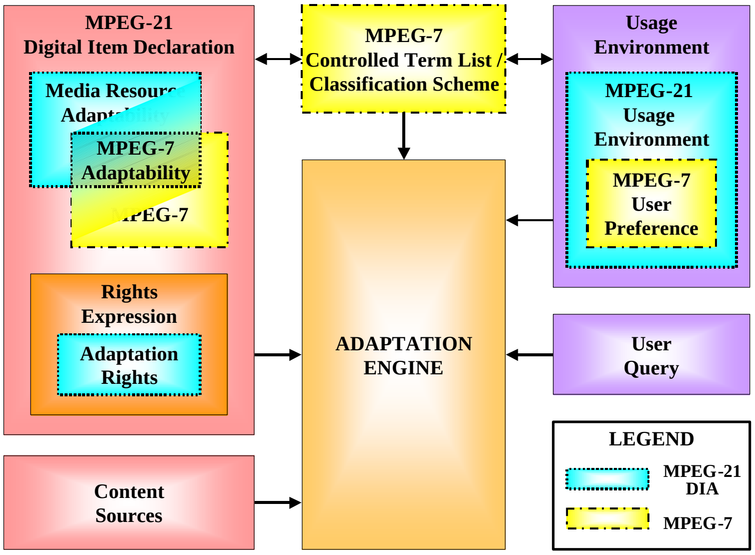


Figure 1: Block Diagram of Digital Item Adaptation Using MPEG-7 and MPEG-21. The adaptation engine matches the content sources and corresponding digital item declarations with the user query and usage environment through a controlled term list in order to (1) optimally select the set of personal content and (2) effectively adapt that dynamic content for the user.

requirements to describe such conditions. The *usage environment* also includes user preferences, which can partially or fully be described by the *MPEG-7 User Preferences* and is adopted as a subcomponent of *MPEG-21 Usage Environment* [10]. Furthermore, the user can request for a specific content in a special application not defined by the standard. This request is represented as the *user query* in the figure.

To match the user's usage environment with the content's digital item declaration, the *adaptation engine* is introduced to personally select the desired content and optimally adapt those content in accordance with the descriptions. Specifically, users specify their requests for content through the user query and usage environment. On the other hand, the digital item declaration encompasses both the rights expression and the content adaptability of the available content. To assist the matching process, a common language can be provided through the *MPEG-7 Controlled Term List*. This list allows specific domain applications to define their taxonomy, their relationships, and their importance. In summary, the adaptation engine takes the user query, the usage environment, the digital item declaration, and the controlled term list to adapt the content sources to generate the personalized content.

In the rest of this section, we describe more details of the MPEG-7 and MPEG-21 components in digital item adaptation for usage environment, media description, rights expression, and content adaptability.

2.1 Media Descriptions

Media descriptions identify and describe the multimedia content from different abstraction levels of low-level features to various interpretations of semantic concepts. MPEG-7 provides *Description Schemes* (DS) to describe content in XML to facilitate search, index and filtering of audio-visual data [7]. The DS's are designed to describe both the audio-visual data management and the specific concepts and features. The data management descriptions include metadata about creation, production, usage, and management. The concepts and features metadata may include what the scene is about in a video clip, what objects are present in the scene, who is talking in the conversation, and what is the color distribution of an image.

MPEG-7 standardizes the description structure; however, technical challenges remain. The generation of these descriptions is not part of the MPEG-7 standard and the technologies for generating them are variable, competitive, and some—non-existent. As such we have developed an annotation tool to describe the semantic meaning of video clips to capture the underlying semantic meaning by a human. These annotators are assisted in their annotation task through the use of a finite vocabulary set. This set can be readily represented by the *MPEG-7 Controlled Term List*, which can be customized for the different domains and applications. For example in a news program, the list can include the anchors' names, topic categories, and camera captures. Our annotation tool is described in Section 4.

2.2 Usage Environment

The usage environment holds the profiles about the user, device, network, delivery, and other environments. This information is used to determine the optimal content selection and the most appropriate form for the user. Other than MPEGs, several standards have been proposed. HTTP/1.1 uses the *Composite Capabilities / Preference Profile (CC/PP)* to communicate the client profiles. The forthcoming Wireless Access Protocol (WAP) proposes the *User Agent Profile (UAProf)* includes the device profiles, which covers the hardware platform, software platform, network characteristics, and browser [3].

The MPEG-7 *UserPreferences DS* allows users to specify their preferences (likes and dislikes) for certain types of content and for ways of browsing [7]. To describe the types of desired content, the *FilteringAndSearchPreferences DS* is used that consists of the creation of the content (*CreationPreferences DS*), the classification of the content (*ClassificationPreferences DS*), and the source of the content (*SourcePreferences DS*). To describe the ways of browsing the selected content requested by the user, the *BrowsingPreferences DS* is used along with the *SummaryPreferences DS*. Furthermore, each user preference component is associated with a preference value indicating its relative importance with respect to that of other components. For instance, we can have a user preference description to encapsulate the preference ranking among several genre categories (*ClassificationPreferences DS*) that are produced in the United States in the last decade (*CreationPreferences DS*) in wide screen with Dolby AC-3 audio format (*SourcePreferences DS*). And, the user is also interested in nonlinear navigation and access of the retrieved summarization content by specifying the preferred duration and preference ranking (*SummaryPreferences DS*).

The MPEG-7 *User Preferences Descriptions* specifically declares the user's preference for filtering, search, and browsing of the requested multimedia content. But, other descriptions may be required to account for the terminal, network, delivery, and other environment parameters. The MPEG-21 *Usage Environment Descriptions* cover exactly these extended requirements even thought the specific requirements are still currently being defined and refined [11]. The descriptions on terminal capabilities include the device types, display characteristics, output properties, hardware properties, software properties, and system configurations. This allows the personalized content to be appropriately adapted to fit the terminal. For instance, videos can be delivered to wireless devices in smaller image sizes in a format that the device decoders can handle. The descriptions on the physical network capability allow content to be dynamically adapted to the limitation of the network. The descriptions on the delivery layer capabilities include the types of transport protocols and connections. These descriptions allow users to access location-specific applications and services. The MPEG-21 *Usage Environment Descriptions* also include *User Characteristics*, that describe service capabilities, interactions and relations among users, conditional usage environment, and dynamic updating.

2.3 Content Adaptability

Content adaptability refers to the multiple variations that a media can be transformed into, either through changes in format, scale, rate, and/or quality. Format transcoding may be required to accommodate the user's terminal devices. For example, an MPEG movie cannot be rendered on the Palm III due to lack of an MPEG decoder thus must be transcoded into bitmap images to be rendered as a slide show. Scale conversion can represent image size resizing, video frame rate extrapolation, or audio channel enhancement. Rate control corresponds to the data rate for transferring the media content, and may allow variable or constant rates. Quality of service can be guaranteed to the user based on any criteria including SNR or distortion quality measures. These adaptation operations transform the original content to efficiently fit the usage environment.

The *MPEG-7 Media Resource Requirement* and the *MPEG-21 Media Resource Adaptability* both provide descriptions for allowing certain types of adaptations. These adaptation descriptions contain information for a single adaptation or a set of adaptations. The descriptions may possibly include required conditions for the adaptation, permissions for the adaptation, and configurations for the adaptation operation.

3. SYSTEM OVERVIEW

Our Video Personalization System comprises of three major components, *the user client*, *the database server*, and *the media middleware*. Figure 2 illustrates the block diagram of these components. The user client component allows the user to specify his or her preference query along with the usage environment, and receives the personalized content on the display client. The database server component stores all the content sources as well as their corresponding MPEG-7 media descriptions, MPEG-21 rights expressions, and content adaptability declarations. The media middleware represents the intermediate component of the figure, and processes the user query and the usage environment with the media descriptions and rights expressions to generate a personalized content, which is appropriately adapted and transmitted to the user's display client.

In the client end of our system, a user can request for personalized content by specifying a user query and communicating his/her usage environment to the media middleware. The user query takes the form of preference topics, certain keywords, and the user's time constraint for watching the content. The usage environment includes descriptions about the client terminal capabilities, physical network properties, and delivery layer characteristics. The display client can range from a pervasive mobile device to a networked high-end workstation. The customized content is optimally selected, summarized and adapted in the middleware and personally delivered to the user.

The database server provides the media middleware with the content sources and their corresponding descriptions. Each content is associated with a set of media descriptions, rights expressions, and adaptability declarations. In the database server, the content sources can include text, audio, image, video, and graphics, and are

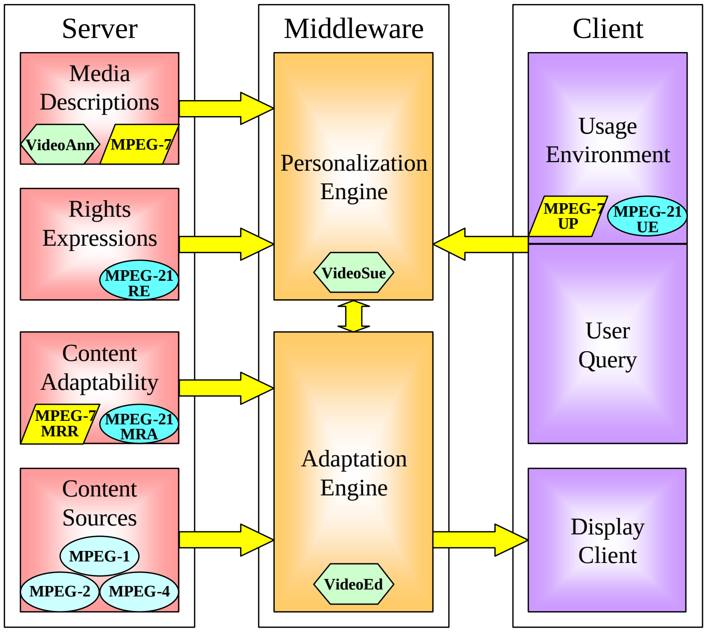


Figure 2: Block Diagram of Video Personalization System.
The three major components are *the user client*, *the database server*, and *the media middleware*.

stored in various formats (*e.g.*, MPEG-1/2/4, AVI, and QuickTime). The corresponding media descriptions include feature descriptions as well as semantic concepts. These semantic descriptions can be semi-automatically generated by our *VideoAnnEx* MPEG-7 annotation tool. For each content, there may be an associated set of rights expressions on the various usages of the media. These expressions define the right for others to view, copy, print, modify, or distribute the original content. Similarly, there is an associated set of content adaptability declarations on the possible variation transformations that can be applied to the media.

The media middleware interfaces the user client and the database server. The middleware consists of the personalization engine and the adaptation engine. The media descriptions and rights expressions are used by the personalization engine to optimally select the required sources for the personalized content. Afterwards, the selected content sources, stored in SMIL, ASX, or MPEG-4 XMT format, and corresponding content adaptability declarations are sent to the adaptation engine to generate the personalized content. In the personalization engine, the user query and usage environment are matched with the media descriptions and rights expressions to generate the personalized content. In the adaptation engine, the results of the personalization engine are used to retrieve the corresponding content sources and content adaptability declarations. Our adaptation engine, called *VideoSue*, determines the optimal variation (i.e., format, size, rate, quality) of the content for the user in accordance with the adaptability declarations and the inherent usage environment. Subsequently, the appropriate adaptation is performed on the media content and delivered to the user's display client. Our adaptation engine consists of multiple transcoders including our *VideoEd* editing and composition tool and *Universal Tuner* video transcoder.

4. DATABASE SERVER

The content sources are stored, described, analyzed and annotated with MPEG-7 and MPEG-21 descriptions. Figure 2 illustrates the database server with four major components. First, the database can store video sources in any format. Our media middleware accepts various kinds of video file types, bit-rates and resolutions. Also, each video is only required to be stored once in any format, because the middleware can dynamically access and transcode the desired video segments in the sequences and generate one composite video in real time. Even though each output video is a composition of multiple video segments, our database do not need to pre-segment the videos. This brings significant benefits to content and storage management.

Second, the database stores media descriptions in MPEG-7 format. A semi-automatic annotation tool assists in the generation of such descriptions. The annotation can range from high-level semantic concepts to low-level feature descriptions. We implemented a *VideoAnnEx* MPEG-7 annotation tool for authors to annotate MPEG video content with semantic descriptions. The *VideoAnnEx* is one of the first MPEG-7 annotation tools being made publicly available. The tool explores a number of interesting capabilities including automatic shot detection, key-frame selection, automatic label propagation to "similar" shots, and importing, editing, and customizing of ontologies and controlled term lists. In the rest of this section, further descriptions are provided in regard to the video segmentation, semantic lexicon, and the *VideoAnnEx* tool.

Third, the database maintains rights management on the usage of content sources as MPEG-21 Rights Expressions. Content owners can specify rights expressions for an individual or group to view, copy, print, modify, and/or distribute the video. Finally, the database also stores content adaptability descriptions as MPEG-7 Media Resource Requirement and MPEG-21 Media Resource Adaptability. Content adaptability specifies requirements on the adaptability parameters and constraints for changing the original content to a personalized derivative for the user.

4.1 Video Segmentation

In general, a short video clip can be annotated by simply describing its content in its entirety. However when the video is longer, annotation of its content can benefit from segmenting the video into smaller units. A video shot is defined as a continuous camera-captured segment of a scene, and is usually well defined for most video content. Given the shot boundaries, the annotations are assigned for each video shot.

Shot boundary detection is performed to divide the video into multiple non-overlapping shots. We use either the IBM *CueVideo* Toolkit or the built-in functionality in *VideoAnnEx* to perform the shot boundary detection. The *CueVideo*, which is based on the multiple timescale differencing of the color histogram, segments our video content into shorter shots, where scene cuts, dissolves, and fades are effectively detected [2]. Our *VideoAnnEx* tool can also perform the video segmentation based on color histogram distributions.

4.2 Video Content Semantic Lexicon

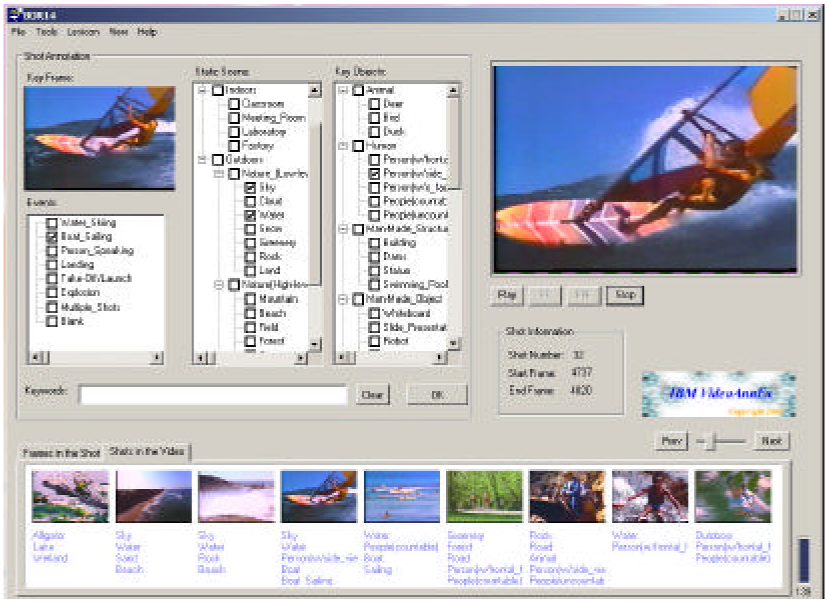
In general, a video shot can fundamentally be described by three attributes. The first is the background surrounding of where the shot was captured by the camera, which is referred to as the *static scene*. The second attribute is the collection of significant subjects involved in the shot sequence, which is referred to as the *key object*. Lastly, the third attribute is the corresponding action taken by some of the key objects, which is referred to as the *event.* These three types of lexicon define the vocabulary for our video content.

According to the characteristics of video corpus, a pre-defined lexicon set can be imported into our IBM *VideoAnnEx*. The shots are labeled with respect to the selected lexicon. Users can also associated each label to the region levels in the keyframe of the shot. Note that the lexicon, whose format is compatible with MPEG-7, is dependent on the summarization application, and can be modified, imported, and saved using the *VideoAnnEx*.

4.3 *VideoAnnEx* – MPEG-7 Video Annotation Tool

We develop an MPEG-7 video annotation tool, *VideoAnnEx*, to assist authors in annotating video sequences with MPEG-7 metadata. Each shot in the video can be annotated according to some lexicon sets. *VideoAnnEx* takes an MPEG video sequence as the required input source. It requires a corresponding shot segmentation file, which can be loaded into the tool from other sources or generated when the input video is first opened.

The *VideoAnnEx* annotation tool is divided into four graphical sections as illustrated in Figure 3. The *Video Playback* window displays the video sequence. As the video is played back in the display window, the current shot information is given as well. The *Shot Annotation* module displays the defined semantic lexicons and the key frame window. The *Views Panel* displays two different previews of representative images of the video. The *Frames in the Shot* view shows all the I-frames as representative images of the current video shot, while the *Shots in the Video* view (as in the bottom of Figure 3) shows the key frames of shots over the entire video. As the annotator labels each



Shot Information
Shot Number: 32
Start Frame: 4737
End Frame: 4820

Figure 3: The IBM VideoAnnEx MPEG-7 Video Annotation Tool is divided into Four Regions: (1) Video Playback, (2) Shot Annotation, (3) Views Panel, and (4) Region Annotation (not shown).

shot, the descriptions are displayed below the corresponding key frames. A more detailed description of the annotation tool as well as its active learning components are shown in [12][14].

In our Video Personalization System, a relevance score is automatically assigned to the video based on the confidence value of the classification. For our system, the annotation process generates a relevance score for the whole video sequence and for each attribute based on the probability of that attribute to the corresponding video unit. After these steps, users can manually correct the annotation as well as the scene boundaries. All these results are then saved as an MPEG-7 XML file.

5. MEDIA MIDDLEWARE

The media middleware consists of the personalization engine and the adaptation engine as illustrated in the system overview of figure 2. In the personalization engine, the user query and usage environment are matched with the media descriptions and rights expressions to generate the personalized content. The adaptation engine determines the optimal variation (*i.e*., format, size, rate, quality) of the content for the user in accordance with the adaptability declarations and the inherent usage environment.

5.1 Personalization Engine

The objective of our system is to show a shortened video that maintains as much semantic content within the desired time constraint. Figure 4 illustrates an overview of the video personalization and summarization engine, where the user preference is matched against the MPEG-7 data to generate the personalized video summary list.

In the database, every video shot is annotated with MPEG-7 descriptions. In addition, each video source is declared with MPEG-21 rights expressions for viewing and distribution attributes. Assuming that the appropriate rights are granted to use the content, the personalization engine extracts the MPEG-7 media descriptions and matches them against the MPEG-7 user preferences, MPEG-21 usage environment, and user query. The matching process results in some ranking scores for the video sources, and is used to determine the selected set of video segments for the personalized video summary. The *VideoSue* engine performs this process and is described next.

5.1.1 *VideoSue* – Video Summarization on Usage Environment

Our *VideoSue* stands for Video Summarization on Usage Environment. This engine takes MPEG-7 metadata descriptions from our content sources along with the MPEG-7/MPEG-21 user preference declarations and user time constraint to output an optimized set of selected video segments, which will generate the desired personalized video summary [14]. Using shot segments as the basic video unit, there are multiple methods of video summarization based on spatial and temporal compression of the original video sequence. In our work, we focus on the insertion or deletion of each video shot depending on user preference.

Each video shot is either included or excluded from the final video summary. In each shot, MPEG-7 metadata describes the semantic content and corresponding scores. Assume there are a total of *N* attribute categories. Let $\vec{P} = [p_1, p_2, \dots, p_N]^T$ be the user preference vector, where p_i denotes the preference weighting for attribute *i*, $1 \le i \le N$. Assume there are a total of *M* shots. Let $\vec{S} = [s_1, s_2, \dots, s_M]^T$ be the shot segments that comprise the original video sequence, where s_i denotes shot number *i*, $1 \le i \le M$. Subsequently, the attribute score $a_{i,j}$ is defined as the relevance of attribute *i* in shot *j*, $1 \le i \le M$ and $1 \le i \le N$. It then follows that the weighted attribute w_i for shot *i* given the user preference $\vec{P}$ is calculated as the dot product of the attribute matrix ***A*** and the user preference vector $\vec{P}$:

$$\vec{W} = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_M \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,N} \\ a_{2,1} & \dots & \dots & \dots \\ \dots & \dots & a_{i,j} & \dots \\ a_{M,1} & \dots & \dots & a_{M,N} \end{bmatrix} * \begin{bmatrix} p_1 \\ p_2 \\ \dots \\ p_N \end{bmatrix}$$

w_i specifies the relative weighted importance of shot *i* with respect to the other shots. Assume shot s_i spans a durations of t_i, $1 \le i \le M$. Consequently, shot s_i is included in the summary if the importance weighting w_i of this shot is greater than some threshold, $w_i > q$, and excluded otherwise. q is determined such that the sum of the shot durations t_i is less than the user specified time constraint. Consequently, each shot is initially ranked according to its weighted importance and either included or excluded in the final personalized video summary according to the time constraint. Furthermore, the *VideoSue* engine generates the optimal set of selected video shots for one video as well as across multiple video sources.

5.2 Adaptation Engine

Our *VideoEd* is a novel MPEG system layer compressed-domain editing and composition technique used to facilitate the delivery and integration of multiple segments of MPEG files, residing on remote databases [8].

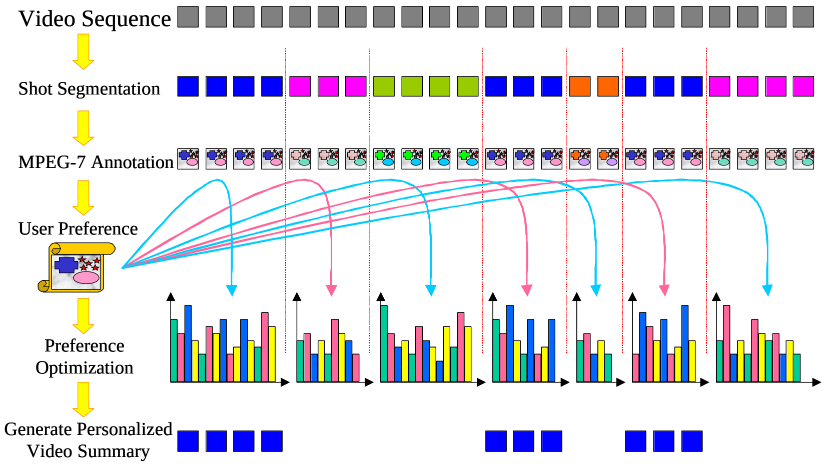


Figure 4: Overview of Video Personalization and Summarization Engine. Each video is segmented into shots that are annotated with MPEG-7 descriptions. Using user's preferences and profiles, the optimally matched set of video shots is selected.

Various multimedia applications, including retrieval and summarization, split MPEG files into small segments along shot boundaries and store them separately. This traditional method requires extra management and storage payload, provides only fixed segmentations, and may not play smoothly. In order to solve this problem, our adaptation engine directly extracts the desired video-audio information from the original MPEG sources and combines them to generate a single MPEG file. Manipulated wholly in the system bitstream domain, this method does not require decoding, re-encoding, and re-synchronization of audio and video data. Thus, it operates in real-time and provides great flexibility. This composite MPEG file can be transmitted and displayed through general Web interfaces.

The other component in our adaptation is the server end of our Universal Tuner. The Universal Tuner system includes two parts -- a software video transcoder at the server end which can transcode MPEG-1/2 video or A/D converted live broadcasting video into client dependent format, and a software client application software on the color or Black-and-White Palm OS PDA.

.5.2.1 *VideoEd* – Real-Time Video Editing and Composition Tool for Content Adaptation

Figure 5 shows the flow chart of the editing tool, *VideoEd.* This tool takes three kinds of input: MPEG files, Frame Map files and Retrieval List. The first two are stored in a database. Frame Map files, which indicate the accessible points of the pack header of GOPs, are used to indicate the randomly accessible positions in an MPEG system stream. The Retrieval List is the XML file generated by the personalization engine.

The highlighted area in Figure 5 shows the case of concatenating multiple video-audio segments from an MPEG file to generate a new one. The first step is to copy and transfer the first one or two packs to a new file. The number of copied packs depends on whether the global information of video and audio are stored separately and the number of packets in a pack. Next, we need to extract video and audio sequence headers. In an MPEG-1 bitstream, there can be more than one video sequence header (VSH). In most cases there is either one VSH for the whole video sequence or multiple VSHs before each GOP. Note that a VSH has to be placed immediately before a GOP. Audio sequence headers have the same properties as that of video.

The content of audio and video data is copied and pasted in the pack levels. According to the retrieval list, we then choose the accessible positions indicated by the Frame Map files. Note that the resolution of segmentation has to be in the GOP. In other words, video and audio can only be cut and pasted at a resolution of nearly 0.5 seconds. After this step, we need to clean up the bits that belong to the previous or next GOP. If there are multiple packets in a pack, we delete the unnecessary packets. After cleaning up, we insert the video sequence header to the beginning of the first video segments. Similar cleaning process is performed for audio packets.

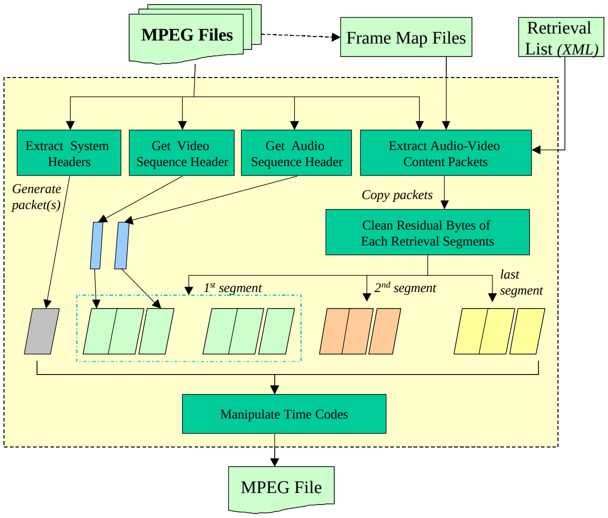


Figure 5: Basic Structure of *VideoEd* System Compressed-Domain Editing and Composition Tool.

The final step of *VideoEd* is the manipulation of time codes. They must be done in the system pack and packet layers. They new sequence is regarded valid regardless of the time code changes at the GOP headers. To concatenate two clips, we have to calculate a temporal offset from the difference of the first SCR in the new clip to the SCR of the last clip, and subtract it from the SCR of all the packs. We have to change PTS and DTS of all the packets in the second clip using a similar strategy.

5.2.2 *Universal Tuner* – Format and Scale Transcoder for Pervasive Devices

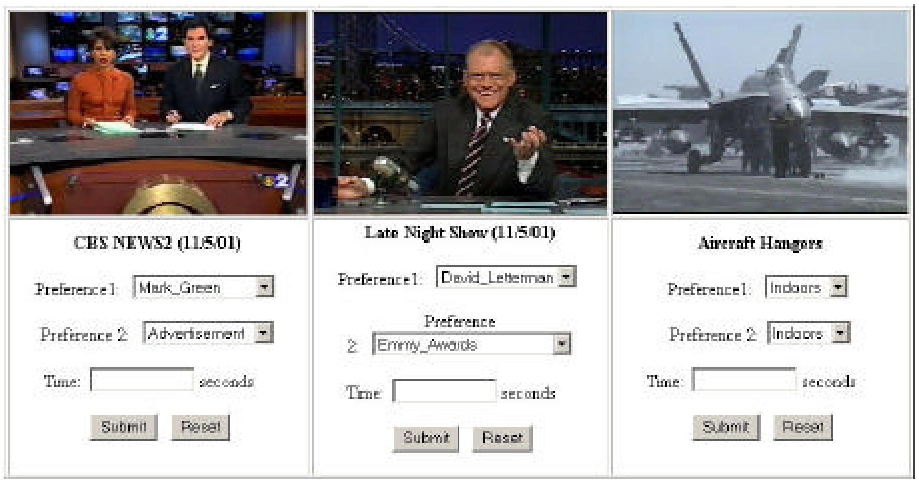
The *Universal Tuner* is a transcoder for pervasive devices [5]. In *Universal Tuner*, the complexity of multimedia compression and decompression algorithms is adaptively partitioned between the encoder and decoder. A mobile client would selectively disable or re-enable stages of the algorithm to adapt to the device's effective processing capability. Our variable-complexity strategy of selective disabling of modules supports graceful degradation of the complexity of multimedia coding and decoding into a mobile client's low-power mode, *i.e*. the clock frequency of its next-generation low power CPU has been scaled down to conserve power.

The input of the transcoder is a stream of video frames of MPEG-1/2/4 file. The video frames are first decoded and resampled to either 80x80 or 160x160 image. Then, depending on whether the PDA device is Black and White or 256 colors (as in Palm IIIc), the color RGB frames are either dithered using halftoning or mapped to 256 colors using a shot-based optimal color map. In the color mode, we added a shot boundary detection functionality in order to update the color codebook for each new shot. After this step, the transcoder can selectively to enable an entropy coding process and a frame differencing process to further reduce the bit rates.

6. PERSONALIZATION APPLICATIONS AND USER CLIENTS

The user client allows the user to retrieve specific content via a user query, provide critical persistent data through the usage environment, and receive the personalized content on the user's display client. The usage environment holds the profiles about the user, device, network, delivery, and other environments. This usage environment descriptions are sent over to the personalization engine of the media middleware, and may include some user profile [i.e., English speaking American, user residing in New York City], device profile [i.e., Palm IIIc with limited color display, no audio capabilities, and memory capacity], and transmission profile [i.e., internet access through 56K modem.] They allow the middleware to dynamically perform the appropriate content adaptation on the requested media. Thus the delivered multimedia can be personalized for the usage environment.

The user submits a request for some content via a user query. The query can take the form of topic preferences, search keywords, media samples, and/or time constraints. The user query then initiates the retrieval for customized



CBS NEWS2 (11/5/01)

Preference 1: Mark_Green

Preference 2: Advertisement

Time: seconds

Late Night Show (11/5/01)

Preference1: David_Letterman

Preference 2: Emmy_Awards

Time seconds

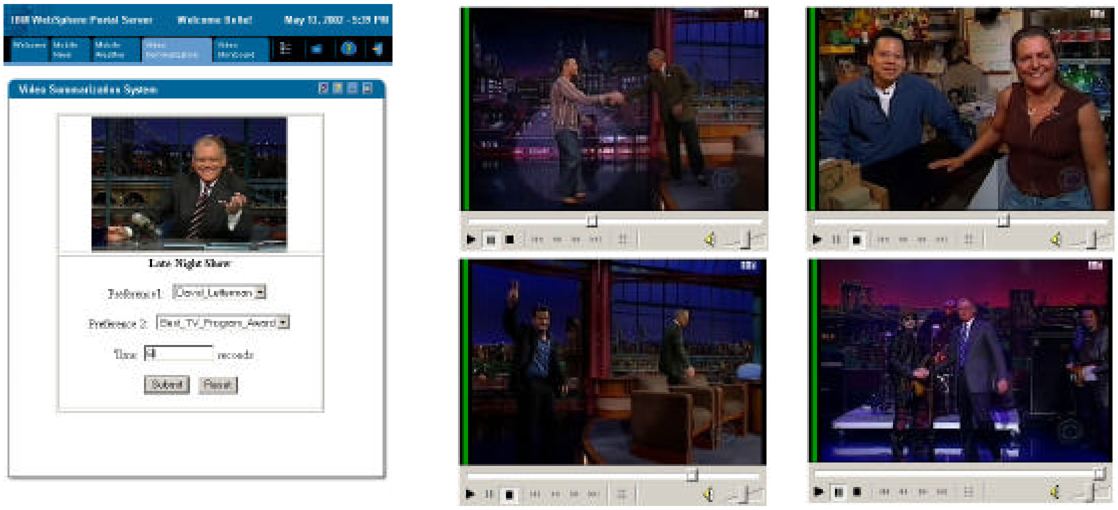
Aircraft Hangers

Preference1: Indoors

Preference 2: Indoors

Time: seconds

Figure 6: User Client Portals corresponding to Three Different Usage Environment Profiles for: (1) News [*left*], (2) Entertainment [*middle*], and (3) Education [*right*]



IBM WebSphere Portal Server Welcome Bella! May 13, 2002 - 5:39 PM

Video Summarization System

Late Night Show

Preference 1: David_Letterman

Preference 2: Best_TV_Program_Award

Time: [illegible] seconds

Submit Reset

Figure 7: The usage example and screen shots of the personalized video summarization for user clients on PC.

content by the media middleware. Following, the personalized multimedia is delivered from the adaptation engine of the middleware to the user and rendered on the display client in accordance with the usage environment and the user query. In the rest of this section, we present two personalization applications.

6.1 User Clients on PCs using IBM Websphere Portal Server

The IBM Websphere Portal Server (WPS) product provides a secure, single point of access to diverse information and applications, personalized to the needs of their users [6]. The WPS uses portals to present users with personalized content according to authenticated user profiles. WPS can also be deployed on many categories of portals that allow access from a wide variety of desktops and mobile devices. This framework provides us with an open, flexible, and scalable infrastructure for creating and deploying our rich media personalization system with usage environment descriptions. Within WPS, we have developed the personalization and summarization portal.

When a user establishes an account with the WPS, the user profiles are created and can be modified later. In our portal, the preference interests of the user are gathered. This includes the topic category and default viewing time. Figure 6 illustrates the user client portals for three different topic categories of *News*, *Entertainment*, and *Education*. Thus the appropriate portal page according to the user profile would automatically be presented when the user logs on to its portal. This illustrates the use of persistent usage environment within WPS. Assuming our user has stored the desired topic category to be *Entertainment*. Then Figure 7 (left) shows the corresponding portal interface of the *Late Night Show*, which is ranked in the Entertainment category.

After our user enters the portal page, she can request for a personalized video summary according to her current interests. There are two preferences and the time limit for her selection. For the *Late Night Show*, the preference choices range from “David_Letterman”, “Paul_Shaffer”, to “Monologue”, “Guest_Interview”, and “Top_10.” Our user submits the user query preferences (1) “Guest_Introduction” and (2) “David_Letterman”, along with a time constraint of 120 seconds. As a result, the *VideoSue* summarization and personalization engine retrieves the optimal set of highly relevant video segments that included David Letterman introducing many guests from three video sources. Subsequently, the *VideoEd* composition tool gathers this set of selected video clips and dynamically generates one personalized video movie. Figure 7 (right) depicts four screen shots of guests as they are introduced.

6.2 User Clients on Personal Digital Assistance (PDA) Devices

Using the Universal Tuner adaptation engine, we can transcode a summarized video and transmit it to pervasive devices. The client can access personalized video with one of these three modes: (a) Video-on-demand with interactive hyperlinks, (b) Summarized video based on preference topics and time constraint, and (c) Summarized video based on query keywords and time constraint. Figure 8 illustrates the user client in mode (b) on the left and mode (c) on the right. A user can first click on the Get Channels button to get a list of hyper-links according to the stored user profile. This list is then displayed on the left of the video playback window. He can then select to view the video by clicking on the hyperlinks. In another scenario, he may then want to see a summarized video based on the preference choices that were sent from the middleware when he selects a video. Then, based on his preferences and time constraint, he will receive a summarized video, which best matches the user preferences and profiles. In addition to the preferences that were recommended by the middleware, users could also key in some keywords and let the middleware find the best matches and generate the personalized video summary.

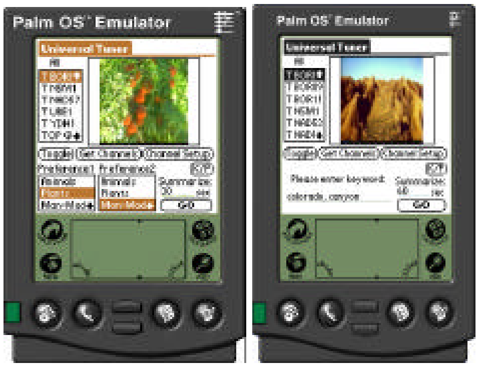


Figure 8: The User Client Interface on the Palm PDA.

7. CONCLUSION

A video personalization and summarization system is presented matching media descriptions with usage environments in a server-middleware-client framework in order to deliver personalized media contents to users. The three-tier architecture provides a standards-compliant infrastructure using our tools and engines to select, adapt, and deliver personalized video summaries to the users effectively. The *VideoAnnEx* MPEG-7 Video Annotation Tool allows annotators to describe video segments using semantic concepts. The *VideoSue* Summarization on Usage Environment Engine determines the optimal selection of media contents according to user preferences and time constraint. The *VideoEd* Editing and Composition Tool and *Universal Tuner* Tool gathers the selected contents and generates one personalized video for the user. Our personalization and summarization systems are demonstrated on top of the IBM Websphere Portal Server (WPS) and for the PDA devices.

8. REFERENCES

[1] K. Aizawa, K. I. Shijima, and M. Shiina, "Summarizing Wearable Video," IEEE ICIP, Thessaloniki, Greece, Oct. 2001.

[2] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, "Using Audio Time Scale Modification for Video Browsing", Hawaii Int. Conf. on System Sciences, HICSS-33, Maui, January 2000.

[3] Mark H. Butler, "Implementing Content Negotiation using CC/PP and WAP UAProf," Technical Report HPL-2001-190, June 2001. [http://www.hpl.hp.com/techreports/2001/HTTPCCPP_2001-190.html]

[4] Y. Gong and X. Liu, "Summarizing Video by Minimizing Visual Content Redundancies," IEEE ICME, Tokyo, Aug. 2001.

[5] R. Han, C.-Y. Lin, J. R. Smith, B. L. Tseng and V. Ha, "Universal Tuner: A Video Streaming System for CPU/Power-Constrained Mobile Devices", ACM Multimedia 2001, Ottawa, Canada, Oct. 2001.

[6] http://www.ibm.com

[7] ISO/IEC JTC 1/SC 29/WG 11/N 4242, Text of 15938-5 FDIS, Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes, Final Document International Standard (FDIS) edition, October 2001.

[8] C.-Y. Lin, B. L. Tseng and J. R. Smith, "Universal MPEG Content Access Using Compressed-Domain System Stream Editing Techniques," IEEE Int'l Conf. on Multimedia and Expos, Switzerland, August 2002.

[9] B. Merialdo, K. T. Lee, D. Luparello and J. Roudaire, "Automatic Construction of Personalized TV News Programs," ACM Multimedia 1999, Orlando, FL, Sept. 1999.

[10] MPEG-7 Tools for MPEG-21 Digital Item Adaptation, ISO/IEC JTC1/SC29/WG11/M8321, Fairfax, VA, May 2002.

[11] MPEG-21 Requirements on Digital Item Adaptation, ISO/IEC JTC1/SC29/WG11/N4684, Jeju, Korea, March 2002.

[12] M. Naphade, C.-Y. Lin, J. R. Smith, B. L. Tseng, and S. Basu, "Learning to Annotate Video Databases", Proc of SPIE on Storage, Retrieval for Media Database, Vol. 4676, San Jose, Jan 2002.

[13] http://www.nttdocomo.co.jp

[14] B. L. Tseng, C.-Y. Lin, and J. R. Smith, "Video Summarization and Personalization for Pervasive Mobile Devices," SPIE Electronic Imaging 2002 - Storage and Retrieval for Media Databases, San Jose, January 2002.

[15] http://www.virage.com

[16] M. Yeung, B. Yeo, W. Wolf and B. Liu, "Video browsing using clustering and scene transitions on compressed sequences," Proc. SPIE on Multimedia Computing and Networking, Vol. 2417, 1995.