PROTOCOLS AND METHODOLOGIES FOR THE UTILIZATION OF MPEG-7 IN MULTIMEDIA DATA STORAGE AND RETRIEVAL

by
Gary Tinker

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science

Chapel Hill North Carolina

December, 2002

Approved by

_____

Advisor

Gary Tinker. Protocols and Methodologies for the Utilization of MPEG-7 in Multimedia Data Storage and Retrieval. A Master's paper for the M.S. in I.S. degree. December, 2002. 44 pages. Advisor: Gregory B. Newby

This paper endeavors to provide a description of the MPEG-7 standards and their applicability to multimedia data storage and retrieval. The concept of multimedia feature extraction is explored as a foundation for the automated collection of data that may provide adequate descriptors for multimedia source material. A demonstration software tool, the IBM MPEG-7 Annotation Tool, is evaluated to determine the viability of automated video keyframe extraction.

Headings:

MPEG-7

Multimedia Metadata

Feature Extraction

Table of Contents

Introduction

Multimedia source material may be generally defined as any combination of audio, visual, and textual media content that conveys a message. These types of source material include audio recordings, motion picture film, still photographs, video tape recordings, and all of the graphic arts. In many instances, collections of normally static media content such as photographs, text, graphic art, and artifacts have been gathered together and combined with music, live action, and narrative voice-overs as a multimedia production piece. The Ken Burns documentary "Jazz" is a prime example of multimedia production. The great quantity of these types of materials currently available in archives maintained by libraries, museums, and commercial collections of all types presents the archivist with the incredible challenge to categorize, index, store and retrieve these materials. The extremely complex nature of multimedia content has led to many different approaches to achieve this goal.

Archivists involved with storage and retrieval of multimedia source materials require the means to substantially improve the accuracy, efficiency, and comprehensive content indexing of those materials. Researchers in a wide range of disciplines are developing methodologies for content-based media analysis to provide the archivist with the automation tools necessary to significantly assist in the process. These efforts are aptly summarized by Chang [8] in the following.

> Tools and systems for content-based access to multimedia and – image, video, audio, graphics, text, and any number of combinations – has increased in the last decade. We've seen a common theme of developing automatic analysis techniques for deriving metadata (data describing information in the content at both syntactic and semantic levels). Such metadata facilitates developing innovative tools and systems for multimedia information retrieval, summarization, delivery, and manipulation.

Innovative content-based analysis tools are indeed required to assist the archivist in processing the vast amounts of metadata inherent in multimedia source material.

This paper will focus on the standardization of the protocols by which multimedia content, which has been converted into digital data, may be described, categorized and indexed through the application of and adherence to the Moving Picture Experts Group (MPEG) standards embodied in the ISO/IEC International Standard 15938, Multimedia Content Description Interface, better known as MPEG-7. A basic description of MPEG-7 will be provided as a means of demonstrating the applicability of the standard as a metadata collection and delivery mechanism. The concept of feature extraction is introduced to illustrate the scope and variety of metadata that may be derived via automation tools. A currently available automation tool that provides automatic feature extraction of MPEG video streams and automatic creation of an MPEG-7 data file is described and presented for evaluation.

The evaluation of the automation tool is undertaken in an effort to reveal the viability of applying automated processes to the indexing of video data. A video test collection which had undergone independent manual indexing was chosen to serve as a baseline for comparative analysis. Test results from the individual videos comprising the

collection are presented in a narrative fashion followed by an appraisal of the overall test results.

MPEG-7

The following is a brief description of MPEG and MPEG standards is taken from MPEG-7 Working Papers. (1)

1.1 What Are the MPEG Standards?

The Moving Picture Coding Experts Group (MPEG) is a working group of the Geneva-based ISO/IEC standards organization, (International Standards Organization/International Electro-technical Committee) in charge of the development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio, and a combination of the two. MPEG-7 then is an ISO/IEC standard being developed by MPEG, the committee that also developed the Emmy Award-winning standards known as MPEG-1 and MPEG-2, and the 1999 MPEG-4 standard.

- ***MPEG-1:*** For the storage and retrieval of moving pictures and audio on storage media.
- ***MPEG-2:*** For digital television, it's the timely response for the satellite broadcasting and cable television industries in their transition from analog to digital formats.
- ***MPEG-4:*** Codes content as objects and enables those objects to be manipulated individually or collectively on an audiovisual scene.

MPEG-1, -2, and -4 make content available. MPEG-7 lets you to find the content you need.
Besides these standards, MPEG is currently also working in MPEG-21 a Technical Report about Multimedia Framework.

1.2 Defining MPEG-7

MPEG-7 is a standard for describing features of multimedia content.

1.2.1 Qualifying MPEG-7

MPEG-7 provides the world's richest set of audio-visual descriptions.

These descriptions are based on catalogue (e.g., title, creator, rights), semantic (e.g., the who, what, when, where information about objects and events) and structural (e.g., the colour histogram - measurement of the amount of colour associated with an image or the timbre of an recorded instrument) features of the AV content and leverages on AV data representation defined by MPEG-1, 2 and 4.

Comprehensive Scope of Data Interoperability.

> MPEG-7 uses XML Schema as the language of choice for content description MPEG-7 will be interoperable with other leading standards such as, SMPTE Metadata Dictionary, Dublin Core, EBU P/Meta, and TV Anytime.

It should be clear from the "Comprehensive Scope of Data Interoperability" statement above that MPEG-7 intends to be an inclusive standard embracing many of the existing multimedia description standards currently in use. The authors of the MPEG-7 standards have also clearly defined the role MPEG-7 is designed to play in multimedia content storage and retrieval applications. [1]

1.3 The Key Role of MPEG-7

> MPEG-7, formally named "Multimedia Content Description Inter-face," is the standard that describes multimedia content so users can search, browse, and retrieve that content more efficiently and effectively than they could using today's mainly text-based search engines. It's a standard for describing the features of multimedia content.

MPEG-7 is positioned to be application neutral and is primarily dependent on the Extensible Markup Language (XML) to achieve this goal. From an application perspective, MPEG-7 descriptors are the message and XML is the messenger. [1]

> MPEG-7 will define a multimedia library of methods and tools. It will standardize:
> • ***A set of descriptors:*** A descriptor (D) is a representation of a feature that defines the syntax and semantics of the feature representation.
> • ***A set of description schemes:*** A description scheme (DS) specifies the structure and semantics of the relationships between its components, which may be both descriptors and description schemes.
> • ***A language that specifies description schemes, the Description Definition Language (DDL):*** It also allows for the extension and modification of existing description schemes. MPEG-7 adopted XML

Schema Language as the MPEG-7 DDL. However, the DDL requires some specific extensions to XML Schema Language to satisfy all the requirements of MPEG-7. These extensions are currently being discussed through liaison activities between MPEG and W3C, the group standardizing XML.

• ***One or more ways (textual, binary) to encode descriptions:*** A coded description is a description that's been encoded to fulfill relevant requirements such as compression efficiency, error resilience, and random access.

Creation of the MPEG-7 standards would at first appear be a daunting task. Fortunately, the MPEG-7 authors have been able to leverage a great deal of the multimedia description work previously done by various standards groups. The SMPTE Metadata Dictionary [2] is comprised of three hundred and fifty-three data element categories with six hundred and seventy individual data element attributes. The Dublin Core Metadata Element Set is fifteen elements with ten attributes per element [3]. The European Broadcasting Union (EBU) P/Meta project extends the SMPTE Metadata Dictionary for compatibility in the European marketplace [4]. Elements of consumer oriented multimedia metadata standards such as TV Anytime and Material eXchange Format (MXF) are also included for compatibility in the consumer market. The primary source of information concerning MPEG-7 is the MPEG Home Page at http://mpeg.telecomitalialab.com/ with the World Wide Web Consortium (W3C) as the primary source for issues concerning XLM.

Feature Extraction

The adoption of MPEG-7 as an international standard for multimedia content description established a flexible framework for the development of applications which

capture metadata for the indexing, storage, and retrieval of multimedia source material. The core of multimedia indexing lies in feature extraction. Features may be segregated into two basic categories, a high and low level depending on complexity and the use of semantics. According to Djeraba [13]:

> Low-level features (also known as primitive features) such as object motion (for video), color, texture, shape, spatial location of image elements (for both images and video), special events, and pitch (for audio) permit queries such as "find clips of objects moving from the bottom left to the bottom right of the frame," which might retrieve video pieces of objects (for example, a ball) following that specific trajectory. . .This level uses features that are objective and directly derivable from the images themselves, but it doesn't refer to any external knowledge base.
>
> High-level features (also known as logical, derived, semantic features) involve various degrees of semantics depicted in images, video, and audio . . . Complex interpretation and subjective judgment can be required by an application domain expert to make the relationship between image content and abstract concepts.

The implication is that automated processes are suitable for low level feature extraction while human intervention is necessary to make semantic sense of the source material. Automated low-level feature extraction may be able to identify a flower arrangement while being unable to distinguish between a wedding bouquet and a funeral wreath or identify a four wheeled vehicle while being unable to distinguish a passenger car from a dump truck. While it may eventually be possible to overcome these limitations to a degree, human intervention may always be necessary for making the crucial semantic judgments that provide context and relationship information for the source material being examined. Manual indexing of multimedia source material is extremely time consuming. Automated tools which assist in the process should reduce the time required for this task

which in turn should encourage a consistent, richer, and more efficient style of multimedia indexing.

The ideal automation tool to assist in metadata creation for multimedia source material would process all low-level visual content, define shot boundaries, extract video keyframes, analyze audio content, perform speech to text conversion, extract keywords, perform optical character recognition on textual video frames, and present a domain expert with a summary of these details for annotation. While a single tool to accomplish those tasks is not yet available, great strides have been made in those individual areas of research. The software subject for evaluation in this regard is the IBM MPEG-7 Annotation Tool v1.4.

MPEG-7 Annotation Tool

The IBM MPEG-7 Annotation Tool v1.4 is available for download from http://www.alphaworks.ibm.com/tech/videoannex for evaluation and may be licensed for commercial use. The only other comparable software that could be located is the Richo MPEG-7 Movie Tool at http://www.ricoh.co.jp/src/multimedia/MovieTool/. However, all attempts to download that tool were unsuccessful.

The description provided for the IBM MPEG-7 Annotation Tool from the download site is as follows:

> The IBM MPEG-7 Annotation Tool assists in annotating video sequences with MPEG-7 metadata. Each shot in the video sequence can be annotated with static scene descriptions, key object descriptions, event descriptions, and other lexicon sets. The annotated descriptions are associated with each video shot and are put out and stored as MPEG-7 descriptions in an XML

file. IBM MPEG-7 Annotation Tool can also open MPEG-7 files in order to display the annotations for the corresponding video sequence. IBM MPEG-7 Annotation Tool also allows customized lexicons to be created, saved, downloaded, and updated.

In practice, the annotation tool pre-processes a MPEG compressed video file to determine shot boundaries, extract all frames in each shot, and selects a keyframe representation for each shot. The tool user may not override the predetermined shot boundaries or select an alternate keyframe. The annotation tool user interface is shown in Figure 1. The user

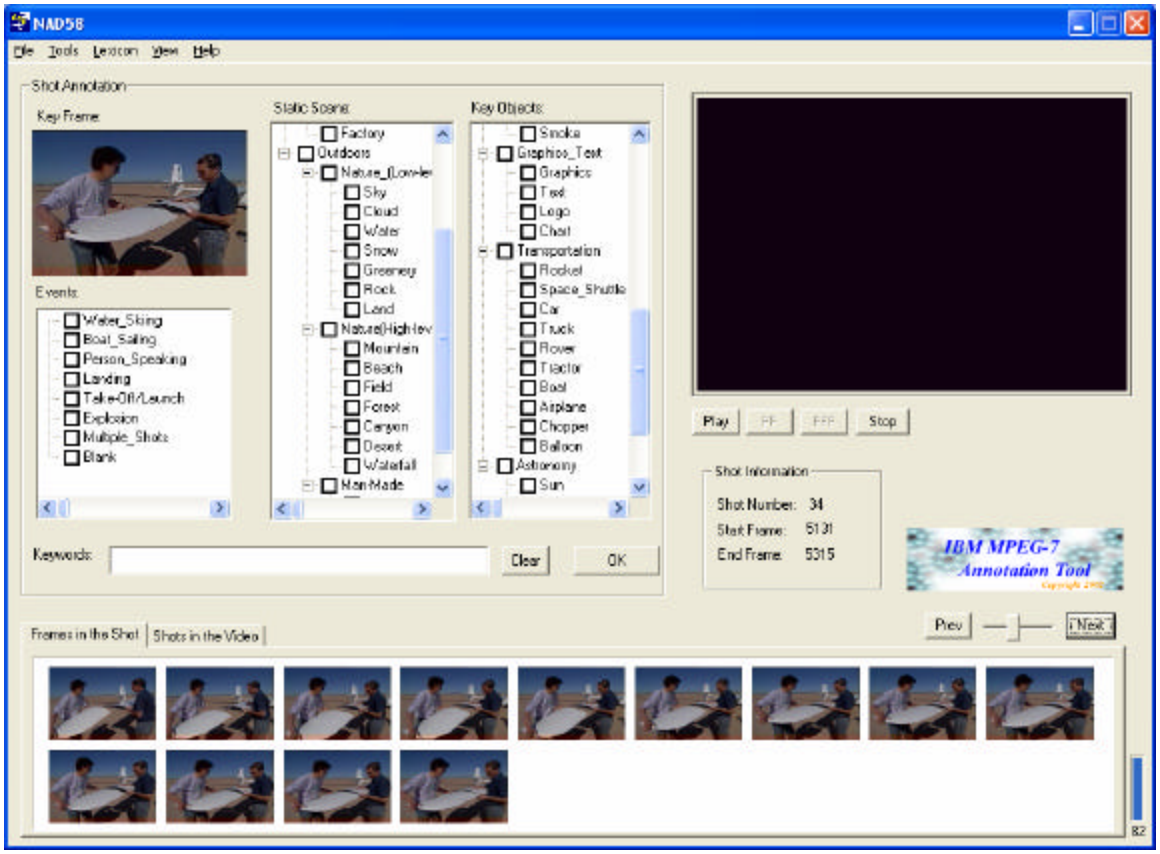

Figure 1 **IBM MPEG-7 Annotation Tool interface**

must first open a MPEG file with the annotation tool. The annotation tool processes the input file as described above the first time the file is opened. This may be very time consuming process depending on the length of the video source material as well as the computer hardware running the software. I was able to determine that software performance during the initial pre-process phase is primarily determined by the speed of the Central Processor Unit (CPU) by monitoring the utilization of system resources. During this phase, CPU utilization was tracked at one hundred percent usage.

The annotation tool interface is “shot” oriented. The interface provides the user with shot annotation section that includes a view of the determined shot keyframe, a window for playing the shot, a shot information section, shot navigation controls, and a tabbed section for viewing all the frames in the current shot or all shot keyframes in the video (you may scroll through the shot keyframes by using the shot navigation controls). Since the shot boundaries and keyframes are determined by the software, the only user control functions are in the shot annotation section.

Shot annotation is facilitated by the use of a default lexicon supplied with the software that may be edited to suit the user’s needs and saved for future use. The lexicon populates the shot annotation section with predefined “Events”, “Static Scene”, and “Key Objects” frames for quickly annotating shots based on those descriptors. The user may also enter freeform data in the “Keywords” frame. Additionally, the may activate “Annotation Learning” and “Region Annotation” from the Tools menu.

Annotation Learning appears to mainly consist of automatically repeating the previous shot keywords presumably to aid in the entry of annotation keywords that the user may require for a series of shots. This is useful when a series of shots comprises a

"scene" with common keyword descriptors. However, the concept of a scene related to a common keyframe is not supported by the annotation tool. The ability to scroll through all the shots in the video aids in determining if a series of shots exist that would benefit from this treatment. The necessity for having foreknowledge of the shot sequences and the additional time required to obtain such knowledge is a definite hindrance in utilizing this feature.

The activation of Region Annotation gives the user the opportunity to select a region of the current keyframe to be associated with the keywords. When Region Annotation is active, clicking the OK button (which saves the annotations in memory) brings up a popup window where the user may use the mouse to select the appropriate keyframe region. This may be useful for effectively eliminating background video data that is not germane to the annotation text. The feature must be activated prior to saving the annotation data and stays active until toggled off.

There are several "save" options available to the user:

Save MPEG-7 XML – saves all the annotation data along with all the temporal data related to each shot to an xml file for future use. Since this is the only place where the annotation data is saved this is highly recommended before ending the annotation session. The saved xml file may be recalled and updated in subsequent annotation sessions.

Save All Keyframes – keyframes are saved as jpeg static images which may be used as thumbnails for a quick video summary or for use as image data in a database.

Save Shot Frames – saving all the frames of any selected shot as static jpeg images.

Save Shot I-Frames – I-Frames appear to be sample frames taken at a fixed interval for each shot perhaps facilitating the creation a "skim" video.

Additionally, the software saves a file with an frp extension which maintains the video shot boundary parameters and is automatically referenced for subsequent annotation sessions. This relieves the necessity of repeating the time consuming shot boundary analysis each time a file is open by the annotation tool. The user may choose to load a previously saved MPEG-7 XML file containing previous annotation data or may create new annotations and save those as well.

Testing the IBM MPEG-7 Annotation Tool

The automatic feature extraction capabilities of the IBM MPEG-7 Annotation Tool are the subject of the testing regime. The capacity to automatically determine shot boundaries and select keyframes for each shot is the primary function of the software. Other software functions such as the recording of shot annotations and the automatic generation of an XML file containing shot boundaries and manually coded shot annotation data rely on the software's ability to determine and manage shot boundaries and keyframes. The "2001 TREC Video Retrieval Test Collection", available for download from the Open Video Project website at http://www.open-video.org/, was chosen to serve as the test collection. Not only was the TREC collection readily available

for use as a test collection but the collection had been manually analyzed and appropriate keyframes manually selected. Manual analysis did not include shot boundary description but the selection of keyframes is expected to provide an indication of the temporal regions of each video that the manual analysis deemed appropriate for description via a keyframe.

Specifications for the MPEG videos that comprise the TREC collection are culled from the Open Video Project website and are given in Table 1. These videos range induration from six minutes and nineteen seconds for the "NASA 25th Anniversary Show, Segment 05" to forty-eight minutes and thirty seconds for "Senses and Sensitivity, Lecture 4". As would be expected, the shortest duration video has the smallest file size (66.99 MB) and the longest duration video has the largest file size (475.00 MB). While variations in video run time and file size are not necessarily indicative of the number of shots / keyframes that will result from any analysis, having this type of variety in the test may produce some notable results. The primary test methodology consists of comparing the shots / keyframes generated by the annotation tool with keyframes generated by manual analysis noting similarities, differences, and significant trends that may become evident.

Processing the Test Collection

Results from processing each video with the IBM MPEG-7 Annotation Tool will be individually evaluated and the accumulated results will be inspected for trend analysis. Please note that the figures provided in Appendix B for the keyframes produced by the Annotation Tool are limited to thirty frames and are provided as a sample of the program

#	Video Title	Duration	MPEG-1
1	A New Horizon	00:16:44	146.00 MB
2	Challenge at Glen Canyon	00:26:57	235.00 MB
3	Giant on the Bighorn	00:14:03	122.00 MB
4	How Water Won the West	00:11:17	98.40 MB
5	Lake Powell	00:27:42	241.00 MB
6	NASA 25th Anniversary Show, Segment 05	00:06:19	66.99 MB
7	NASA 25th Anniversary Show, Segment 06	00:09:13	97.66 MB
8	NASA 25th Anniversary Show, Segment 09	00:06:50	72.57 MB
9	NASA 25th Anniversary Show, Segment 10	00:17:27	184.83 MB
10	Report #259	00:14:20	127.00 MB
11	Report #260	00:14:31	125.00 MB
12	Report #262	00:07:06	128.00 MB
13	Report #264	00:07:06	65.00 MB
14	Report #265	00:07:42	67.20 MB
15	Senses And Sensitivity, Lecture 3	00:25:30	473.00 MB
16	Senses And Sensitivity, Lecture 4	00:48:30	475.00 MB
17	Space Works 3	00:29:26	257.00 MB
18	Space Works 5	00:29:49	260.00 MB
19	Space Works 6	00:29:09	254.00 MB
20	Space Works 7a	00:29:03	253.00 MB
21	Space Works 8	00:27:41	241.00 MB
22	Take Pride in America	00:11:32	101.00 Mb
23	The Colorado	00:19:59	174.00 MB
24	The Great Web of Water	00:28:07	245.00 MB
25	The Story of Hoover Dam	00:27:35	240.00 MB
26	Wetlands Regained	00:14:10	124.00 MB

Table 1 The 2001 TREC Video Retrieval Test Collection

output. The thirty frame limit was established in order to fit both figures relating to a video onto a single page. All generated keyframes were inspected to produce the following commentaries.

Video #1 **"A New Horizon"** – the downloaded video is unusable due to significant video as well as audio noise. The Annotation Tool appeared to process the video normally. The resultant keyframes were viewable in the Shots preview window but the program terminated prematurely upon attempting to save the keyframes and no data was retained for analysis. Of the one hundred and seventy keyframes produced, the first six were heavily distorted and the remaining one hundred and sixty-six were identical copies. The Open Video website contained no Segmentation Frames (keyframes) for this video.

Video #2 **"Challenge at Glen Canyon"** – the downloaded video was successfully processed by the Annotation Tool (AT). The Open Video (OV) website displays twenty-seven keyframes to represent this video. The annotation tool created two hundred and thirty-two keyframes. Appendix B, Figure 1a illustrates the OV keyframes and Figure 1b displays the first thirty AT keyframes. All OV keyframes have similar or identical counterparts in the AT keyframe set.

Video #3 "Giant on the Bighorn" – the downloaded video was successfully processed by the AT. The OV website displays twenty-seven keyframes to represent this video. The AT created one hundred and forty-three keyframes. Appendix B, Figure 2a illustrates the OV keyframes and Figure 2b displays the first thirty AT keyframes. Nearly all OV keyframes have similar or

identical counterparts in the AT keyframe set. The exceptions being part of an identifiable frame sequence.

Video #4 "How Water Won the West" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately ninety percent of the video. A scan of the video revealed significant video noise beginning at approximately the ten minute mark of this eleven minute and seventeen second video. Presumably the video noise was sufficient to cause the process termination.

Video #5 "Lake Powell" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately fifty percent of the video. Video was not viewable with any available player. The OV website does not show any keyframes for this video.

Video #6 "NASA 25th Anniversary Show, Segment 05" – the downloaded video was successfully processed by the AT. The OV website displays forty keyframes to represent this video. The AT created forty-one keyframes. Appendix B, Figure 3a illustrates the OV keyframes and Figure 3b displays the first thirty AT keyframes. A total of twenty-one keyframes are common to both sets of keyframes. This short, six minutes and nineteen seconds, video is composed of several scenes that quickly change point of view or distance aspect. Both the OV and AT display several similar keyframes for

certain, though different, scenes. The AT generated keyframes have a tendency to show this characteristic but it is unexpected for manually selected keyframes.

Video #7 "NASA 25th Anniversary Show, Segment 06" – the downloaded video was successfully processed by the AT. The OV website displays nineteen keyframes to represent this video. The AT created fifty-nine keyframes. Appendix B, Figure 4a illustrates the OV keyframes and Figure 4b displays the first thirty AT keyframes. Seventeen of the OV keyframes are duplicated, or nearly so, in the AT keyframe set. The two OV frames not included in the AT keyframe set are significant, unique still shots in an area of the video with almost no noticeable video noise.

Video #8: "NASA 25th Anniversary Show, Segment 09" – the downloaded video was successfully processed by the AT. The OV website displays thirty-seven keyframes to represent this video. The AT created sixty keyframes. Appendix B, Figure 5a illustrates the OV keyframes and Figure 5b displays the first thirty AT keyframes. All but six of the OV keyframes are represented in the AT keyframe set. These are keyframes for fairly well defined, static shots. Video quality is good throughout.

Video #9: "NASA 25th Anniversary Show, Segment 10" – the downloaded video was successfully processed by the AT. The OV website displays thirty-three

keyframes to represent this video. The AT created one hundred and thirty-two keyframes. Appendix B, Figure 6a illustrates the OV keyframes and Figure 6b displays the first thirty AT keyframes. All but one of the OV keyframes are represented in the AT keyframe set, a well defined, static shot. Video quality is good throughout.

Video #10: "Report #259" – the downloaded video was successfully processed by the AT. The OV website displays ten keyframes to represent this video. The AT created one hundred keyframes. Appendix B, Figure 7a illustrates the OV keyframes and Figure 7b displays the first thirty AT keyframes. All OV keyframes are represented in the AT keyframe set. The selected OV keyframes are inadequate to properly represent this video. The video is in fact three short videos covering three distinct topics. Each video section is prefaced by a title shot. The OV keyframe set only displays one title frames and omits significant shot frames from all four sections.

Video #11: "Report #260" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately ten percent of the video. A scan of the video revealed significant video as well as audio noise beginning at approximately the two minute mark of this fourteen minute and thirty-one second video. Presumably the video noise was sufficient to cause the process termination.

Video#12: "Report #262 " – the downloaded video was successfully processed by the AT. The OV website displays nine keyframes to represent this video. The AT created one hundred and seventy-five keyframes. Appendix B, Figure 8a illustrates the OV keyframes and Figure 8b displays the first thirty AT keyframes. All OV keyframes are represented on the AT keyframe set. The selected OV keyframes are inadequate to properly represent this video. The video is in fact four short videos covering four distinct topics. Each video section is prefaced by a title shot. The OV keyframe set only displays two title frames and omits significant shot frames from all four sections.

Video #13: "Report #264" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately fifteen percent of the video. A scan of the video revealed significant video as well as audio noise beginning at approximately the one minute and thirty second mark of this seven minute and six second video. Presumably the video noise was sufficient to cause the process termination.

Video #14: "Report #265" – the downloaded video was successfully processed by the AT. The OV website displays forty-one keyframes to represent this video. The AT created two hundred and fifty-two keyframes. Appendix B, Figure 9a illustrates the OV keyframes and Figure 9b displays the first thirty AT keyframes. All OV keyframes have similar or identical counterparts in the AT keyframe set.

Video #15: "Senses And Sensitivity, Lecture 3" – the downloaded video was successfully processed by the AT. The OV website displays twenty-four keyframes to represent this video. The AT created two hundred and twenty-five keyframes. Appendix B, Figure 10a illustrates the OV keyframes and Figure 10b displays the first thirty AT keyframes. All OV keyframes have similar or identical counterparts in the AT keyframe set.

Video #16: **"**Senses And Sensitivity, Lecture 4" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately thirty percent of the video. A scan of the video revealed significant no detectable video or audio noise in the suspect region of the video. No indication of why the process failed to complete.

Video #17: "Space Works 3" – the AT was unable to process the downloaded video. The software simply halted after processing approximately twenty percent of the video. A scan of the video revealed significant video as well as audio noise throughout the video.

Video #18: **"Space Works 5"** – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately twenty percent of the video. A scan of the video revealed significant video as well as audio noise beginning at approximately the seven minute and thirty second mark.

Video #19: "Space Works 6" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately sixty percent of the video. A scan of the video revealed significant video as well as audio noise beginning at approximately the sixteen minute and thirty second mark.

Video #20: "Space Works 7a" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately eighty percent of the video. A scan of the video revealed significant video noise beginning at approximately the twenty-two minute and fifteen second mark.

Video #21: "Space Works 8" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately twenty-five percent of the video. A scan of the video revealed significant video noise beginning at approximately the four minute and thirty second mark.

Video #22: "Take Pride in America" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately eighty percent of the video. A scan of the video revealed significant video noise beginning at approximately the ten minute mark.

Video #23: "The Colorado" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately seventy-five percent of the video. A scan of the video revealed significant video noise beginning at approximately the fifteen minute mark.

Video #24: "The Great Web of Water" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately twenty-five percent of the video. A scan of the video revealed significant video noise beginning at approximately the six minute mark.

Video #25: "The Story of Hoover Dam" – the AT was unable to process the downloaded video. The software terminated abnormally and abruptly after processing approximately eighty-five percent of the video. A scan of the video revealed significant video noise beginning at approximately the twenty-two minute mark.

Video #26: "Wetlands Regained" – the downloaded video was successfully processed by the AT. The OV website contained no keyframes for this video. The AT created one hundred and seventeen keyframes. The video exhibited substantial video noise throughout but was not severe enough to cause the AT to terminate abnormally. However, the video was not viewable with any player.

Table 2, shown below, presents a generalized summary of the results of applying the Annotation Tool to the video test collection. Only results from a successful application of the Annotation Tool are displayed. The most notable feature of the summary results is the lack of consistent metrics relative to duration and OV keyframes.

#	Duration	OV Keyframes	AT Keyframes	OV Keyframes Included in AT Keyframe Set	OV Keyframes Not Included in AT Keyframe Set
2	00:26:57	27	232	27	0
3	00:14:03	27	143	27	0
6	00:06:19	40	41	21	19
7	00:09:13	19	59	17	2
8	00:06:50	37	60	31	6
9	00:17:27	33	132	32	1
10	00:14:20	10	100	10	0
12	00:07:06	9	127	9	0
14	00:07:42	41	252	41	0
15	00:25:30	24	225	24	0
	Totals	267	3362	239	28

Table 2 Summary of Annotation Tool Test Results

Analyzing the Results

The Annotation Tool was used to process a total of twenty-six videos. The most obvious collective result is the failure to process sixteen of the videos due to poor video quality. However, the successful processing of the remaining ten videos shows a great deal of promise for this type of automated multimedia feature extraction technology. A

strictly numerical analysis of the successful results does not appear to bestow any viable statistical significance. This is particularly apparent given the subjective nature of manual keyframe extraction and the widely variable nature of the source material.

A comparison between the manually selected keyframes and the automatically selected keyframes provokes the most interest. Two hundred and sixty-seven keyframes were manually extracted from the ten successfully processed videos. Twenty-nine of those keyframes were not included in the automatically extracted keyframe sets. Of those twenty-nine keyframes, twenty were from video number six.

The Annotation Tool does a remarkably good job of determining shot boundaries. There are occasions when shot boundaries unnecessarily segment a seemingly continuous shot and produce very similar keyframes for the resulting segments. Significant shots are rarely omitted as indicated in the preceding paragraph and noted in the individual video commentaries. The approach to shot segmentation appears to be very aggressive.

As shown in Appendix A, the Annotation Tool does produce MPEG-7 compliant XLM file as a primary end product of the annotation process. The XML code appears to track the shot boundaries. The first time a video is loaded by the user is informed that there is “No Frame Map file (.frp) and Shot Segmentation, Generating new ones”. Generating the frame map file and shot segmentation is the pre-processor function of the Annotation Tool in preparation for creating the XML output and is a relatively time consuming process. When the video is loaded on subsequent occasions, the information saved in the .frp files allow for quick retrieval of all video frames. Some of this information is likely included in the generated XML file as indicated by the following human readable yet still arcane code snippet.

```
<VideoSegment>
  <TextAnnotation type="scene description" relevance="1" confidence="1">
  </TextAnnotation>
  <MediaTime>
    <MediaTimePoint> T00:03:49:22886F30000 </MediaTimePoint>
    <MediaIncrDuration timeUnit="PT1001N30000F"> 587 </MediaIncrDuration>
  </MediaTime>
  <SpatioTemporalDecomposition>
    <StillRegion>
      <MediaIncrTimePoint timeUnit="PT1001N30000F"> 7179 </MediaIncrTimePoint>
      <SpatialDecomposition>
      </SpatialDecomposition>
    </StillRegion>
  </SpatioTemporalDecomposition>
</VideoSegment>
```

Conclusions

MPEG-7 is the standardized collection of metadata definitions for describing digitized multimedia source materials. Standardization creates an environment where the description of multimedia data will be portable between software applications. This is possible because the resultant metadata are delivered to applications via XML data files. XML is the messenger and encapsulated metadata is the message. The defining quality of XML is that it is extensible. Thus, it is fairly simple to accommodate any additional metadata descriptors that may be subsequently introduced into the MPEG-7 standard. MPEG-7 is therefore flexible, adaptable, and able to embrace technological innovation as well as commercially desirable features.

Feature extraction is the process of identifying the components of multimedia data that we wish to describe, applying a suitable description to those components, and storing those descriptions for later retrieval. The general subdivision of features into low-level

and high-level features is indicative of a significant factor that must be addressed in any indexing methodology. When is a rose not a flower but a symbol on a coat of arms. When is a cross not a religious symbol but a political icon. Automated feature extraction technologies will be relegated to the low-level feature domain for the foreseeable future. It is, and will continue to be, the task of the archivist to interpret the relationships between objective data and the semantics of the social and historical context depicted therein. Automated feature extraction technologies are tools to assist the archivist. Tools to relieve the tedium of repetitive tasks and allow the archivist to concentrate on the semantics, become more efficient, and produce more consistent results when indexing multimedia data.

Commercial research labs are now demonstrating the capabilities of currently available multimedia feature extraction tools. The IBM MPEG-7 Annotation Tool is one such software application that has the capability of extracting keyframes from MPEG video streams. Testing the Annotation Tool with a video collection that had been manually processed to extract keyframes demonstrated the power of tool and revealed its limitations.

The power of the Annotation Tool is evinced by its capacity to determine shot boundaries and extract keyframes for those shots. That the tool performs this task even reasonably well is justification for granting much respect to the tool's programmers and is an indication of that this technology is indeed viable. The application of the underlying technology to producing program output may best be described as aggressive or should perhaps be characterized as too much is better than not enough. The automatically generated MPEG-7 compliant XML output makes this tool useful for metadata storage to

the extent of the tool's possible output. IBM has made this tool available for evaluation and it is not intended as a production tool. However, this tool would provide a welcome assist to an archivist working with a relatively small collection of video data. It is easy to envision that this tool could become immensely more attractive with the addition of a few additional capabilities and the extension of those already present.

An obvious addition to the Annotation Tool's capabilities would be the integration of a speech to text engine. Though speech to text technology is still a work in progress, automatic transcription to the extent that keywords may be effectively extracted would be very attractive even in a demonstration software product and would be a logical next step in the evolution of this type of software. Since the preprocessing of video data by the tool is very time consuming, an option to batch preprocess a collection video data files would greatly enhance the usability of the product.

The Annotation Tool's deficiencies are mainly due to an excess of data returned to the user in terms of shots and keyframes to be annotated. The tool does not recognize multiple shots as parts of a scene. As a result, numerous identical or nearly identical keyframes may be returned for a series of shots in which each shot is merely a change in point of view. Repetitions also occur when the video zooms from a distance to near close-up and vice versa or when a moving object moves sufficiently far off for the tool to recognize this as a new shot. A solution to this situation would allow for a certain amount of user intervention allowing multiple shot sequences to be identified as a single shot while allowing multiple keyframes per sequence.

One notable result of the test procedure was the fact the Annotation Tool did not tolerate poor video quality. Of the twenty-six videos available for processing, only ten

videos were successfully processed by the tool. In all but one of those sixteen cases, poor video quality was observed at the approximate instance when the tool failed. In the remaining case, no indication for cause of program failure was discovered. Of the sixteen cases where processing was unsuccessful, in only one case did the video fail to be viewable with available mpeg video viewers. The tool should be designed to handle a greater level of bad mpeg data without causing the program to abort. Of course the onus could also be placed on the archivist to ensure that acceptable quality mpeg data be provided as program input.

The comparison of automatically selected keyframes with manually selected keyframes for the test collection points out the need for automated assistance in the determination of shot boundaries and keyframe selection. Some of the manually analyzed videos were carefully reviewed with appropriate keyframes selected to adequately represent the video content, often they were not. The basic failing in manual keyframe extraction was the lack of consistency. While video duration was not an absolute indicator of that an insufficient number of keyframes would be proffered to describe any particular video, there was definitely a tendency for the longer videos to be insufficiently described. This lack of consistency is one of the primary issues that is adequately addressed by the Annotation Tool.

Multimedia source material may be analyzed and described in myriad different ways. MPEG-7 formalizes the metadata structures for those descriptions. Feature extraction tools will automatically populate those metadata structures with data and XML data files will deliver the data to applications that will organize and store that data for

search and retrieval. The archivist will have the tools necessary to effectively and efficiently categorize, describe, index, store, and retrieve multimedia source materials.

Appendix A

Sample XML Output From IBM MPEG-7 Annotation Tool

```
<Mpeg7 type="complete" xmlns="http://www.mpeg7.org/2001/MPEG-7_Schema"
xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
xsi:schemaLocation="http://www.mpeg7.org/2001/MPEG-7_Schema">
  <ContentDescription xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="VideoType">
      <Video>
        <TemporalDecomposition>
          <VideoSegment>
            <TextAnnotation type="scene description" relevance="1" confidence="1">
              <FreeTextAnnotation>
                Outer_Space
              </FreeTextAnnotation>
              <FreeTextAnnotation>
                Human
              </FreeTextAnnotation>
              <FreeTextAnnotation>
                Person(w/o_face)
              </FreeTextAnnotation>
              <FreeTextAnnotation>
                Transportation
              </FreeTextAnnotation>
              <FreeTextAnnotation>
                Rocket
              </FreeTextAnnotation>
              <FreeTextAnnotation>
                Take-Off/Launch
              </FreeTextAnnotation>
              <FreeTextAnnotation>
                astronaut
              </FreeTextAnnotation>
            </TextAnnotation>
            <MediaTime>
              <MediaTimePoint> T00:00:00:0F30000 </MediaTimePoint>
              <MediaIncrDuration timeUnit="PT1001N30000F"> 1106
</MediaIncrDuration>
            </MediaTime>
```

```
          <SpatioTemporalDecomposition>
            <StillRegion>
              <MediaIncrTimePoint timeUnit="PT1001N30000F"> 552
</MediaIncrTimePoint>
              <SpatialDecomposition>
                    <StillRegion>
                      <TextAnnotation>
                        <FreeTextAnnotation>
                          Outer_Space
                        </FreeTextAnnotation>
                      </TextAnnotation>
                      <SpatialLocator>
                        <Poly>
                     <CoordsI> 0 0 0 0 0 0 0 0 </CoordsI>
                    </Poly>
                  </SpatialLocator>
                 </StillRegion>
                      •
                      •
                      •
                      •
                      •
                      •
          </SpatioTemporalDecomposition>
         </VideoSegment>
       </TemporalDecomposition>
     </Video>
   </MultimediaContent>
  </ContentDescription>
</Mpeg7>
```

Appendix B

Figure 1a: Manual Keyframes for Video #2

Figure 1b: Automatically Generated Keyframes for Video #2, 1 – 30 of 236

Figure 2a: Manual Keyframes for Video #3

Figure 2b: Automatically Generated Keyframes for Video #3, 1 – 30 of 143

Figure 3a: Manual Keyframes for Video #6

Figure 3b: Automatically Generated Keyframes for Video #6, 1 – 30 of 41

Figure 4a: Manual Keyframes for Video #7

Figure 4b: Automatically Generated Keyframes for Video #7, 1 – 30 of 59

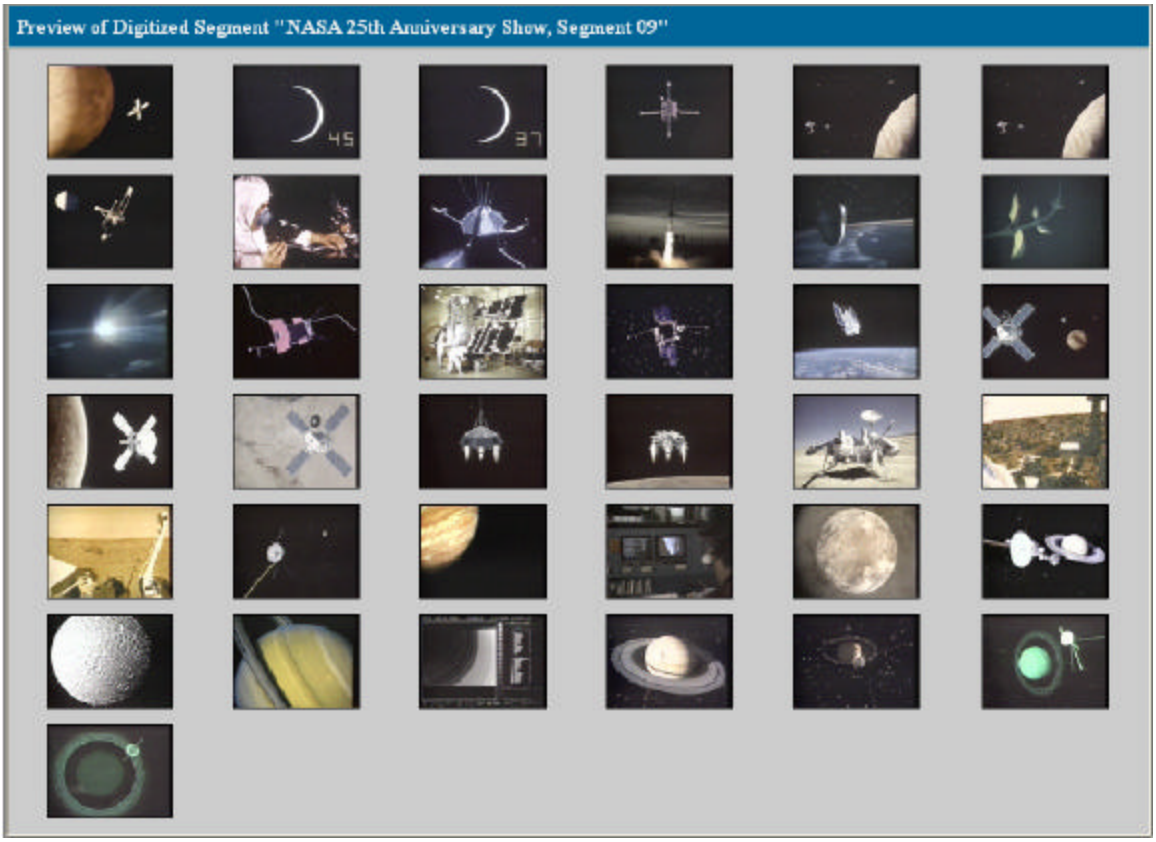

Figure 5a: Manual Keyframes for Video #8

Figure 5b: Automatically Generated Keyframes for Video #8, 1 – 30 of 60

Figure 6a: Manual Keyframes for Video #9

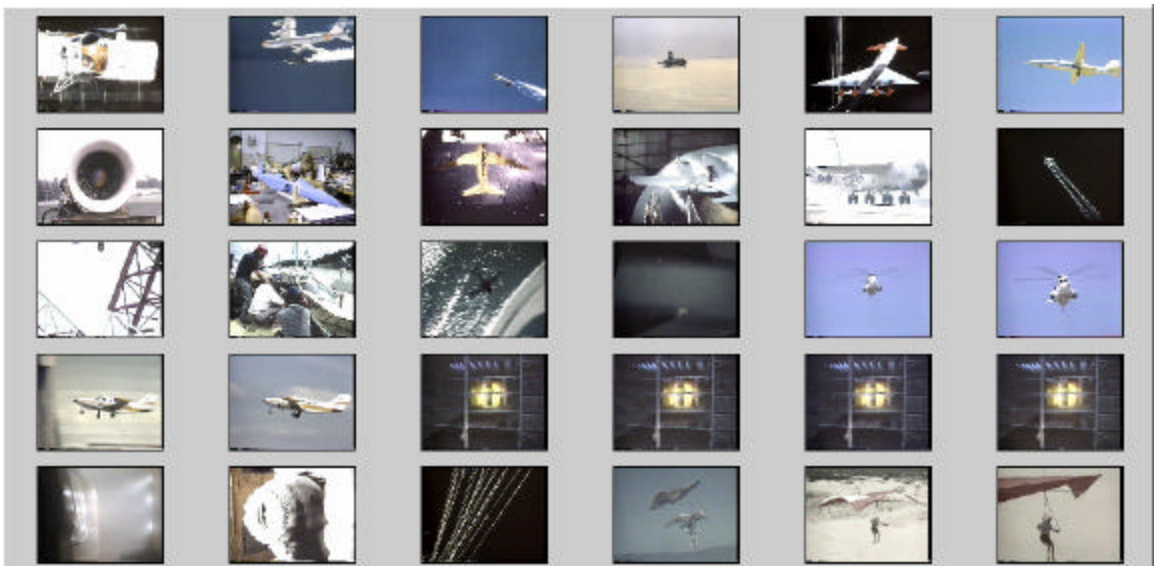

Figure 6b: Automatically Generated Keyframes for Video #9, 1 – 30 of 132

Figure 7a: Manual Keyframes for Video #10

Figure 7b: Automatically Generated Keyframes for Video #10, 1 – 30 of 100

Figure 8a: Manual Keyframes for Video #12

Figure 8b: Automatically Generated Keyframes for Video #12, 1 – 30 of 175

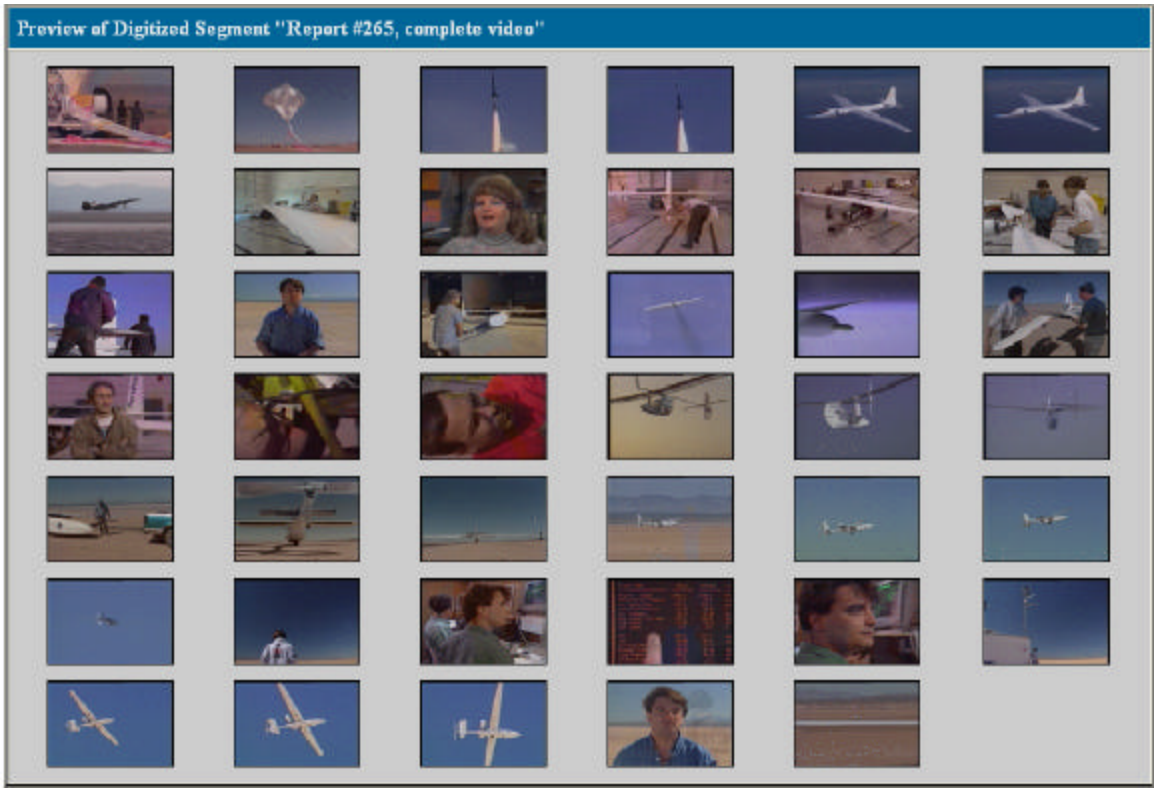

Figure 9a: Manual Keyframes for Video #14

Figure 9b: Automatically Generated Keyframes for Video #14, 1 – 30 of 252

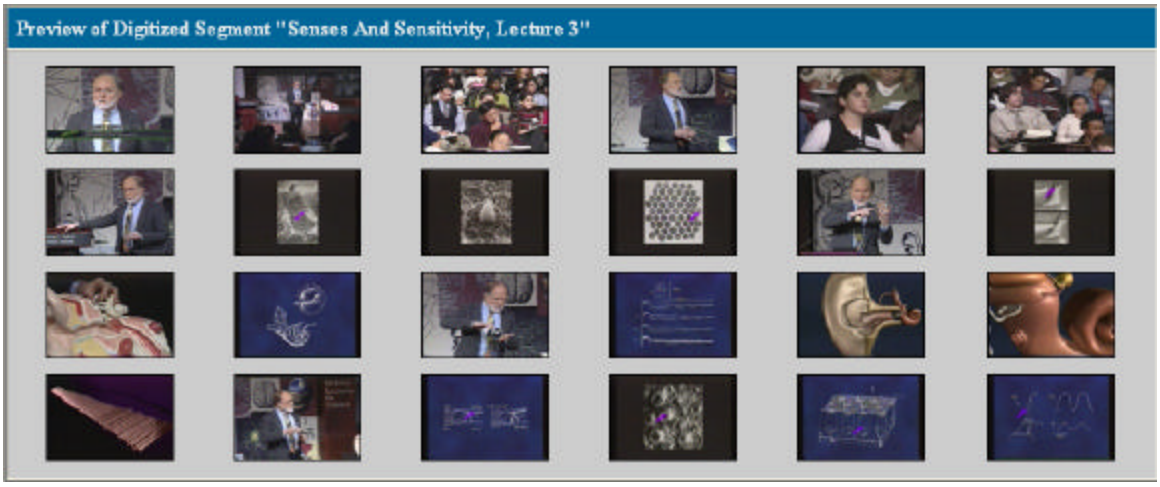

Figure 10a: Manual Keyframes for Video #15

Figure 10b: Automatically Generated Keyframes for Video #15, 1 – 30 of 225

References

1. *ISO/IEC JTC1/SC29/WG11 N4675 Introduction to MPEG-7 (v4.0),* N. Day & J. M. Martinez, ed., MPEG Requirements Group, Jeju, Mar. 2002

2. *SMPTE Metadata Dictionary (RP210.2) (including RP210.1) - Merged version December 2001 post trial publication of RP210.2,* SMPTE Registration Authority, LLC, http://www.smpte-ra.org/mdd/

3. *The Dublin Core Metadata Element Set*, Dublin Core Metadata Initiative http://dublincore.org/documents/dces/

4. *EBU Technical Review No. 284,* September 2000 http://www.ebu.ch/trev_284-hopper.pdf

5. *ISO/IEC JTC1/SC29/WG11/N2467 Description of MPEG-7 Content Set*, MPEG Requirements Group, Atlantic City, Oct. 1998

6. *ISO/IEC JTC1/SC29/WG11/N4981 MPEG-7 Requirements Document V.17***,** F. Pereira, ed., Klagenfurt, July 2002

7. An overview of the MPEG-7 description definition language (DDL), Hunter, J. Circuits and Systems for Video Technology, IEEE Transactions on , Vol.11, Iss.6, 2001, Pages: 765- 772

8. The holy grail of content-based media analysis, Shih-Fu Chang Multimedia, IEEE, Vol.9, Iss.2, 2002 Pages: 6- 10

9. Content-based multimedia indexing and retrieval Djeraba, C. Multimedia, IEEE, Vol.9, Iss.2, 2002 Pages: 18- 22

10. MPEG-7 the generic multimedia content description standard, part 1 Martinez, J.M.; Koenen, R.; Pereira, F. Multimedia, IEEE, Vol.9, Iss.2, 2002 Pages: 78- 87

11. Standards - MPEG-7 overview of MPEG-7 description tools, part 2 Martinez, J.M. Multimedia, IEEE, Vol.9, Iss.3, 2002 Pages: 83- 93