RhythmPix: A Multimedia Composition and Albuming System for Consumer Images

Alexander Loui, Bryan Kraus, and Jon Riek[#]
Photographic Science & Technology Center
Eastman Kodak Company, Rochester NY 14650-1816 USA
alexander.loui@kodak.com

Abstract

With the advent of new digital consumer electronic devices, the interest and demand for audiovisual media and content authoring has increased in recent years. The RhythmPix project is to explore, develop, and demonstrate technology for albuming and authoring of multimedia image content for playback on consumer electronic devices such as DVD players, as well as PCs. In this paper, we describe an end-to-end multimedia albuming system that we have developed as a platform for combining images and sound/music content. The three main functional modules – media composition, media encoding, custom recording, and related implementation issues, are described. A set of APIs has been developed for rapid system adaptation. Various user studies and focus groups have been conducted in the US and Asia to validate the concept and functionality of the system.

1. Introduction

The popularity of home theatre equipment converging music and imagery, and the unprecedented diffusion of digital versatile disc (DVD) and video CD (VCD) players world wide, have created new opportunities for multimedia products and services [1-3]. Together with the increasing use of digital imaging in general consumer applications, there is a great interest to develop new output products that increase the value and enjoyment level of viewing digital images in consumers' living rooms. The RhythmPix project was initiated on multimedia authoring to explore the opportunities in this area with an emphasis on composing still images with sound, including music and spoken annotations.

The goal of this work is to research and develop efficient image and multimedia albuming tools and systems that will enable playback of multimedia CDs using popular consumer electronic equipment. New image and multimedia composition tools have been developed to facilitate the implementation of an end-to-end authoring system. The availability of such multimedia CDs will greatly enhance the value of digital pictures, as it provides an alternate path and a new experience to view digital photographs in consumers' living rooms rather than on their PCs. The technology will enable the viewing of photographs (with multimedia enhancement) in consumer electronic devices such as VCD, SVCD (super VCD), and DVD players, as well as on PCs.

The vision of RhythmPix aims to enrich image sharing with music and style. This work is focused on offering services and products that enable users to collect, combine, and compose images, music, and other relative content, to create expressive, easy-to-share presentations. The media presentation will leverage creative composition to integrate the various forms and provide enjoyment for one or many viewers simultaneously. The goal is to deliver an enjoyable and exciting experience on the first cut, with very little user input and effort. This paper is organized as follows. Section 2 describes the RhythmPix multimedia albuming system that has been developed. Three main modules, namely the media composition module, the media-encoding module, and the custom-recording module, are detailed. The system prototype and user study are presented in Section 3. Finally, a summary is given in Section 4.

2. RhythmPix Albuming System

The core of the RhythmPix authoring system consists of three modules: 1) media composition (including automated image grouping), 2) media encoding, and 3) custom recording. A block diagram of the authoring system is depicted in Figure 1. The three modules are communicated via a Windows-based user interface. The media composition module accepts various input media sources including images, video, audio, text, and graphics. The input media sources can be grouped, edited, and composed together (e.g., images with audio and voice annotations). The output of the media composition module will be fed to the media-encoding module where the visual and audio content will be compressed according to the MPEG-1 video compression standard in this case. The various compressed bitstreams may be fused together to produce a final output bitstream. Finally, the output-compressed bitstreams will be

[#] was with Kodak, now with VirtualScopics LLC, Pittsford, NY

0-7803-9332-5/05/$20.00 ©2005 IEEE

written to the specific playback media, e.g., CD-R or DVD-R, via the custom recording module. The authoring system also allows the inclusion of other content (e.g., original high-resolution images) as well as other image rendering and processing applications in the same disc medium.

2.1 Media Composition

2.1.1 Image Composition

The media composition module consists of two sub-module. The image-composition sub-module uses a set of image processing and transform operations developed in-house for all image I/O, editing, and other image-processing related tasks. The VCD standard specifies an image size of 352 × 240 for the NTSC system. Image prefiltering and resampling operations are used to resize images. The resampling is set up to resize the input so that it will fit in the desired image size. If the image was resampled using the SHRINK mode, it is likely that it is smaller than the specified size. In this case, a black image of the specified size is created, and the two images are composed together using an image-composition operation. If the image was resampled using the EXTEND mode, it is likely to be larger than the specified size. In this case, the image is cropped to the specified size via an image-cropping operation. The specific region to be cropped from the image is determined by user input. To make the images look better on a television, a small amount of sharpening is applied using an image-unsharp masking operation. The kernel used is a 5 × 5 box kernel and the gain in the unsharp mask is set to 0.5. To further improve the image quality, an enhanced black printing (EBP) algorithm was applied. The EBP algorithm maximizes the contrast of the image so that the entire range from 0 to 255 is utilized. When sharpening the image in very bright regions, this can cause clipping to occur. By sharpening the image prior to enhancing the contrast, the amount of clipping is minimized.

Another important function of the image composition sub-module is to automatically organize the set of input images into event groups for different compositions, e.g., each event will use different music. This is accomplished by incorporating an event clustering algorithm that automatically groups images by the use of a two-means clustering method with a block-based histogram correlation technique as described in [4]. This feature greatly reduces the time spent organizing and ordering the input images for final authoring.

2.1.2 Audio Composition

One of the main tasks of the audio-composition sub-module is to combine multiple-audio input streams (including speech annotations, music, and other audio) in different file formats and specifications from various input sources. The basic workflow of the audio-composition module is as follows. The module takes as input, two types of audio streams: 1) a list of speech annotations (directly from a microphone or from a file), and 2) a list of music files. In addition, a parameter file specifying the display characteristic of the images is input to the composition sub-module. The default display time will be calculated by dividing the length of the sound/music file by the number of images for authoring. The output of the module is a single 16 bits/sample, 44.1 KHz, stereo PCM stream in MS wave format. The timing parameters are also passed on to the MPEG encoding module for audiovisual compression.

The major steps for audio composition include: 1) extraction, e.g., taking music from audio CD tracks; 2) normalization, which converts all kinds of audio streams to the same normalized specifications of the sample precision, sampling rate, and number of channels; 3) alignment, which determines timing parameters; and 4) composition, which merges all audio streams into a single stream by combining the samples at the same time instance together. In case of the presence of speech annotations, the combination step will emphasize them with respect to the background music. The current audio-composition module is capable of handling most popular audio formats.

2.2 Media Encoding

The VCD specification [5] is based on MPEG-1. Therefore, the compression routine generates an MPEG-1 sequence. There are certain limitations on the MPEG-1 bitstream contained within the VCD specification. For NTSC, the picture rate is 29.97 Hz, and for PAL, it is 25 Hz. The constrained bitstream parameter must be set to true. This implies a constant bit rate that cannot exceed 1,151,929.1 bits/second. In addition, the VBV buffer size cannot be greater than 327,680 bits (40 Kbytes) [5]. This implies that no one frame can take more than 327,680 bits to encode.

To encode an image to remain on the screen for N seconds, it needs to be encoded $29.97N$ times for NTSC and $25N$ times for PAL. To encode a still image in MPEG, the image is encoded as an intra-coded picture. This is very similar to JPEG in that the image is converted to YC_bC_r, the chroma channels are subsampled, and the image is encoded using 8 × 8, quantized DCTs. The intra-coded pictures in MPEG are referred to as "I" pictures. Another encoding type available within MPEG is a predicted or "P" picture. In a P picture, each 16 x 16 macroblock may be encoded as in an I picture, or it may be predicted from the previous I or P picture. To leave the image on the screen as it was encoded in the I picture, all that is required is to encode each macroblock as predicted. If all the macroblocks are predicted, MPEG provides a facility to skip all the macroblocks except the first and the last. It results in only 232 bits being required in an NTSC P picture to predict the

entire image from the previous picture. For PAL, 256 bits are required.

This creates a small rate control problem. Our buffer can only hold 327,680 bits, and bits are being read into the buffer at a rate of approximately 1,152,000 bits per second or 38,400 bits per NTSC picture, and 46,080 bits per PAL picture. The bits are only being read out of the buffer at a rate of 232 bits per NTSC P picture and 256 bits per PAL P picture. If the I picture took 300,000 bits to encode, it would only take about seven NTSC P pictures or six PAL P pictures before the buffer was overflowing. This is illustrated in Figure 2. To leave an image on the screen for five seconds, it would require either 125 or 150 pictures. From a compression-efficiency point of view, we would like to encode one I picture followed by 124 or 149 P pictures. Since a constant bit rate is required, every time the buffer begins to fill up, the entire picture needs to be recoded as an I picture. Referring to practical experience, this can be as few as every three to four pictures. For efficiency sake, the I picture is stored in memory and copied out to the bitstream every time it is needed. The only parameter that needs to be changed is the temporal reference of the I picture.

2.2.1 Coding Transitions in MPEG

The authoring application is capable of producing various transitions between images. Currently, the transition effects are created by using a technique that takes advantage of the bi-directionally "B" predicted pictures in the MPEG encoding structure. Specifically, a B picture can copy macroblocks from the previous and next I or P picture. One way to do a transition in MPEG that is not data-dependent and does not require any motion estimation is to "uncover" the second image. The idea is that we are currently displaying a first image, and selectively change (16 × 16) macroblocks to display the second image. When all the macroblocks have been changed to display the second image, the transition is complete.

An example (see Figure 3) may help to clarify this idea. If the transition we would like to encode uncovers the image from top to bottom, the first picture is currently displayed on the screen. To perform the transition, the first row of macroblocks is switched to display the first row of macroblocks from the second picture. Next, the second row of macroblocks is switched to display the second row of macroblocks from the second picture, and so on, until all the rows of macroblocks correspond to the second image. Basically, each block contains a number representing a macroblock. The ones labeled "1" correspond to macroblocks in the first image. The ones containing a "2" correspond to macroblocks in the second image. Each intermediate B picture contains just macroblocks that have a flag to copy the macroblock from either the first picture or the second picture. Any type of uncover transition can be created just by determining when to uncover which macroblocks. Transitions that are more complicated can also be created by coding of motion vectors.

2.2.2 High-Resolution Still Encoding

The VCD 2.0 standard has a provision to encode a higher resolution of the input image frame. This option will allow higher quality display of images for menus, graphics, and professional content. In the high-resolution mode, the MPEG I-frame is encoded at four times the normal resolution – 704 × 480 for NTSC or 704 × 576 for PAL. The high-resolution encoding is realized via the Mixed Resolution Still Picture stream, as described in the VCD 2.0 standard. This implies that the high-resolution (or the Mixed Resolution Still Picture) stream is not exactly compliant with the MPEG-1 standard. The high-resolution still pictures will be treated as a segment play item and stored in the SEGMENT directory on a Video CD. The details of the VCD high-resolution still encoding and rendering can be found in [6].

2.3 Custom Recording

The custom recording module is comprised of two sub-modules: 1) a custom library called VcdGen that creates a disc image consisting of two files — a bin and a cue file, and 2) a CD recording library that uses the bin/cue files to create the final VCD. The VcdGen software library allows properly encoded sequence and segment items (i.e., MPEG bitstreams) to be built quickly and easily into VCD image files. The library provides full control over each image, including the content, organization, and playback options. This control is exercised through a custom CD description XML file created by the library. The second sub-module is implemented using a commercial CD writing library.

The CD recording module has a provision to include original image content (e.g., high-resolution JPEGs) in the same disc. Another option is implemented to co-locate an image-rendering and processing application in the same disc for PC playback. When this option is checked, all the images in the active workspace area of the authoring application will be converted to the appropriate format, such as JPEG to be included in the appropriate directory for the PC application. Finally a menu option is implemented to facilitate some albuming function. This will allow the user to organize different event groups (from the composition module) into different tracks in the VCD for viewing. With the menu option, the user can choose between a default background, or a user-specified image as the menu page.

3. System Prototype

A prototype system of the RhythmPix multimedia authoring application described in the above sections has been

designed and implemented to demonstrate end-to-end functionality and performance. A snapshot of the current user interface for MS Windows operating systems is shown in Figure 4. The prototype application is built using Mircorsoft Visual C++. The prototype application is compatible to all current Windows platforms. In addition, we have created a set of APIs for system developers. The RhythmPix libraries are released as building blocks such that they can be integrated into other multimedia authoring applications and systems. The input includes a number of still images, music, and speech annotation files, and the output will be a VCD-compliant MPEG-I system bitstream and disc image files. The API includes several static and dynamic libraries.

A number of user studies and focus groups have been conducted in the US and Asia to validate the concept and functionality of the RhythmPix authoring system. Most users were enthusiastic and appreciated the picture-sharing concept through a DVD player in the living room. Particularly some users indicated that this was better than video, as watching a video seems to drag on for most people. In addition, most users indicated that the tool is relatively easy to use, especially with the automated event organization feature. In terms of speed performance, we observed that the compression time varies, depending on the format of the sound/music content and, to a lesser extend, the image format. The CD recording time is fairly fast at approximately 100 seconds for CD-R media and around 170 seconds for CD-RW media using a HP 12X CD writer, for a set of 25 JPEG images with a 205 seconds music file.

4. Summary

In this paper, we presented the RhythmPix multimedia albuming system that has been developed as an end-to-end platform for combining images and sound/music content. The functionality and implementation issues of the three main modules – media composition, media encoding, and custom recording are discussed. A set of APIs has been developed for system adaptation. Various user studies have validated the composition functions and authoring process developed. We are conducting more research to further automate the authoring process and to improve the user interface design based on the feedback from user studies.

5. Acknowledgments

The authors would like to thank Zhaohui Sun and Phoury Lei for their contributions to this work.

6. References

[1] Ulead DVD PictureShow application, 2003.
[2] Oak Technology SimpliCD application, 2003.
[3] Dazzle OnDVD 2.0 application, 2003.
[4] A. Loui, and A. Savakis, "Automated event clustering and quality screening of consumer pictures for digital albuming," *IEEE Trans. on Multimedia,* Vol. 5, No. 3, Sept. 2003.
[5] Video CD Specification Version 2.0. Philips Consumer Electronics B.V., April 1995.
[6] Z. Sun, J. Riek, and A. Loui, "High resolution multimedia slide show composition for Video CD and DVD rendering," *Proc. IEEE ICME '03*, Baltimore, MD, July 6-9, 2003.

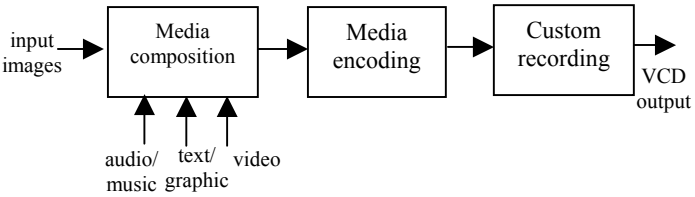

Figure 1. RhythmPix system block diagram

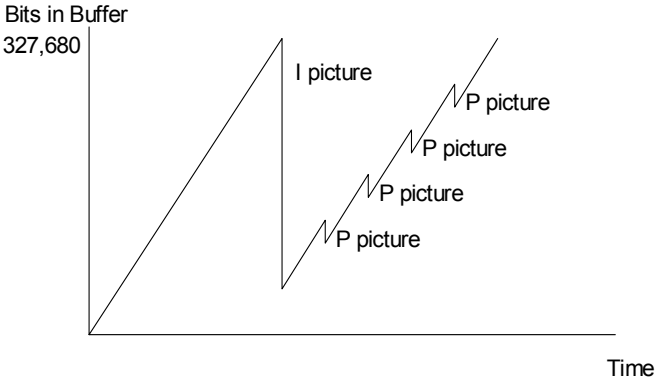

Figure 2. Pictorial description of the decoding buffer

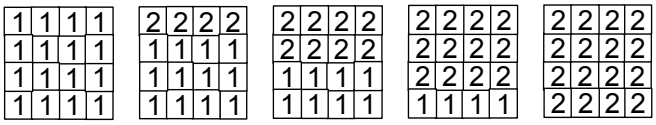

Figure 3. Top to bottom uncover transition

Figure 4. A snapshot of the RhythmPix albuming application