MusicStory: a Personalized Music Video Creator

David A. Shamma, Bryan Pardo, Kristian J. Hammond

Intelligent Information Laboratory
Northwestern University
1890 Maple Avenue, 3rd Floor
Evanston, Illinois 60201 USA
{ayman, pardo, hammond}@cs.northwestern.edu

ABSTRACT

In this paper, we describe MusicStory, a system that automatically creates videos to accompany music with lyrics. MusicStory uses common search engines, photo-sharing websites, and simple analysis of the dynamics and tempo of the music to create personalized photo-narratives. Video pacing and content is based on the content of the song and structure of the image repositories selected. The image associations MusicStory presents amplify the emotional experience by externalizing the imagery in song lyrics with the content found within a social network. The resulting work juxtaposes the meanings inherent in the social network with those in the song.

Categories and Subject Descriptors

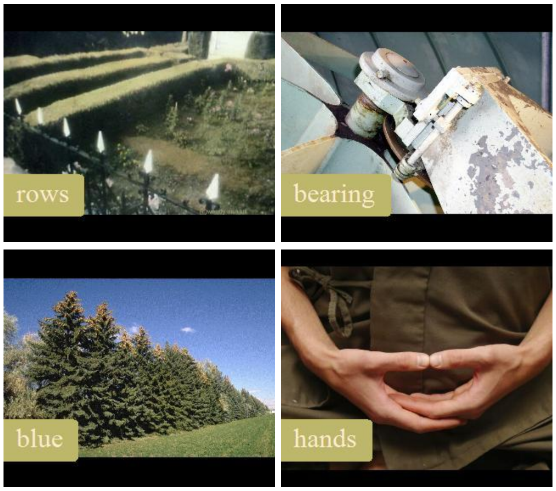
J.5 [**Arts and Humanities**]: Performing arts (e.g., dance, music); H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Web-based interaction*

General Terms

Performance, Design, Human Factors

Keywords

Media Arts, Music Video, Network Arts, Photo Sharing, Software Agents, World Wide Web



Rows of houses all bearing down on me
I can feel their blue hands touching me

Figure 1: MusicStory finds images of songs and makes a slide show for either a portable MP3 player or for a venue's back display. Pictured here are the images and lyrics of the Radiohead song *Street Spirit (Fade Out)*.

1. INTRODUCTION

Our personal media collections, from the music on an iPod to the images in a photo album, reflect who we are as individuals. Public media collections, from the Library of Congress, to commercial image repositories, to the songs put up on a local band's website, reflect who we are as a culture.

Today, media collections are increasingly stored on-line; public images can be found using web search engines. Music sites, such as iTunes and Rhapsody bring large public collections of music to the web. Weblogs, personal photo software (like Picasa), and social photo-sharing websites (like Flickr) bring personal media into the web-searchable, digital realm. This provides an unprecedented opportunity to explore the relationships between personal media and public media in the context of automated, web-based, installations. This genre of installation is called Network Arts [14].

At the core of Network Arts are technological advancements in information retrieval, social networks, and video/audio processing, and a new cultural understanding of meaning, impact, and artistic portrayal. In recent Network Arts pieces, Ruberry et al. [10] explored new ways of experiencing weblogs through automated "live acting" and Shamma and Hammond [12] experimented with the emotional imagery inherent in televised political speeches. For the work covered in this paper, we set out to build an autonomous music video creator that takes an audio file (mp3, wma, etc.) as input and output a music video file (MPEG, wmv, or 3gp). The video creator, MusicStory, builds a video for a piece of music using relevant public or personal images. Demonstration videos are available at `http://www.infolab.northwestern.edu/musicstory/`.

2. RELATED WORK

Much work has been done in automatic music video creation. Our work bears the most likeness to P-Karaoke [4]. Both Mu-

MM'05, November 6–11, 2005, Singapore.
Copyright 2005 ACM 1-59593-044-2/05/0011 ...$5.00.

sicStory and P-Karaoke rely on the Microsoft DirectX framework for the video creation and both systems examine the beat of the audio track. P-Karaoke filters undesirable images (blurred, bad exposure, duplicates, etc.) and selects capriciously from the final candidate list. P-Karaoke also relies on exact beat transitions which includes the necessity for a syllable by syllable time-stamped lyric file (found online or manually entered). MusicStory uses an estimate of the beat (actually a multiple) which provides enough for a software agent to direct the remainder of the video performance. Additionally, MusicStory discovers images (locally or online) linked to the words in the lyrics. The end result is a video which brings new and unexpected imagery to the viewer, based on images, textually indexed, and related to the song itself.

In addition to P-Karaoke, there are other automatic music video systems, such as Muvee's AutoProducer and Microsoft's PhotoStory. They require the user to specify the song's speed and to hand select their local photos on disk. These systems provide assisted music video creation; they construct photo narratives with an audio soundtrack. Since MusicStory's images are semantically tied (via web-indexing and community tagging) to the song and its lyrics, it brings a new experience, a musical narrative with discovered imagery.

3. MUSIC NARRATIVES

Like a human listener, MusicStory processes the lyrics in the music and these lyrics bring forth associations with images. The imagery chosen by MusicStory is defined by the set of links between words drawn from the lyrics and image-word associations contained in a social network, either private or public. As the images are found, it presents them to the audience, creating an on-the-fly music video, heightening, clarifying, and exposing the connections between words, ideas, and images that we are often unaware of, until shown. Figure 1 shows a slide-by-slide expansion of the imagery from the lyrics of a Radiohead song.

When listening to the music, the image associations that MusicStory presents amplify the emotional experience and heighten its visceral appeal by externalizing the concrete imagery intrinsic in the lyrics. Some images depict the expected relation found in the song, while others present a juxtaposition between the song's meaning and the meaning found within the social network. Our approach focuses on the creation of a photo narrative, to compliment the music and not the strict alignment of images to lyrics as we have shown in the Imagination Environment [7, 12, 16]. Artistically speaking, the strict alignment of images-word pairs to lyrics or spoken dialog provides an amplification of meaning through free association [11]. For MusicStory, we rely on this amplification of meaning in the context of the song itself, and not the individual words being communicated.

MusicStory uses public media to retrieve images with popular relevance (relying on web frequency as a measure of familiarity and salience [13]), returning images that reflect current pop-culture meanings. More personal images can be found by focusing the retrieval to smaller social networks, such as personal photo-sharing sites.

For example, the word 'home' from the song *Sweet Home Alabama* retrieves different associations from different repositories. Google's ranking returns a canonical photo of a home from a realtor's website. Flickr, in contrast, returns photos people took in their home, in this case of their child, see figure 2. The combination of image repositories (popular, canonical, and personal) provide a balance of associations which the agent uses to aid in the art's creation and assists in the audience's understanding of the work [11, 5].

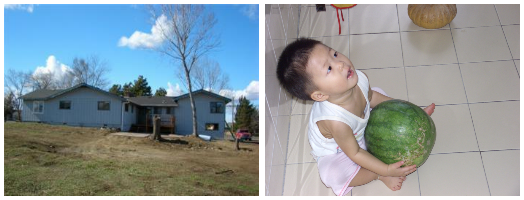


Figure 2: Image search for 'home' returns two types of associations: one from a popular search engine, another from a photo sharing website.

4. SOFTWARE AGENT ARCHITECTURE

While it is possible to use simple search to create a sequence of images overlaying music, that does not make for a successful piece. Successful integration of sound and image relies on an intimate knowledge of the media itself, considering available image resources (repositories, number of images per term), musical parameters (tempo, dynamics, density, lyrics) and the output format (screen size, playback bit rate). To do this, MusicStory assumes the role of a director, concerning itself with the overall flow and pacing of the resulting multimedia performance.

4.1 Artistic Information Agent

MusicStory's director is an Artistic Information Agent. This agent is a variant of the Information Management Assistant [1] used in Information Retrieval. The agent's architecture shows several adapters that let it access online information sources. The core functionality is separated into four basic components, three of which are borrowed from a previous architecture. [14]

The Artistic Analyzers for the MusicStory consist of a *listener* and a *presenter*. The listener feeds in the audio information from a source and the meta-data (some meta-data, such as lyrics, is not carried within the source file). The presenter controls how the final movie is created.

Internally, the listener agent queries for images as sources are delivered via a set of application adapters. Once a set of candidate media (to display) is created, the agent decides what to present based on the current flow state. The overall flow state is determined by analyzing the input stream from the listener, the output stream to the presenter and available real estate on the presenting media itself. The flow of imagery moves with the pace of the song, providing quick transitions through fast songs, and leisurely transitions through slower songs.

4.2 Finding Pace

When building a Network Arts installation, one must remember the audience. It is important the audience be engaged and connected with the installation and its performance. Keeping this connection will allow the piece to create more emotional and captivating moments with the viewer, following Kandinsky's model of expressionism [5].

Foremost, the agent must determine the pacing of the installation and hence performance. There are two pacing metrics, the tempo of the source (input) media and the desired tempo of the overall (output) performance.

The tempo of a slow ballad does not match that of a live speech or a fast hip-hop song. MusicStory bases its rate for presenting images on the pace of the media. To accommodate several media sources, we created a model of presentation for the agent. The model's presentation pace is set to complement the pace of the source media. As a result, an effective flow state for the overall installation is achieved. To keep the flow state engaging, thresholds are set to keep the images from changing too quickly or too slowly, which prevents the audience from being overwhelmed or becoming bored.

A simple operationalization of Mihaly Csikszentmihalyi's flow model [3] suffices for our purposes. While, Csikszentmihalyi describes human activities as a compromise between two components, challenges and skills, we focus on the flow channel itself and the neighboring anxiety and boredom outside the channel.

MusicStory needed a descriptor of the pace of the song to operationalize our modified Csikszentmihalyi pacing model. In a previous Network Arts installation, which used broadcast media, Shamma et al. estimated pace by using the rate of the closed captioning feed [14]. Given the input rate and knowledge of how fast information can be displayed (output rate), Csikszentmihalyi's flow model can be put to use by the agent to optimize the display and interaction.

In our current work, pacing is affected by song tempo and the agent's artistic intent. The directing agent adjusts how long a image is displayed and the speed of transition between images, influenced by placement of peaks in the volume of the audio. The final slide show pacing does not map strictly to beat-by-beat image transitions, but visually moves at a speed complimentary to a multiple of the beat.

To make this adjustment, the agent first needs to know the general pace of the song. For many kinds of popular music, good synchronization points for the video correspond to peaks in the root-mean-squared amplitude of the audio signal (RMS). To use RMS for this application, we must look at the structure of the digitally encoded song. For simplicity, a compressed audio file is converted to linearly encoded PCM audio.

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}, \text{ where } \left\{ \begin{array}{lcl} x & = & \text{sample's amplitude} \\ n & = & \text{window size} \end{array} \right. \tag{1}$$

Finding the beat in music is often problematic [8], especially in cases of odd-meter and shifting metric levels. It is, however, simple to find the average pace at which percussive events occur in a passage of music. For the purposes of video pacing, this turns out to be an important and useful measure. To find the event pace, we compute the RMS amplitude of the audio at time t by applying Equation (1) to a 100 millisecond window, centered on time t, where n is the number of samples in the window and x is the amplitude of a single sample.

Figure 3, shows the RMS amplitude for the first 10 seconds of Michael Jackson's *Billie Jean*. An average of the peak distances (1025ms) yields the overall pace for the song (about 59 pulses per minute). By walking through the entire song, one can easily detect sections whose pacing gives the music a half-time feel. This works quite well in the Pop/Rock genre. More complex music styles require more sophisticated techniques [9].

4.3 Finding Lyrics

Speech-to-text on singing is an unsolved problem [6], due to the non-standard nature of the speech and the large amount of background noise (read: the musical accompaniment) present in the recording. In fact, many humans have great difficulty in performing this task[1]. For this reason, we concentrated on finding song lyrics in on-line lyric repositories.

There are many strategies for finding song lyrics from audio meta-data like artist, title, and album information. Using a general purpose search engine to find lyrics introduces difficulties. As an alternative, MusicStory uses Leos Lyrics, an online lyrics library. Specialized search engines allow a direct lookup from the meta-data.

Hint	*T-Frame Count*	*Time per Frame*
Slow ($\leq$ 55 bmp)	33.3 ms	90 frames
Medium (55–85 bmp)	25.0 ms	80 frames
Fast ($>$ 85 bpm)	16.6 ms	70 frames

Table 1: The frame duration time and transition frame count for each given *hint* from the RMS pace estimate. Each photo is displayed for 60 frames plus the transition times for that rate.

Direct search does not always work. For example, the song *Smells Like Teen Spirit* by Nirvana appeared first on the 1991 album Nevermind, then on numerous compilations. If MusicStory looks up *Smells Like Teen Spirit* lyrics from Nirvana's self-titled 2002 compilation, our lyrics database will return no results. Similarly, Tori Amos's cover of this song may not also return any results. When a search failure is encountered, the agent performs a roll back strategy, dropping the album from the query. If no results are returned again, the artist is dropped and only the song is queried. This ensures the most exact lyric match can be found. If the song is not in the database, we prompt for lyrics.

4.4 Making Movies

Lyrics and pace in hand, the MusicStory agent is now ready to build the music video. First, a stop-list of common words in the English language is removed from the lyrics (such as 'and,' 'if,' 'he,' 'the'). Images associated with the remaining salient words are retrieved from the desired public and/or personal media space. If a variety of resources is used, the agent simply shows an equal number of images per repository. This balancing of repositories has been shown to be an effective approach [7].

The current version of MusicStory creates a photo slide show of the lyrics set to the source music. Here, we use the hint from our RMS pace estimate.

From previous work, we find that images need to be visible for at least 900 ms for the viewer to be able to see and gather the image association [14, 7].

The agent adjusts each photo's duration and the dissolve transition speed between photos using the RMS pace hint. The initial version of MusicStory uses the hint to select one of three categories: slow, medium, and fast. Table 1 shows the average display time per frame and number of transition frames for each category. Each category has a preset slide duration and transition speed. The duration and transition times for each category follow common video direction practice [15]. The output video file contains the audio in the same or similar format as the source audio. The video can be encoded with any variable bit-rate suitable for JPEG based videos, we chose to use Microsoft's ImageVideo Codec v9, which was designed for this style of application.

5. TWO INSTALLATIONS

Our motivations for MusicStory centered on autonomous music video creation for new media devices. The hand held music/MP3 player has grown both in its ability and ubiquity. Stand alone players, now play DVD-like quality video. Integrated players, like cellphones, have large color displays also capable of video playback. The use case for these video enabled devices provides something to hold and look at, and not something to hide in a jacket pocked

[1] see `http://www.kissthisguy.com/` for examples

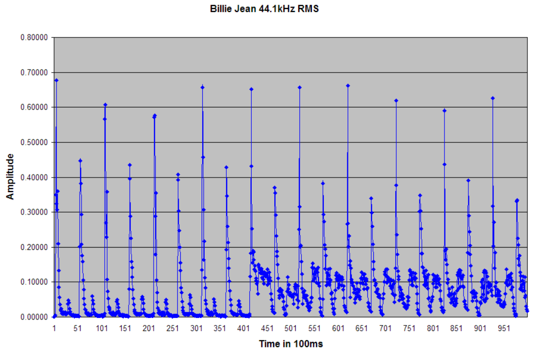


Figure 3: The RMS values for the intro to *Billie Jean*. An average of the peak distances (1025ms) yields an overall pace for the song.

during playback. MusicStory transforms the pure audio into a compelling multimedia experience.

Moving from the small to the silver screen, we deployed MusicStory in a large-scale, concert venue. We developed a five song set list which was presented as one of three acts at Wired Magazine's NextMusic show on June 22, 2005. Jeff Tweedy of the band Wilco, the NextMusic's curator, used MusicStory to create lyric based photo narratives and presented the music videos between two live acts.

6. FUTURE WORK

Currently, MusicStory connects our personal audio with visuals from public and personal media sources accessible through search engines (Google, Yahoo!, etc.). We are incorporating MusicStory with photo sharing websites, like Flickr.com, where tagged images are shared amongst a social network. There, personal music will be joined with personal images or the images from only their friends. Each music video will be unique to one's personal experiences and photos—creating one's own personalized music video soundtrack.

We are working on incorporating more sophisticated techniques for finding *hints*. Instead of using an overall multiple of the beat of the song, we wish to find where the song changes time or introduces an interlude or a break and have the agent direct the visual performance accordingly. Additionally, we will introduce work on song structure identification to find such things as verse/chorus boundaries [2] in the song to provide further direction. MusicStory could preform a call back to a prior image during a later chorus in the song.

We also plan to incorporate our existing work on affect detection to steer the direction of the music video. High affect lyrics can be matched to similar high affect images, which can be found from reading the affect in the image's captions, tags, and comments as well as image analysis.

Acknowledgments

The authors thank Victor Friedberg and Melanie Cornwell from NextFest & Wired Magazine, Jeff Tweedy from the band Wilco, and Microsoft for supporting MusicStory. In addition, continuing thanks to the guiding comments of Larry Birnbaum, Sanjay Sood, and other members of the Intelligent Information Laboratory at Northwestern University.

7. REFERENCES

[1] J. Budzik. *Information Access in Context: Experiences with the Watson System*. PhD thesis, Northwestern University, June 2003.

[2] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2003.

[3] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper & Row, New York, NY, USA, 1990.

[4] X. S. Hua, L. Lu, and H. J. Zhang. P-karaoke: personalized karaoke system. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 172–173, New York, NY, USA, 2004. ACM Press.

[5] W. Kandinksy. Point and line to plane: A contribution to the analysis of pictoral elements. In K. C. Lindsay and P. Vergo, editors, *Complete Writings on Art*, pages 528–699. Da Capo Press, USA, first edition, 1994 edition, 1926.

[6] M. Mellody, M. A. Bartsch, and G. H. Wakefield. Analysis of vowels in sung queries for a music information retrieval system. *Journal of Intelligent Information Systems*, 21(1):35–52, 2003.

[7] M. Mirapaul. Art unfolds in a search for keywords. *The New York Times*, page E5, 17 June 2004. Vol. CLIII, No. 52883, Circuits Section.

[8] B. Pardo. Tempo tracking with a single oscillator. In *ISMIR 2004, 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 10–14 2004.

[9] A. Pikrakis, I. Antonopoulos, and S. Theodoridis. Music meter & tempo tracking from audio. In *ISMIR 2004, 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 10–14 2004.

[10] M. Ruberry, S. Owsley, D. A. Shamma, J. Budzik, and C. Albrecht-Buehler. Affective behaviors for theatrical agents. In *IUI 2005 Workshop on Affective Interactions: The Computer in the Affective Loop*, San Diego, CA USA, January 2005.

[11] D. A. Shamma. *Network Arts: Defining Emotional Interaction in Media Arts and Information Retrieval*. PhD thesis, Northwestern University, Evanston, IL, USA, December 2005. (in prepairation).

[12] D. A. Shamma and K. J. Hammond. Imagination environment: Using the web as a source of popular culture. In *ACM SIGGRAPH 2004 Emerging Technology track*. ACM Press, 2004.

[13] D. A. Shamma, S. Owsley, S. Bradshaw, and K. J. Hammond. Using web frequency within multi-media exhibitions. In *Proceedings of the 12th International Conference on Multi-Media*. ACM Press, 2004.

[14] D. A. Shamma, S. Owsley, K. J. Hammond, S. Bradshaw, and J. Budzik. Network Arts: Exposing cultural reality. In *Alternate track papers & posters of the 13th international conference on World Wide Web*, pages 41–47. ACM Press, 2004.

[15] *The Simpsons*. DVD, 2002. Creator, director and producers notes from the season commentary.

[16] Nextfest.2005. *Wired Magazine*, 13(6):24, June 2005. The Wired World's Fair, June 24-26, Chicago. Insert Booklet.