Preprint for ACM Siggraph '06 Video Game Symposium Proceedings, Boston, 2006

Emotional Remapping of Music to Facial Animation

Steve DiPaola
Simon Fraser University
steve@dipaola.org

Ali Arya
Carleton University
arya@carleton.ca

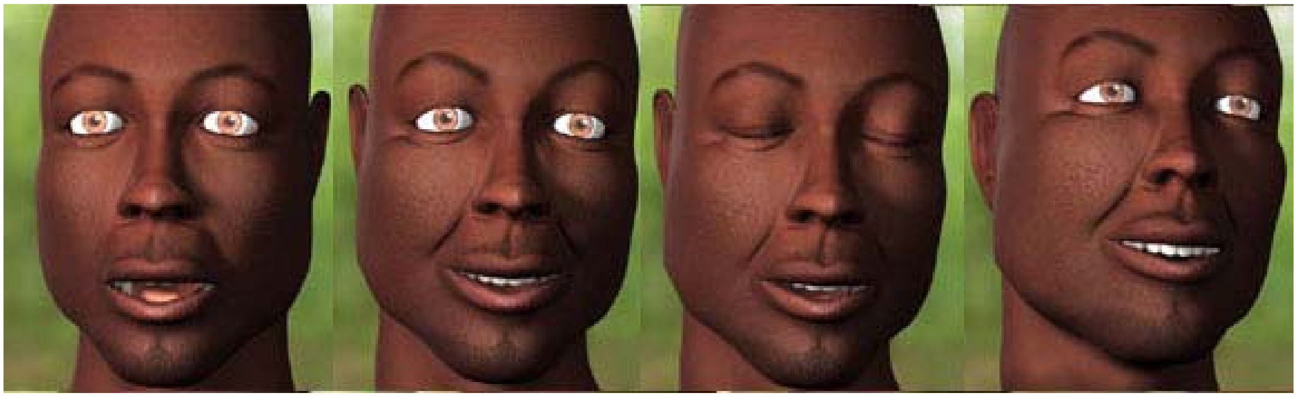


Figure 1. Stills from "“Concerto for Virtual Strings and Faces”" animation created by remapping affective data from a music score.

Abstract

We propose a method to extract the emotional data from a piece of music and then use that data via a remapping algorithm to automatically animate an emotional 3D face sequence. The method is based on studies of the emotional aspect of music and our parametric-based behavioral head model for face animation. We address the issue of affective communication remapping in general, i.e. translation of affective content (eg. emotions, and mood) from one communication form to another. We report on the results of our MusicFace system, which use these techniques to automatically create emotional facial animations from multi-instrument polyphonic music scores in MIDI format and a remapping rule set.

Keywords: affective communication, facial animation, data driven animation, procedural art.

1 Introduction

Human beings communicate their feelings, sensations, and ideas through a variety of channels. Such channels are, usually, designed to be aesthetically appealing to make the communication more attractive and pleasant. Although some might argue that artistic creativity, primarily, serves the creative needs of the artist, it is reasonable to say that, almost always, a motion picture, a painting, a piece of music, and any other aesthetic creation share a common theme that we call “affective communication”. They express and/or cause different emotional states [Levinson et al 1999]. Affective communication complements (and frequently overlaps with) “literal communication” which is essentially based on storytelling (through audio, visual, or textual means).

They both map content to certain elements and structures, but while the former is concerned with the expression of feelings and sensations, the latter is mainly related to describing events and making statements.

Although translation, as remapping some content from one form to another, has long been used for literal communication, its application to affective communication has been studied only recently, after the secrets of emotions themselves have been revealed by psychology and cognitive science [Ekman 2003, Levinson et al 1999]. A motion picture made based on a novel, and a document translated to a new language, are examples of the operation that remaps communication material from one domain to another. It is possible because the material (events, statements, etc) which were primarily mapped to structural elements of source communication medium, can later be remapped to a target medium. In other terms, this possibility is due to the fact that the content, communication media, and their relation are well-known.

Emotions and the way we express them through body/facial actions [Ekman 2003, Levinson et al 1999] and external media such as music [Juslin and Sloboda 2001, Krumhansl 2002] are the subject of a growing research. This increasing knowledge has also motivated “affective computing”, i.e. computer systems such as software agents that can work with affective processes, for instance recognize and express emotions. In this paper, we introduce the concept of “affective communication remapping”. Our knowledge about affective issues (emotions, moods, and the way they are expressed and perceived through different communication channels) enables us to extract affective information from source channels and express them in a target channel, as shown in Figure 2.

Emotional remapping is based on the idea that each affective communication consists of a medium-dependent form and a “general” or “abstract” affective or sentic form [Clynes 1992]. The remapping is done by relating affective forms of source medium to the affective forms of the target. Although some researchers have argued that these affective forms are independent of communication channels and remain unchanged [Clynes 1992], but due to the different nature of two media, the relation is not necessarily a one-to-one correspondence, and might need user interaction to select desired mapping. To demonstrate affective communication remapping, we have developed a music-driven emotionally expressive face animation system called MusicFace.

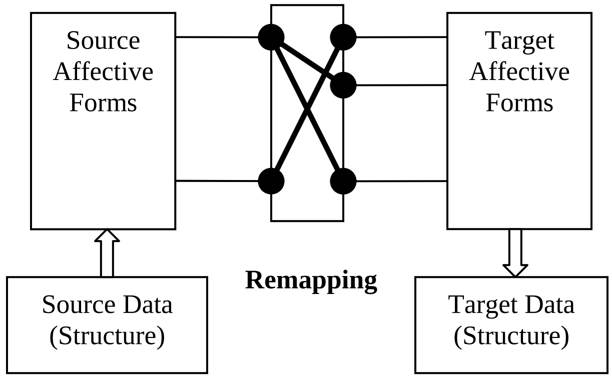


Figure 2. Affective Communication Remapping: Affective forms are extracted from source and then remapped to target (possibly different) forms.

We use emotional aspects of a given piece of music, provided in MIDI or audio data, to animate a face. Based on designer's interaction, the animation system can illustrate the emotional states expressed in music, simulate the possible emotional responses, or demonstrate a free "facial choreography" based on the piece of music. This is part of our Interactive Face Animation – Comprehensive Environment (iFACE) project that aims at developing a framework for "face objects" and "facespaces" [Arya and DiPaola 2004, DiPaola 2002]. Face object is a software component that represents human faces with a variety of functionality (talking, moving expressions, etc) through interfaces that can be accessed by users or other applications (web pages, GUI programs, etc). Facespace is a multi-dimensional space that holds different states and variations of a face object. iFACE allows programming face object and navigating through the facespace. Music-driven Emotionally Expressive Face (MusicFace) is a sample iFACE client that controls the face object and creates a "facial choreography" or "moving along a path in facespace" driven by musical input. In addition to artistic values, MusicFace can be used for creating visual effects in movies and animations, and also realistic characters in computer games and virtual worlds. But most important aspect of MusicFace is to provide an example and prototype for more complicated applications of affective communication remapping between media such as music, dance, animation, and painting.

In Section 2 some related work on music, emotions, and facial expressions are reviewed. Sections 3 to 5 explain the affective structure we have considered for music and face, and also our proposed remapping mechanism. In Sections 6 and 7 some experimental results and conclusions are presented.

2 Related Work

Although human emotions and affective behaviours have long been subjects of psychological and physiological studies [Darwin 1872, Ekman 2003, Levinson, Ponzetti and Jorgensen 1999], emotions and especially their facial expression are still active research topics [Ekman 2003]. During the 1980s, several models for emotional states and moods were proposed which were mainly based on two-dimensional mood classification. Among them Russell's circumplex model of emotion [Russell 1980] and Thayer's mood model [Thayer 1989] can be mentioned, illustrated in Figure 3. The former is based on activation and pleasure as two dimensions, and the latter energy and stress (two different types of activation). More than a hundred years after Charles Darwin proposed the idea of universal human expressions, Ekman and Friesen [1978] showed that basic human emotions and their facial expressions are universal among different cultures, and defined Facial Action Coding System (FACS) to represent major facial expressions and actions. In computer graphics and animation, FACS has been used to create emotionally expressive characters [Arya and DiPaola 2004], and acted as the conceptual foundation for Motion Picture Expert Group (MPEG) Face Animation Parameters (FAPs) in MPEG-4 standard [Ostermann 1998].

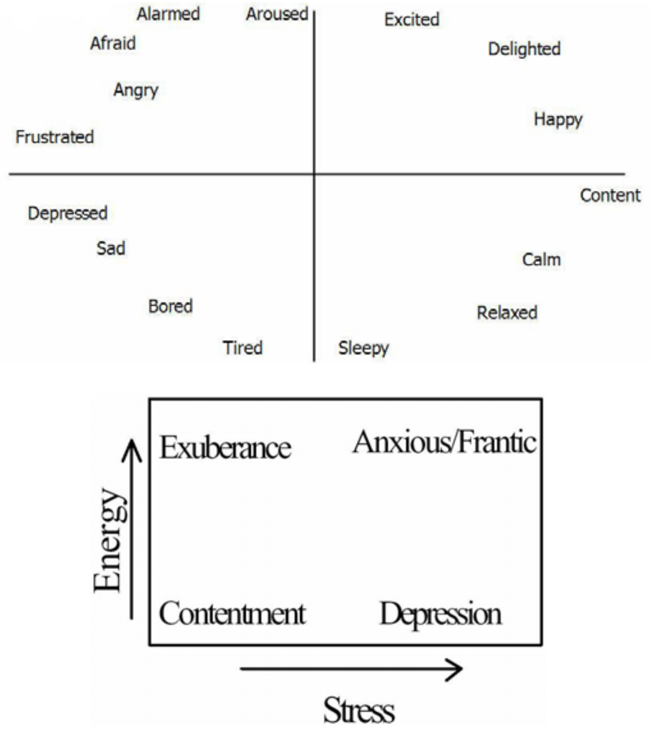


Figure 3. Emotion Models. Russell [1980] (top) and Thayer [1989] (bottom).

Emotional aspects of music have also been studied for a long time [Farnsworth 1958, Hevner 1936, Juslin and Sloboda 2001]. Despite the well-established body of knowledge about musical structures especially for western music [Lerdahl 2001, Temperley 2001], the effect of musical structure and performance on listeners' emotional response and perception is still a challenging research area [Juslin and Sloboda 2001]. Hevner [1936] grouped listeners' emotional response to pieces of music by asking them to write adjectives describing the music, and then clustering these adjectives. Fransworth [1958] refined and regrouped these adjectives into ten groups. Extracting audio features and classifying music based on these adjective groups is used to detect emotional states of music pieces [Li and Ogihara 2003]. Other researchers have used two dimensional mood models to achieve emotion detection in music. Liu et al. [2003] use music intensity to represent the energy dimension of Thayer model, and timbre and rhythm for stress. While psychologists and music therapist [Juslin and Sloboda 2001, Krumhansl 2002] have associated different structural aspects of music to certain emotions through experiments, some have focused more on performance-dependent aspects. Juslin [2000] considers tempo, sound level, frequency spectrum (as a representative of timbre), and articulation as performance cues and relate them to intended and perceived emotions. Table 1 summarizes some of the features extracted from music and their correlation with emotional state, as observed by Juslin [2000] and Bresin and Friberg [1999]. On the other hand, Clynes [1992] has proposed the idea of "sentic forms" as dynamic forms common to all emotional expressions (music, performance, etc), and used them for creating "emotional" music performances by machines (http://www.microsoundmusic.com).

Emotion	Music Feature	Value
Fear	Tempo Sound Level Articulation	Irregular Low Mostly non-legato
Anger	Tempo Sound Level Articulation	Very rapid Load Mostly non-legato
Happiness	Tempo Sound Level Articulation	Fast Moderate or load Airy

Table 1. Example Relation: Music Features and Emotions

Music Visualization has long been a topic of research in computer graphics [Cardle et al 2002, Kim et al 2003, Lytle 1990, Mitroo et al 1979]. Among the earliest and pioneering works, Mitroo et al. [1979] and Lytle [1990] can be mentioned who used musical attributes such as pitch, notes, chords, velocity, loudness, etc, to create color compositions and moving objects, respectively. Cardle et al. [Cardle et al 2002] have extracted musical features from MIDI and audio sources to modify motion of animated objects in an ad-hoc way. Pitch, level, chords, beat, and time durations are extracted from MIDI, and accompanied by power and bandwidth data extracted from audio signal to provide extra information not available in MIDI. Commercial products are now available in form of visualization plug-in for media player programs (http://www.winamp.com) or limited music-based animation tools (http://www.animusic.com).

With popularity of video games and web-based services, emotionally expressive software agents (characters) are another active area of research in computer graphics and multimedia systems. Valentine [1999] and DiPaola [2002], among others, have studied the concept of "facespace" as the set of all possible faces that can be created by changing spatial and temporal parameters governing geometry and behaviour. Cassell et al. [2001] propose a toolkit for suggesting non-verbal behaviour from a given text to be spoken. Arya and DiPaola [Arya and DiPaola 2004] introduced the concept of Face Multimedia Object that is based on a multi-dimensional head model, including a hierarchical geometry, and personality and mood meta-dimensions.

3 Extracting Emotional Information from Music

Affective information is embedded through a variety of structural elements in music [Juslin and Sloboda 2001]. Emotions expressed in a piece of music come from two different sources: composer and performer. The perceived emotional state also depends on listener's cultural and emotional context, and even physical conditions (not to mention environmental issues such as noise and audio quality). The emotional response of a listener (the change of state caused by music) depends on these factors as well, and is not necessarily the same as the perceived emotions. In this section, we discuss the major music features that are considered responsible for expressing emotions. We will show how MusicFace perceives emotional state of a given piece of music.

Based on existing studies, following musical features are extracted for affective information:

- Rhythm; Beats are detected as the peaks of he amplitude envelope made from the lowest sub-band (bass instrument such as drum). After detecting beats, the average tempo is calculated by dividing the total duration by the number of beats.
- Sound Level (Power); Signals root mean square (RMS) level for each sub-band and the sum of them are used.
- Timbre; Frequency spectrum is analyzed for timbre-related cues. High-frequency energy index [2000] is the simplest and most practical correlate of perceived timbre. It can be defined as the relative proportion of energy found above a certain cut-off frequency. Juslin [2000] has used 3 KHz for this cut-off value. Different values can be tried for optimal cut-off frequency as explained in Section 6. More detailed timbre features can be extracted from frequency spectrum for a better analysis, as shown by Liu et al. [2003].
- Articulation; Structure/phrasing of music refers to notes being smoothly connected (legato) or not (staccato). Two durations for each note can be measured in this regard (the average ratio will be an index of articulation):
    - From onset of a note to onset of the next note
    - From onset of any note to its offset
- Melody; Density and spread of each note (pitch) are the main melody-related feature.
- Tonality; Chords (harmony) used in the music and the key information (major/minor, etc) are the main tonality-related features.
- Duration; For each note attack, sustain, release, and decay times are measured.

The feature extraction mainly uses MIDI but audio data is also used for additional information. This is due to two reasons:

- The sound from a synthesizer is a function of MIDI input and the program resident inside the instrument, so MIDI data does not include all information.
- Performance-dependent variations in power, timing, and frequency spectrum do not show in musical score such as MIDI.

Liu et al. [2003] hierarchical method is used primarily to detect mood of the music according to Thayer model [Thayer 1989]. Intensity is used first to classify the music as content/depressed or exuberant/anxious. The timbre and rhythm are used to detect the mood category within each group. Cues suggested by Juslin [2000] and Bresin and Friberg [1999], and also Russell's two dimensional mood classification [Russell 1980] are then used to break down four mood categories of Thayer model into more detailed emotional states, similar to those in Russell's circumplex model, including:

- Contentment: content, satisfied, pleased, happy
- Depression: distressed, annoyed, frustrated, miserable, sad, depressed, bored
- Anxious: excited, astonished, aroused, afraid, angry
- Exuberance: sleepy, tired

A fuzzy rule base is used with extracted features as input and moods as output. In cases when fuzzy membership value to a mood class is not sufficiently higher than others, more than one mood will be selected with different weights. Figure 4 illustrates the extraction of affective information in MusicFace system. In addition to detection of high-level moods, music features are also used for controlling other affective gestures such as head movement and blinking, in an ad-hoc interactive way defined by the animation designer. This process is explained in more details in Section 5.

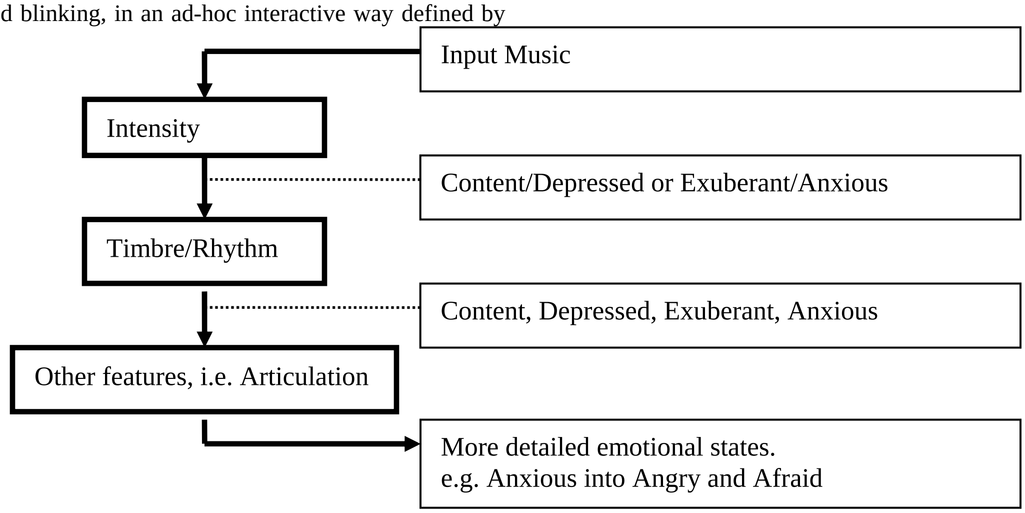

Input Music
Intensity
Content/Depressed or Exuberant/Anxious
Timbre/Rhythm
Content, Depressed, Exuberant, Anxious
Other features, i.e. Articulation
More detailed emotional states.
e.g. Anxious into Angry and Afraid

Figure 4. Mood Detection

4 Affective Behavior In iFace

iFACE system is based on the concept of "communicative face". For a large group of applications, facial presentations can be considered a means of communication. A "communicative face" relies and focuses on those aspects of facial actions and features that help to effectively communicate a message. We believe that the communicative behavior of a face can be considered to be determined by the following factors:

- Geometry: Creating and animating different faces and face-types are done by manipulating the geometry that can be defined using 2D and/or 3D data (i.e. pixels and vertices).
- Knowledge: Behavioral rules, stimulus-response association, and required actions are encapsulated into Knowledge. In the simplest case, this can be the sequence of actions that a face animation character has to follow. In more complicated cases, knowledge can be all the behavioral rules that an interactive character learns and uses.
- Personality: Long-term modes of behavior and characteristics of an individual are encapsulated in Personality.
- Mood: Certain individual characteristics are transient results of external events and physical situation and needs. These emotions (e.g. happiness and sadness) and sensations (e.g. fatigue) may not last for a long time, but will have considerable effect on the behavior. Mood of a person can even overcome his/her personality for a short period of time.

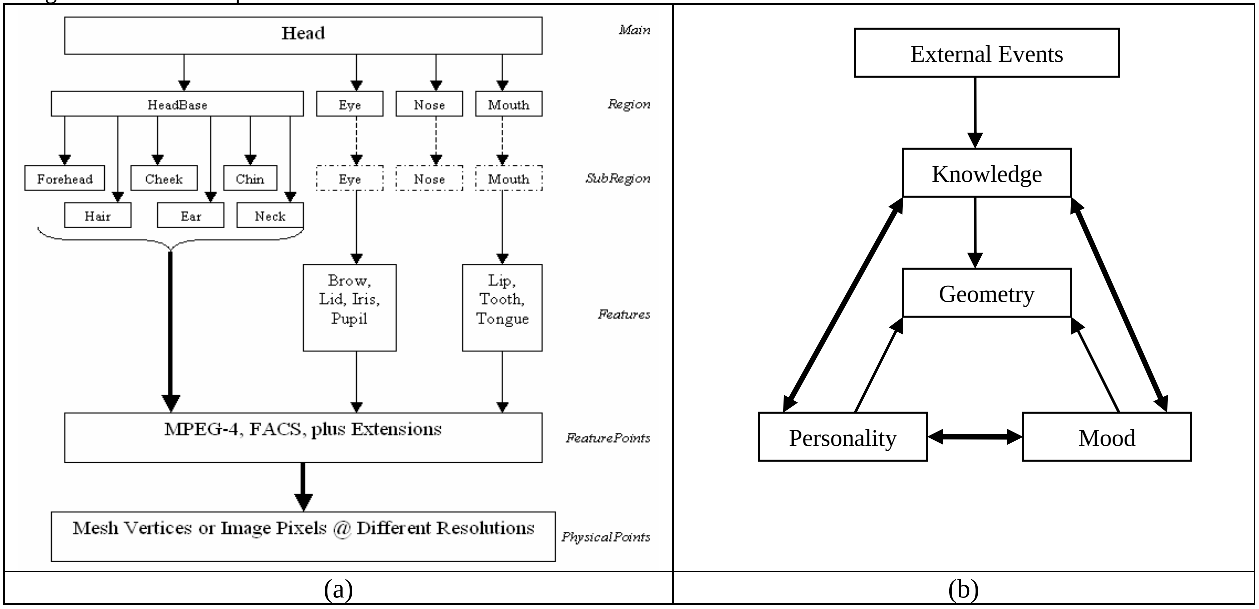
The geometry allows different levels of detailed control over various regions of head/face, through an object model. This object model is designed to allow users and client applications to access the underlying head data (3D or 2D) at different layers of abstraction (Figure 5a). This provides a multi-dimensional parameter space where objects at each level expose properly abstracted parameter sets. For instance, the Head object provides parameters for controlling the head as a whole without the need to know about or work with details of facial regions and features. Every object can be an aggregation of lower level (child) objects which are accessible through the higher level (parent) and provide more detailed parameters. Defining lower level objects is done only where and when necessary. For example, the model can include and expose detailed parameters only for the mouth area.

Knowledge meta-dimension encapsulates all the information required in order to define the requested/expected "actions," such as talking, movements, and expressions. Following the model in expert systems, we represent the knowledge in form of a rule base where the inputs are external events and outputs are functionality invoked in geometric objects. Following is a simple example of the interaction between Knowledge and Geometry. Here, Knowledge represents the following script written in Face Modeling Language (FML) [Arya and DiPaola 2004]. FML is designed exclusively for face animation and used in our system.

```
<action>
   <seq> <!-- sequential -->
      <hdmv dir="1" val="30" />
      <talk>Hello World</talk>
   </seq>
</action>
```

If Knowledge specifies "what" characters do, Personality and Mood help determine "how" they do it. In simplest cases, they act as modifiers for the rules in Knowledge. For instance, if the default response to the external event of "being greeted by another character" is to say "hello", the shy personality may only nod and the talkative one may say "hello, how are you". The main reason for having two meta-dimension for such individualization is to separate the effects of long-term and short-term parameters.

Personality, and especially Mood, can also affect Geometry in a more direct way. For example, an external event can cause Knowledge rule base to alter the Mood which in turn accesses and changes the Geometry by showing new facial expressions. Figure 5b shows our meta-dimensions and their relation to each other and the external events. The moods are defined based on Thayer and Russell models explained earlier, with energy and valence as parameters, and their facial displays are according to FACS. Common animation practices and ad-hoc methods, and also some visual cues suggested by psychologists have been used in iFACE to create personalities such as nervous, shy, assertive, and perky. This is done by considering movements of eye, brows, lips, and general head as the parameters.



(a) (b)

Figure 5. Parameterized Head Model

5 Displaying an Affective Face

As mentioned in Section 3, affective information is grouped into following categories:

- High-level group made of relatively long-term emotional states, i.e. moods
- Low-level group made of affective features extracted from source structural elements

Affective communication remapping is based on applying moods to the target medium, and also applying rules of activation to activate low-level target forms. Moods can be applied to the animated face based on three parameters:

- Strength of the mood is primarily calculated from the music piece but for the sake of smoothness in facial animation, the mood changes can be reduced.
- Detection time is the time the system spends on detecting moods. Longer detection times cause more stable moods.
- Transitions time is the time spent on a mood change. Longer transitions result in a smoother, and possibly more realistic, facial animation

Low level affective elements in source and target media can also be associated to each other. A typical example is rhythmic head movement related to music beat. Unlike moods, such association is not exclusively defined and in MusicFace will be controlled by user (animation designer). Each detected music feature (as described in Section 3) can be associated with facial actions in form of general expressions, eye and brow movement (especially blinking), 3D head movement, and lips movement (in general all MPEG-4 FAPs). This can be done in a periodic or one-time way.

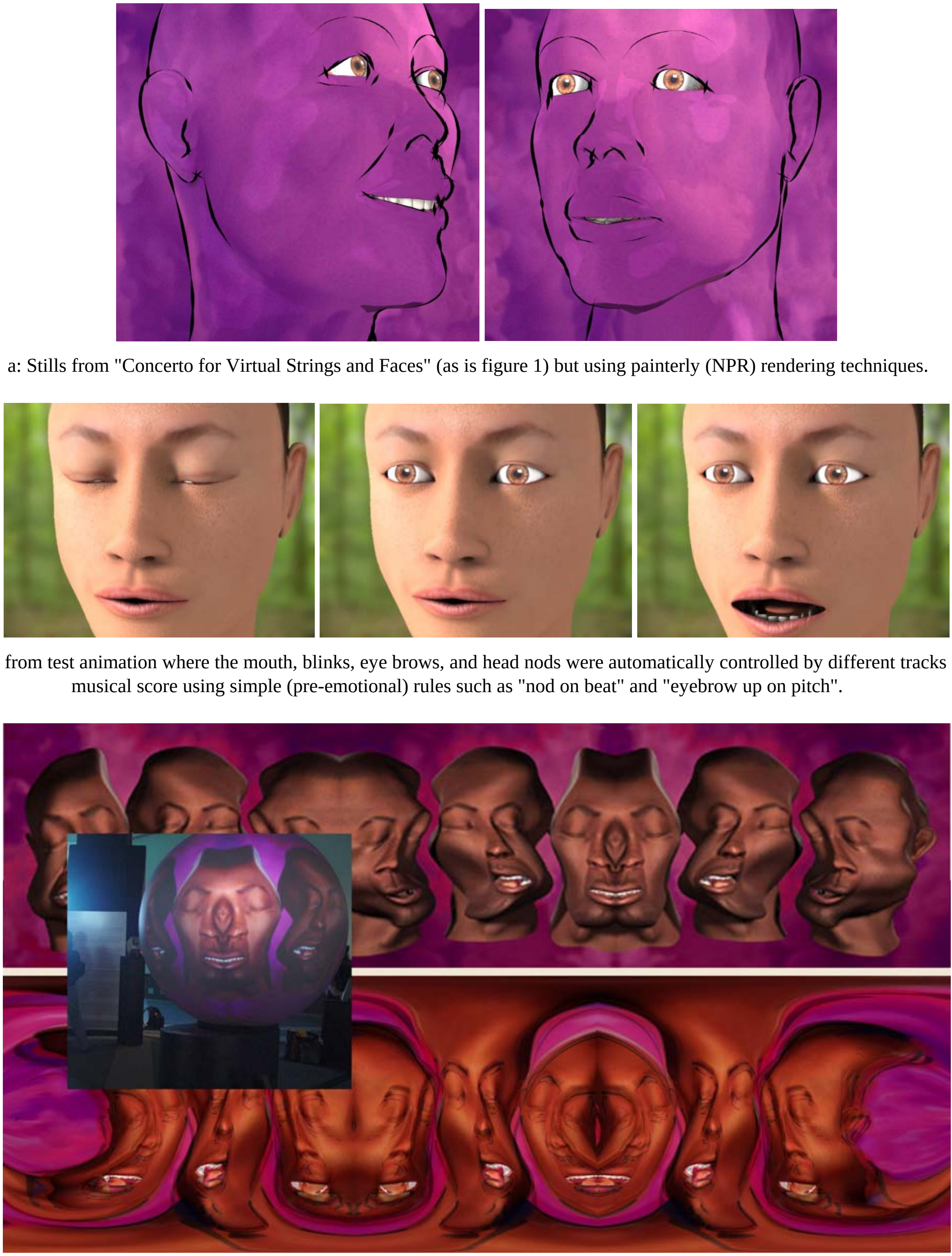
6 Results

Figure 1 and 6 show some typical facial actions driven by music features, generated by MusicFace system taken from our simple test music score (Figure 6b), from the multimedia piece “Concerto for Virtual Strings and Faces” (Figure 1 and 6a) and from the art installation piece "insideOut" (Figure 6b). MusicFace has been used in several art animation based projects including Vancouver’s NewForms Festival and in New York City’s A.I.R. Gallery in a piece called “insideOut” which was displayed on a 6 foot projected globe. In both cases a simple rule set was created by the artist. All animation was generated by the MusicFace system via the music in MIDI format and a rule set. A subset of rules discussed in this paper were used, but were musically associated with beat, volume, note length, melody and tonality. One rule created by the artist had the face close it’s eyes and lean back contemplatively for a length of time when the score’s note pitches were higher than average for a short sustained sequence and were generally harmonic. This rule would fire whenever this musical condition occurred additively with other rules creating a very natural mapping from music to face animation.

7 Conclusion

MusicFace starts by receiving the input music in MIDI format (or Audio format that provides extra information about the music). Structure and expressive cues such as tempo, loudness, and timing are combined with emotion color cues from extracting harmonic tension and rhythm, which then can be translated into emotional states based on the associations defined by our observations and other existing studies. A fuzzy rule-based system, using a two dimensional mood model, is responsible for deciding on facial emotions and actions (e.g. blinking). The requested facial actions are then sent to iFACE framework components in order to create the visual effects.

To create visual content, MusicFace mainly uses a parameterized 3D head model, provided by iFACE framework. This model has a multi-layer parameter space that allows control of the head movements and facial actions at low (vertex), medium (facial feature), and high (feature group) levels, combined with parameters for mood and personality. Extensions to MusicFace (and iFACE) involves non-photo-realistic animation or “moving paintings” where the emotions are applied to a painting rather than a photograph or computer model (Figure 6a) as well as additional driving time-based facial shape distortions and color/image effects of the animation.



a: Stills from "Concerto for Virtual Strings and Faces" (as is figure 1) but using painterly (NPR) rendering techniques.

b: Stills from test animation where the mouth, blinks, eye brows, and head nods were automatically controlled by different tracks of the musical score using simple (pre-emotional) rules such as "nod on beat" and "eyebrow up on pitch".

c: Stills (top and bottom) of our second generation MusicFace system where 3D shape distortion and color/image effects were also automatically driven by the music score's emotional content via the remapping rules. This piece was shown in art galleries in NYC and LA on a 6 foot internally projected sphere (inset).

Figure 6. Stills from different test and final animations created automatically by a musical score and remapping rule set.

8 Future Directions

The current system is modular, allowing any of the emotional music models discussed in this paper to be tested and compared. Our future work is to test and compare these models in a more systematic way. The remapping subsystem is also modular (shown in Figure 2), making it is possible to use another output system such as a character animation system or an abstract 2D painting system. Other output (or input) systems would need to be knowledgeable of a level of emotional parameterization like we have built into iFace or our music extraction system to be mapped efficiently. We are interested in working with other researchers and artists on other emotional remapping input/output scenarios.

We have worked with track-based MIDI data for the music scores because the remapping to date needs full symbolic representation (key structure, chord information, etc.) to be fully emotionally utilized. While we have discussed in this paper techniques to use audio data alone, it is still a significant research problem to extract separate note/track/instrument information from an audio source to derive significant structural and emotional content. To date even reliable beat extraction from a complicated audio source is still a non trivial problem. We have begun researching analyzing audio and MIDI data together which benefits by giving us both full symbolic data from the MIDI combined with subtle dynamic/performance data from the audio.

In application areas, we are working with a theatre designer, to research and develop a system using MusicFace that can drive the visual set design based on live music, much like dancers are affected by the live music played slightly differently each night. It would emotionally drive live animation that is projected onto the theatre walls and screens. Additional being able to use music on an affective level to control portions of real-time or offline animation has uses in several fields which could benefit from more complicated and dynamic visuals in ambient or secondary tracks, such as it gaming. This type of animation correlated to the music and emotional element of a scene could either effect gaming animation parameters in real-time to make a character more emotional syncopated with the background music/mood, or drive fully background elements with subtle animation.

References

Arya, A. and DiPaola, S. 2004. Face As A Multimedia Object, *Intl Workshop on Image Analysis for Multimedia Interactive Services*.

Bresin, R. and Friberg, A. 1999. Synthesis and Decoding of Emotionally Expressive Music Performance, *IEEE International Conference on Systems, Man, and Cybernetics*, Tokyo, Japan,

Cardle, M., Barthe, L., Brooks, S. and Robinson, P. 2002. Music-driven Motion Editing, *Eurographics-02*, UK.

Cassell, J., Vilhjálmsson, H. and Bickmore, T. 2001. BEAT: the Behaviour Expression Animation Toolkit, *ACM*.

Clynes, M. 1992. Time-Forms, Nature's Generators and Communicators of Emotion, *IEEE International Workshop on Robot and Human Communication*, Tokyo, Japan.

Darwin, C. 1872. *Expression of the Emotions in Men and Animals*, John Murray, London.

DiPaola, S. 2002. FaceSpace: A Facial Spatial-Domain Toolkit, *Sixth International Conference on Information Visualisation.*

Ekman, P. and Friesen, W.V. 1978. *Facial Action Coding System*, Consulting Psychologists Press Inc.

Ekman, P. 2003. *Emotions Revealed*, Henry Holt and Company, New York.

Farnsworth, P. 1958. *The social psychology of music*. The Dryden Press.

Hevner, K. 1936. Experimental studies of the elements of expression in music. *American Journal of Psychology*.

Juslin, P.N. AND Sloboda, J.A. 2001. *Music and Emotion: Theory and Research*, Oxford University Press, New York.

Juslin, P.N. 2000. Cue Utilization in Communication of Emotion in Music Performance: Relating Performance to Perception, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 6, pp 1797-1813.

Kim, T., Park, S. and Shin S. 2003. Rhythmic-Motion Synthesis Based on Motion-Beat Analysis, *ACM Trans Computer Graphics*, July.

Krumhansl, C.L. 2002. Music: A Link Between Cognition and Emotion, *Current Directions in Psychological Science*, vol. 11, no. 2, pp 45-50, April.

Lerdahl, F. 2001. *Tonal Pitch Space*, Oxford University Press, New York.

Levinson, D., Ponzetti J. and Jorgensen, P. 1999. eds. *Encyclopedia of Human Emotions. Vol. 1 and 2*. Simon & Schuster.

Li, T. AND Ogihara, M. 2003. Detecting Emotion in Music, ISMIR-03

Liu, D., Lu, L. and Zhang, H. 2003. Automatic Mood Detection from Acoustic Music Data, *ISMIR-03*

Lytle, W. 1990. Driving Computer Graphics Animation from a Musical Score, *Scientific Excellence in Supercomputing: The IBM 1990 Contest Prize Papers*.

Mitroo, J., Herman, N. and Badler, N. 1979. *Movies From Music: Visualizing Musical Compositions, ACM SIGGRAPH-79*.

Ostermann, J. 1998. Animation of Synthetic Faces in MPEG-4, *Computer Animation Conference*.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161-1178.

Temperley, D. 2001. *The Cognition of Basic Musical Structures*, MIT Press.

Thayer, R.E. , 1989. *The Biopsychology of Mood and Arousal*, Oxford University Press, New York.

Valentine, T. 1999. Face-Space Models of Face Recognition. *In Computational, geometric, and process perspectives on facial cognition: Contexts and challenges.* Wenger, M. J. & Townsend, J. T. (Eds.), Lawrence Erlbaum Associates Inc.