2009

Common Emotion Modeling in Distinct Medium Analysis and Matching

Amy Cho
San Jose State University

Common Emotion Modeling in Distinct Medium Analysis and Matching

A Writing Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Amy G. Cho

December 2009

Abstract

With the ever growing amount of digital information and multimedia on the World Wide Web and the current trend towards personalizing technology, users find themselves wanting a more intuitive way of finding related information, and not just any information but relevant information that is personal to them. One way to personalize and filter the information is by extracting the mood affectation, allowing the user to search based on current mood. The artificial intelligence field has done extensive research and continues to discover and improve current mood extraction techniques for each distinct medium. This paper will explore how to link and integrate the mood extraction of several distinct mediums—audio, image, and text—by utilizing a common emotion model that is customizable to the user. This project will allow the user to provide an input medium and find a matching output of a different medium based on default settings or user customization.

Acknowledgements

The following acknowledgement goes to the people involved in my academic efforts: endless thanks to Dr. Robert Chun of San José State University for his suggestions, enthusiasm, patience, and encouragement while overseeing this project. Many thanks to Dr. Chris Tseng for being a part of my thesis committee and for his input. Plenty of thanks to Mr. Frank Butt for his support, suggestions, and follow-up the entire time I was trying to balance school and work.

The following acknowledgement goes to my family: my gratitude goes to my brother Alvin Cho for his unwavering faith and support and always putting up with me. He always provides me the fortitude to push forward. My gratitude goes to my parents Yui Tsing Cho and Woon Gee Cho for their encouragement and wisdom.

The following acknowledgement goes to the people at work: again thank you to Frank Butt for his guidance. Many thanks to my manager Demai Ni for his understanding nature and working around my overwhelming schedule. Thanks to my team lead Cherri Vidmar for her patience, guidance, and knowledge, and thanks to my co-workers Lois Desplat and Vandana Goyal who helped pick up where I left off.

The following acknowledgement goes to the survey participants: thank you all for providing your valuable feedback which helped make this project complete.

Thank you, everyone.

Dedication

To my big brother Alvin K. Cho, for always encouraging me to think big, to think fun, to think the impossible is possible, and then to make things happen.

Table of Contents

LIST OF FIGURES

1.0 Introduction

The field of study in artificial intelligence is indeed broad and attempts to either create an intelligent machine or mimic human beings' intelligence. What makes up a human being is not just his knowledge but how he perceives the knowledge he obtains, the objects he encounters, and the people he meets. The perception of those things are usually associated with an emotion.

When mimicking human intelligence the study may involve natural language parsing or mood perception and conveyance. This concept of mood or affect extraction is of great interest since mankind envisions a more interactive relationship between humans and robots/machines. The mood extraction field mainly understands and mimics human beings by quantifying what is normally and intuitively thought by human beings to be unquantifiable such as personality, moods, and emotions.

Mood extraction can be done by analyzing a human being directly by observing the facial expression and body language. Another way to is to analyze them indirectly by understanding what human beings produce: art, music, novels, movies, etc. There has been extensive interdisciplinary research into the extraction of moods from these various mediums. For the most part, mood extraction is based on human patterns that appear as quantifiable patterns in the medium. For example, in a work of art, the dominant use of a color such as red can invoke a certain emotion such as anger or love. Yet another example, would be music: the frequent high pitch could mean tension and fear. Mediums such as movies, songs, and text at a glance may seem disparate but actually share objects and traits that invoke emotions and moods.

By being able to extract moods automatically, the copious amounts of information and objects available in the world and Internet can be sifted, CEMT and narrowed better because more relevant and pertinent pieces of information are available.

1.1 Purpose

Ultimately, the purpose of this project is to be able to connect and integrate distinct mediums (audio, image, and text) by using a common model of mood representation and then find a match output for an input medium. The project involves the modification and integration of mood extraction methods from image and audio files. The analyses of the mood extractions not only result in numerical values but are also supplemented with textual values as well in order to find the closest match(es). The mood extractions can also be customized according to the user's personal preferences. With the implemented system, a user will be able to find the corresponding match(es) of an audio file from a finite set of image files based on the mood of the content and vice versa. The result set of such a match can also be CEMT and narrowed down by additional annotations and textual information. The core lies within the integration of all these analyses from different components based upon a common mood model.

2.0 Motivation

Human beings and computer machines are certainly very different from one another but the field of artificial intelligence is determined to blur the lines between the two entities. The ability to have emotions is a major capability of humans that machines lack. But can emotions be simulated by machines? How can machines attempt to simulate emotions? To answer these questions, one must first understand what are emotions and how emotions are portrayed by human beings. There have already been studies in the emotional states of human beings, more aptly called emotional intelligence. Many studies have categorized and classified emotions into various models. Now knowing those states, how can those emotions be conveyed? Human beings show their emotions physically with their facial expressions and body language. Another expressive way is through what human beings produce: art, music, stories, and movies.

There has been extensive research and utilization of certain emotion models for mood extraction in images. Art consists mainly of texture, shapes, and color. When interpreting digital images, the method of determining texture is its own research field but does not express emotion as easily as other characteristics such as shapes and colors. Yet another whole field of study exists alone on object recognition in images and reality, which may not even include the emotion that is evoked by the shape. Then finally what remains is color as the main expression of emotion in digital images and is commonly used in research when extracting mood from images since it is most readily accessible. The method to extract mood based on color in an image will be the primary method described later in 4.0 Design & Implementation.

Music is fondly said to be a universal language and consists of many qualities: some are low-level features that are easily measured by machines such as frequency or pitch as well as other musical qualities that are measured by human beings such as melody, tempo, and rhythm. These characteristics of an audio file will be called audio features. The method to extract mood based on a set of audio features will be the primary method described later in 4.0 Design & Implementation.

Stories primarily consist of words when not accompanied by images and music, and words often act as a bridge between the other mediums of images, songs, and movies, meaning that people often describe an image, song, or movie in words; they are able to summarize and assign a mood in words to a non-textual medium. These mediums can all be connected via moods and words.

This research project excludes video and movies because adding videos would widen the scope of the project too much. To answer the question of whether it is possible to integrate the distinct mediums in order to allow distinct medium matching, it is not necessary to include videos when text, image, and audio are adequate enough. Besides, a video is comprised of many images synchronized to an audio file, so it is sufficient to research just the image, audio, and text.

What emotions and how intense the emotion varies from person to person when experiencing art, music, and stories. However, there are commonalities shared among people, cultures, and expressive mediums. This project will aim to utilize those commonalities and leaves room for the uncommon traits by allowing customizable rules into the system when interpreting the distinct mediums. For instance, in one culture the color red could represent passion and love more than it does in another culture where red could represent anger and aggression more.

The motivation to match distinct mediums initially all started on the thought: sometimes when a person listens to a song, he says,“That reminds me of this movie” or “That reminds me of this picture.” And the other way around when a person sees an image or movie and says, “That reminds me of this song.”

2.1 Potential Applications

The concept of mood extraction from various mediums alone has great potential because it can be utilized in commercial and non-commercial scenarios. In non-commercial settings, libraries of images, songs, movies, and books can be sorted, organized, and CEMT based upon mood. In commercial settings, it can also be used for the same purpose. Amazon.com, for instance, could show recommendations of products based on moods. Some examples of the concept has already been implemented: open-source software that allows you to create music playlists based upon mood and Moody for iTunes which allows the user to annotate the mood manually by marking the mood upon a colored version of the arousal-valence graph[32]**.**

To my knowledge, there has been little research about bridging the gap between distinct mediums by matching them. The concept of matching distinct mediums was originally based on my desire to assist the physically-disabled who visit websites. Although there are text-to-speech mechanisms that help the visually-impaired to hear the text of a web page and alternative captions of images intended for visually-impaired visitors, it would be useful if an audio file was matched to the image based on similar mood invocation so that the visitor does not lose out on the intended emotion of the image. For the hearing-impaired visitors, if a website is playing a song, then a matching image would be displayed instead so that the visitor can feel a similar emotion that the song would have evoked. If a website has many images or audio, it would be convenient for the website creator to have an automated solution of finding a matching medium so that the physically-disabled visitor can fully appreciate the website.

The concept of matching distinct mediums has other potential uses as well[19]. It can be utilized on a user's large library of images and songs when creating a slide show or presentation for family, friends, and peers during events such as weddings, birthdays, and other celebrations. The concept can also be used for blogs. When a user creates a new entry to his journal, the overall mood of the entry can be determined so that an image, emoticon or not, and song can be matched to the entry. It can also be utilized for social networks, recommendation systems, and commercial applications like

Amazon.com as mentioned above. There has been an incredible demand and growing trend from the marketing, advertising, and entertainment industry [1, 18] for such mood extraction technology because it will help these industries better understand and anticipate targeted audiences.

If this technology is extended and customized extensively to not only match files of different mediums but generate entirely new ones based upon an input, then it can also be utilized by novices who want to try their hand at another medium. For instance, music is not necessarily easy to learn so a person who wants to generate a song can draw an image and put that image input through the system and generate a unique song to that image and vice versa. In the commercial scenario, an advertisement team have a particular image they have in mind for a commercial, but the team needs a matching song or an entirely new song altogether[22]. This technology could potentially be able to complete projects like commercials and music videos much more quickly.

3.0 Background & Related Work

3.1 Mood & Emotions

In some papers mood and emotions may be considered distinctly different due to the temporal sense, but in this paper, mood and emotions are considered interchangeable[11,]. Research studies have used two categories of emotion modeling: textual listings and dimensional models. The more popular model is dimensional, specifically two-dimensional. A textual listing model would look something like Kate Hevners Adjective Circle [9] which displays eight groupings of related adjectives, see Figure 1. This model is used in the mood classification of photos in Chen et al's study [5]. This type of model allows generalization of several users' mood perception and allows mapping.

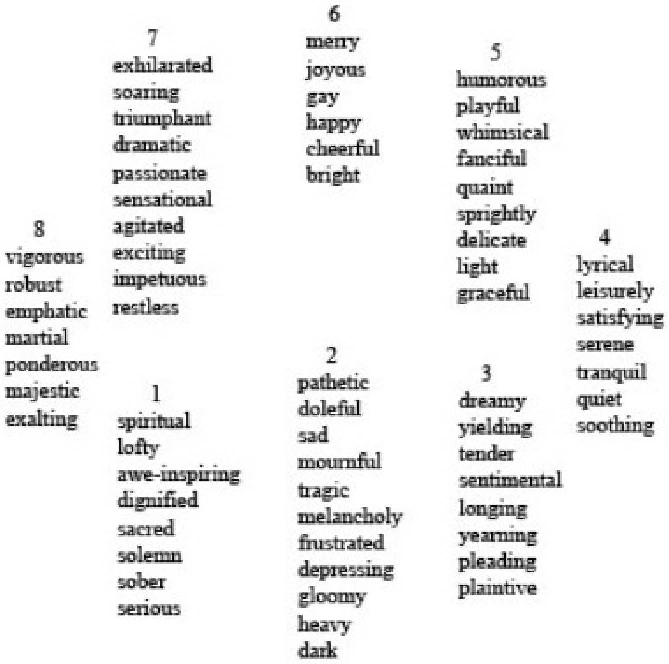


Figure 1: Kate Hevner's Adjective Circle[9]

Dimensional models may consist of three dimensions such as Wundt's mood model seen in Figure 2. The first axis represents pleasure and displeasure while the second axis represents arousal and non-arousal, and lastly the third axis indicates stress and relaxation. In other terms, these axes can be called “valence/arousal/dominance,” but the third axis dominance is usually not necessary because valence and arousal axes are enough to represent human emotions in most empirical studies [13] as seen in Thayer's mood model in Figure 3, which is a popular mood model. Representing mood on dimensional models allow more granularity of mood representation and better visualization and distribution of moods to a user.

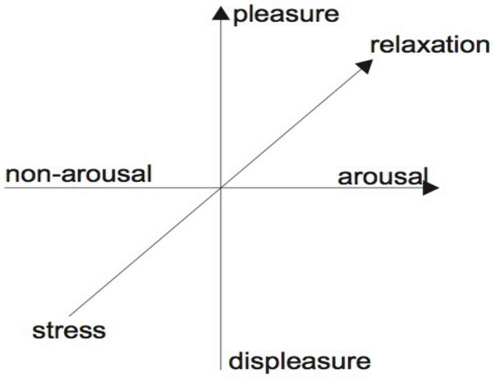


Figure 2: Wundt mood model[9]

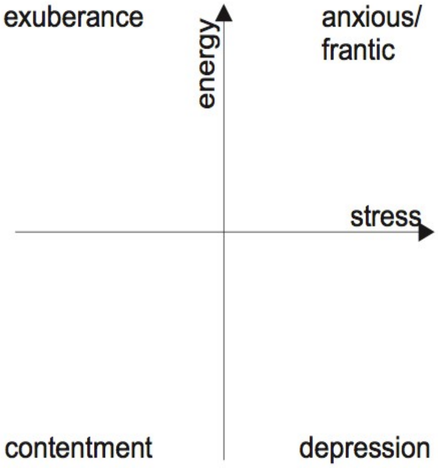


Figure 3: Thayer mood model[9]

3.2 Audio Analysis & Mood Extraction

As bandwidth and usage of the Internet increases, so does the use, sharing, and searching of multimedia. Audio files, usually consisting of music, may or may not be associated with metadata (tags, comments) and so to filter and search the vast multimedia library, automatic classification of audio is convenient. The typical procedure consists of determining the set of audio features that need to be extracted that help determine mood first, a classifier must be designed or chosen next, then a set of files must be used as a training set. The training set is manually annotated and classified to a certain class. The classifier is then trained based on the manually classified training set by extracting the determined set of necessary audio features. The necessary pre-processing is then completed in order to determine the class of untrained audio files. The untrained audio file would then be extracted for the audio features and fed through the classifier [10, 24, 25].

When analyzing audio files, certain aspects of the audio files called audio features are extracted. Each audio feature is correlated with a certain mood based on empirical studies [10, 16, 17]. Commonly extracted audio features are the following: Timbre Features, Shift Input, Fanout, Zero Crossings, Windowing, Power Spectrum, PowerSpectrumNet1, STFT Features, Centroid, Rolloff, Flux, MFCC, Spectrum2Chroma, SCF, SFM, Filter, LPC, LSP, LPCC, and Texture Stats. Popular classifiers include Gaussian Model (GM) and Support Vector Machine (SVM)[9].

3.3 Image Analysis & Mood Extraction

Similar to audio files, due to the enormous amount of image media available on the Internet, automatic classification of images is useful for searching and filtering relevant information since not all images have useful metadata. In some situations, no metadata is available even in the filename, surrounding webpage, or tags. Usually image metadata is not stored within the file like tags ID3 tags for music. When classifying images, the procedure is very similar to automatic audio classification: a set of relevant image features must be determined, the type of classifier must be determined, then a set of images must be selected as a training set. The training set is manually classified and the necessary image features are extracted and analyzed in order to build the classifier. Once the classifier is built, untrained images are then extracted for their features and fed through the classifier[8].

Images can consist of everyday objects, real scenes, and abstract objects. There are various features of images that can be extracted. Most common are color, shape, and texture. Color can be composed of hue, brightness, and lightness[7]. This trait can immediately set the mood. Certain combinations of colors is also a factor in setting mood[14]. Shapes such as real-life objects or symbols can be extracted. Texture such as the smoothness of the ocean or layout of farm crops can also be determined. Common classifiers include SVM[8].

3.4 Text Analysis & Mood Extraction

Stories, novels, and blogs all contain text, most likely attempting to express an emotion[12]. Songs with lyrics attempt to express an emotion as well. LyricAlly[21] works to synchronize the lyrics with the spoken words in the song but does not take emotions into account. When automating classification of music or images, including the metadata such as tags or lyrics improves the accuracy of the classification because emotion can be detected in text[2, 3, 6, 20, 23]. SentiWordNet[27] provides numeric values for each word with sentimental value. The emotion from a word can be manually annotated or it can be semi-automated by utilizing synonyms. Like image or audio classification, a set of items must be trained and a connection must be established between training set and untrained item. If the word "happy" is assigned numeric values indicating emotional value, then similar words such as "elated" can be assigned similar numeric values by deriving these values from the trained word "happy."

3.5 Related Work

3.5.1 Freeware

One closed-source freeware application called Moody[32] allows a user to annotate his songs based on mood. This software works en tandem with iTunes. As a user listens to a song in his iTunes library, he can use the Moody interface to annotate the current song with the mood simply by clicking on one of the colored squares. The interface consists of a 4x4 grid that represents various moods. Beginning from the bottom left corner, the mood changes from left to right (sad to happy) and the intensity changes from bottom to top (calm to intense). The annotation is then saved in the mp3 file's ID3-tag under "Comments" or "Composer." The value saved is a two-character value representing one of the squares in the grid. The rows are alphabetically named from top to bottom with A, B, C, and D. The columns are numerically identified as 1, 2, 3, and 4. For example, the top left, red corner would be represented as "A1," and to the user that square would represent "anger." This interface is intuitive and easy to understand to the user, see Figure 4.

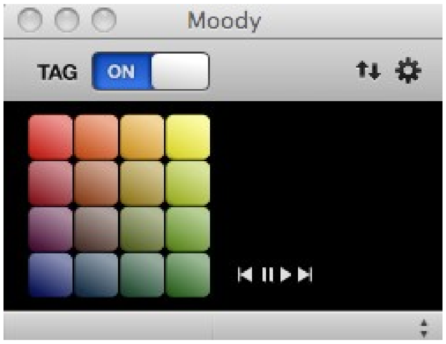


Figure 4: Moody interface

3.5.2 Research

In Dunker et al's work [9], they create an emotion model based that can be used by music and images but they do not take text into account. They focus solely on the content of the test files and not any associated metadata which is not realistic because there are many files available on the Internet that are tagged with useful information that can help further determine the mood. Also this method consists of retraining sets of data if preferences and interpretation of moods change. See Figure 5 for their process.

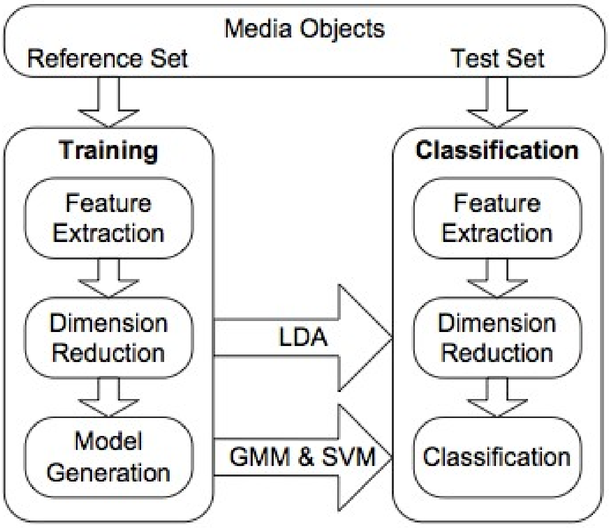


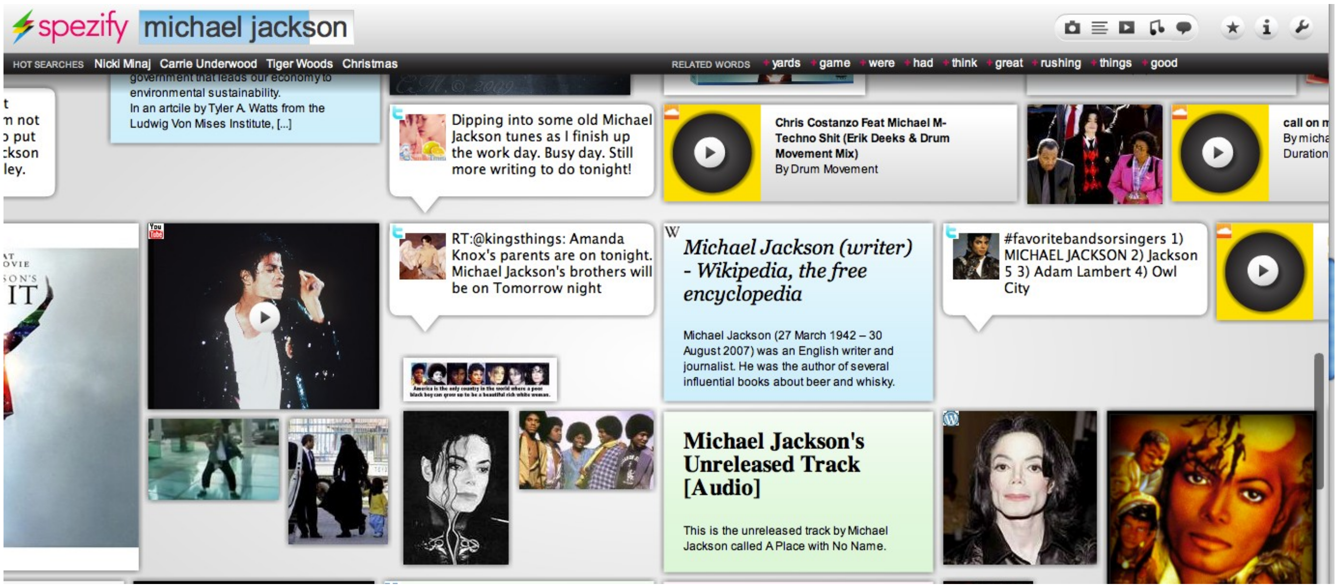
Figure 5: Dunker et al's framework model[9]

In Chen et al's work [5], they use the adjective circle based on Kate Hevner's Circle as an emotion model for image files. This paper also does not incorporate additional metadata for audio files and image files. They do not allow a quick user customization of the determined feature extractions.

In Kim et al's work[15], they propose the concept of creating image icons to represent certain categories of songs that fall under a particular mood. A user takes a survey and sets gives his preferences to establish the characteristics of an icon that should fit a particular mood. The attempt to create these icons have not been completed as indicated by the paper.

3.5.3 Commercial

The commercial sector of the Internet has several major search engines like Google.com and Yahoo.com that do not specialize in a specific medium search; they allow a text input and return various mediums based on a ranking algorithm; moods are not necessarily considered within the search process. Then there are specialized search engines like Microsoft's recently unveiled Bing 2.0 [26] that is coined as a “visual search engine” because a user inputs some text and mainly visual image output appears. Then there are even more specialized searches that focus on a specific user experience and medium searching. Spezify.com [33] is a website that focuses on user experience by organizing various mediums into an intuitive layout. The website allows textual input and returns various mediums that match the text input and allows the user to filter based on medium type as apparent in Figure 6.



government that leads our economy to environmental sustainability.
In an artcile by Tyler A. Watts from the Ludwig Von Mises Institute, [...]

Dipping into some old Michael Jackson tunes as I finish up the work day. Busy day. Still more writing to do tonight!

Chris Costanzo Feat Michael M- Techno Shit (Erik Deeks & Drum Movement Mix)
By Drum Movement

RT:@kingsthings: Amanda Knox's parents are on tonight. Michael Jackson's brothers will be on Tomorrow night

Michael Jackson (writer) - Wikipedia, the free encyclopedia

Michael Jackson (27 March 1942 – 30 August 2007) was an English writer and journalist. He was the author of several influential books about beer and whisky.

#favoritebandsorsingers 1) MICHAEL JACKSON 2) Jackson 5 3) Adam Lambert 4) Owl City

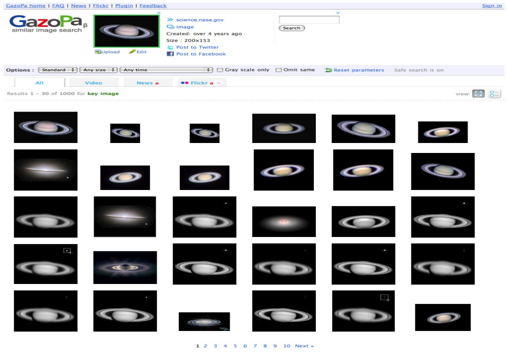
America is the only country in the world where a poor black boy can grow up to be a beautiful rich white woman.

Michael Jackson's Unreleased Track [Audio]

This is the unreleased track by Michael Jackson called A Place with No Name.

Figure 6: Spezify.com search results[33]

Gazopa.com [34] is a specialized search engine that focuses on image similarity as can be seen in Figure 7. The website allows an image as input and returns a set of images that are similar to the input and provides a similarity percentage for each image. These two websites do not allow other medium types such as audio and music to be used as an input.



Created: over 4 years ago
Size : 200x153
Options : Standard | Any size | Any time | □ Gray scale only | □ Omit same
Results 1 – 30 of 1000 for key image

Figure 7: Gazopa.com search results[34]

3.6 Software

A variety of technologies was utilized to complete this project. The initial audio analysis was primarily handled by Marsyas v0.2 [30] which is built on CMake 2.8. Slight modifications of Marsyas was handled by using XCode development environment. The core of this project was implemented as an application that sits on a GlassFish Enterprise Server v2.1 with Java SDK 1.5 and the graphical user interface was implemented with Java Server Pages, HTML, and CSS. The analysis provided by Marsyas was then processed and handled with Java and inserted into a Derby database. Image analysis was handled by Java code and the output was inserted into the Derby database as well. Much of the matching and filtering is done by the Derby database.

4.0 Design & Implementation

4.1 System Overview

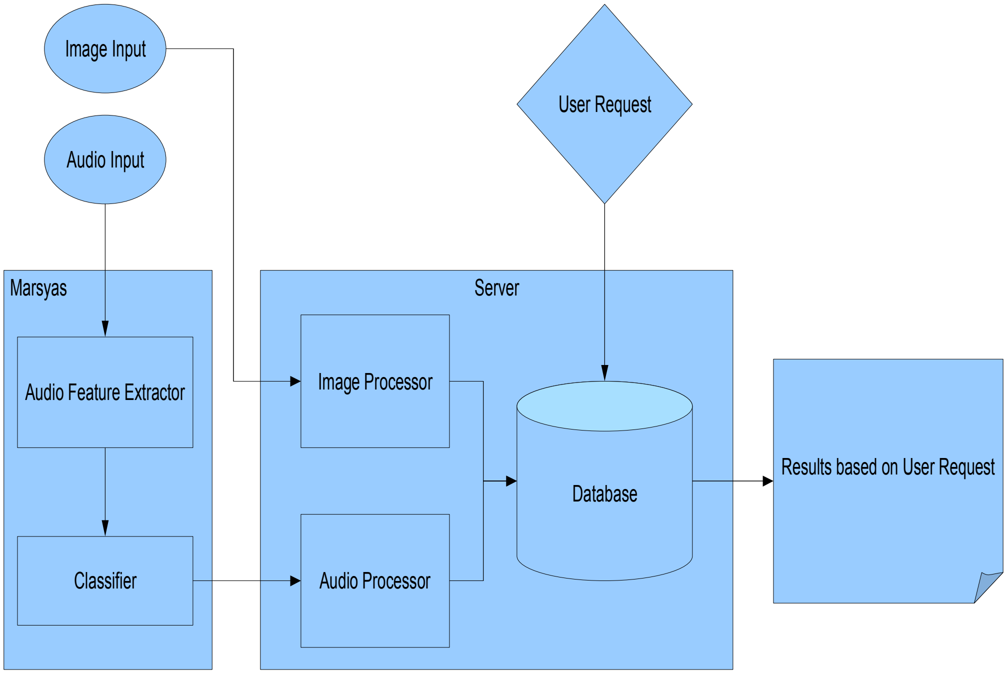


Figure 8: System Architecture

The project's entire system in Figure 8 consists of several major components, see Figure 8: data sets, pre-processors, rules, annotation, and medium matching. The system allows you to add new files, edit existing database values, and delete existing database values. Two major operations occur in this system: pre-processing and distinct medium matching. Pre-processing consists of processing audio and image files prior to any distinct medium matching. The pre-processor analyzes the files and creates textual and quantitative output that is stored in the database under the appropriate tables which will act as the data sets. Distinct medium matching consists of matching two distinct mediums based on the dominant mood whether is manually annotated or retrieved automatically by the system and filtering the matching results based on any textual information that is provided. The user is also allowed to edit and any existing rules that will be used later on in the medium matching prior to or after medium matching. The user is also allowed to edit the database values for each files associated metadata before or after medium matching to influence the matching results during the matching and filtering process.

4.1 Emotion Models

Emotions intuitively seem difficult to quantify but there are models available and they are commonly used by researchers in the psychology and artificial intelligence field. This project proposes a common 16-sector arousal-valence model to map out the emotions when comparing and integrating several distinct medium types.

4.1.1 Arousal-Valence Model

There are several emotion models available, some are three-dimensional, but most are of the two-dimension variation derived from Thayer's arousal-valence model[9]. The X-axis represents valence, the basic emotion while the Y-axis represents the arousal, strength of the emotion. (The third axis Z is usually dominance but emotions are represented well enough with just two dimensions as seen through empirical studies[13].) The bounding values of X and Y are -1 to 1 with (0,0) being the center. The four quadrants are normally associated with four broad emotions. Quadrant -X,+Y typically represents feelings of anger, fear, and disgust. Quadrant -X,-Y represents feelings of sadness and boredom. Quadrant +X,+Y represents feelings of happiness, pleasantness, and excitement, while quadrant +X,-Y represents calm and neutral feelings.

4.1.3 Hue Wheel

The hue wheel is a common mechanism for understanding colors in pictures, see Figure 9. The hue values are contained within the range from 0 to 360 degrees because it forms a circle. The basic colors of red, orange, yellow, green, blue, purple, and pink can be broken into bins or even more depending upon one's subjective needs. Black, white, and grey colors are not a part of the hue wheel since those colors are merely high or low saturations or lightness values of a certain hue, so the system's image processor does not handle lightness and saturation.

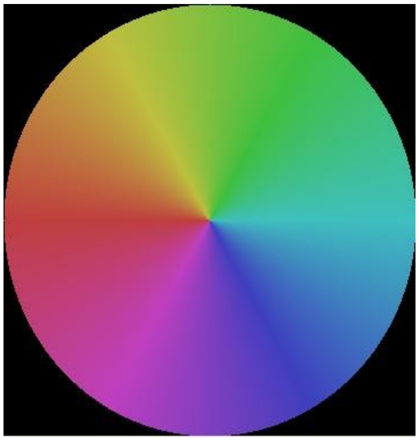
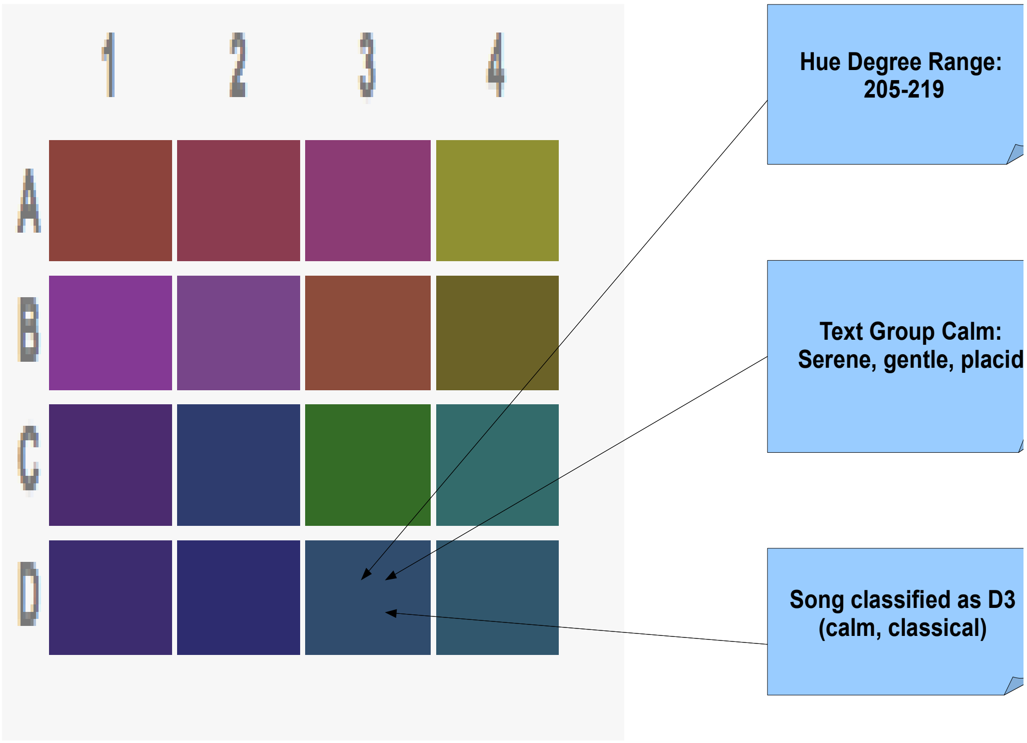


Figure 9: Hue Wheel[31]

4.1.4 Common Arousal-Valence Model

This paper introduces a 16-sector arousal-valence model as the common emotion model for matching distinct mediums and is the critical point to how matches are made initially. Thayer's valence-arousal model is broken up into 16 equally sized sectors, each quadrant has four sectors. Each sector is assigned a value to locate the sector. The vertical location is indicated by A, B, C, D from +Y to -Y. The horizontal location is indicated by 1, 2, 3, 4 from -X to +X. The top-left corner sector would be called “A1.” All audio, image colors, and text can be mapped to each sector according to rules saved in the database. For instance, angry music, angry text, angry images with a lot of red color could be mapped to sector “A1.” See Figure 10 to see how audio, text, and images can be mapped together on the 16-sector arousal-valence model. Any further mentioning of the arousal-valence model in this paper will refer to this common 16-sector arousal-valence model.



Hue Degree Range: 205-219

Text Group Calm: Serene, gentle, placid

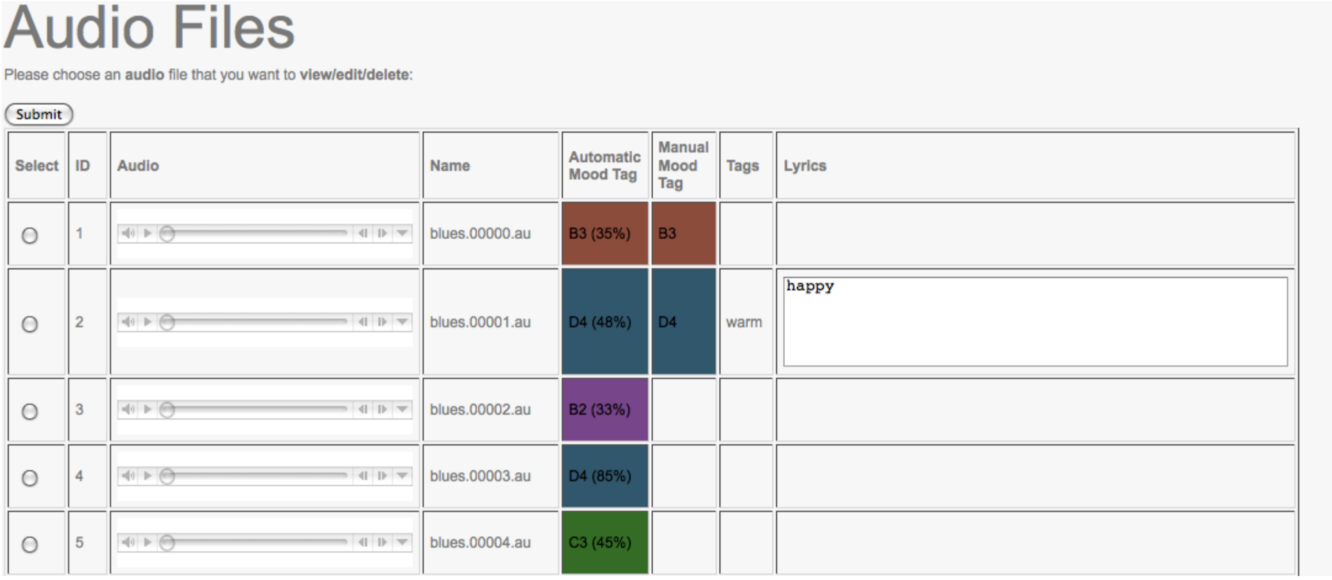
Song classified as D3 (calm, classical)

Figure 10: Mapping Values to Emotion Model

4.2 Data Collection

4.2.1 Audio Files

This project uses the audio files from this project [32]. The entire, original audio data set consists of 1,000 songs and 100 songs from each of the following 10 genres: blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, and rock. Each song is a 30-second clip in .AU format which is what the software Marsyas accepts because it is a highly structured audio format. This project does not utilize all 1,000 songs but merely a subset of 10 songs from each genre, totaling 100 songs total, which were randomly chosen. See Figure 11 for a sample of the data set.



Audio Files

Please choose an **audio** file that you want to **view/edit/delete**:

Select	ID	Audio	Name	Automatic Mood Tag	Manual Mood Tag	Tags	Lyrics
○	1		blues.00000.au	B3 (35%)	B3		
○	2		blues.00001.au	D4 (48%)	D4	warm	happy
○	3		blues.00002.au	B2 (33%)			
○	4		blues.00003.au	D4 (85%)			
○	5		blues.00004.au	C3 (45%)			

Figure 11: Audio Files Data Set

4.2.2 Image Files

The project uses the image files randomly taken from the Internet. The entire, original image data set consists of over 2,000 images and were categorized in the following subjects: coast, mountain, forest, highway, inside city, street, and tall building. Again, only a subset of these images were used for the survey. Images with obvious dominant hues were chosen to see if this project can handle the more basic test cases first. The data set consists of scenic images mainly. The more obvious files tended to be mountain and coast imagery. See Figure 12 for a sample of the data set.

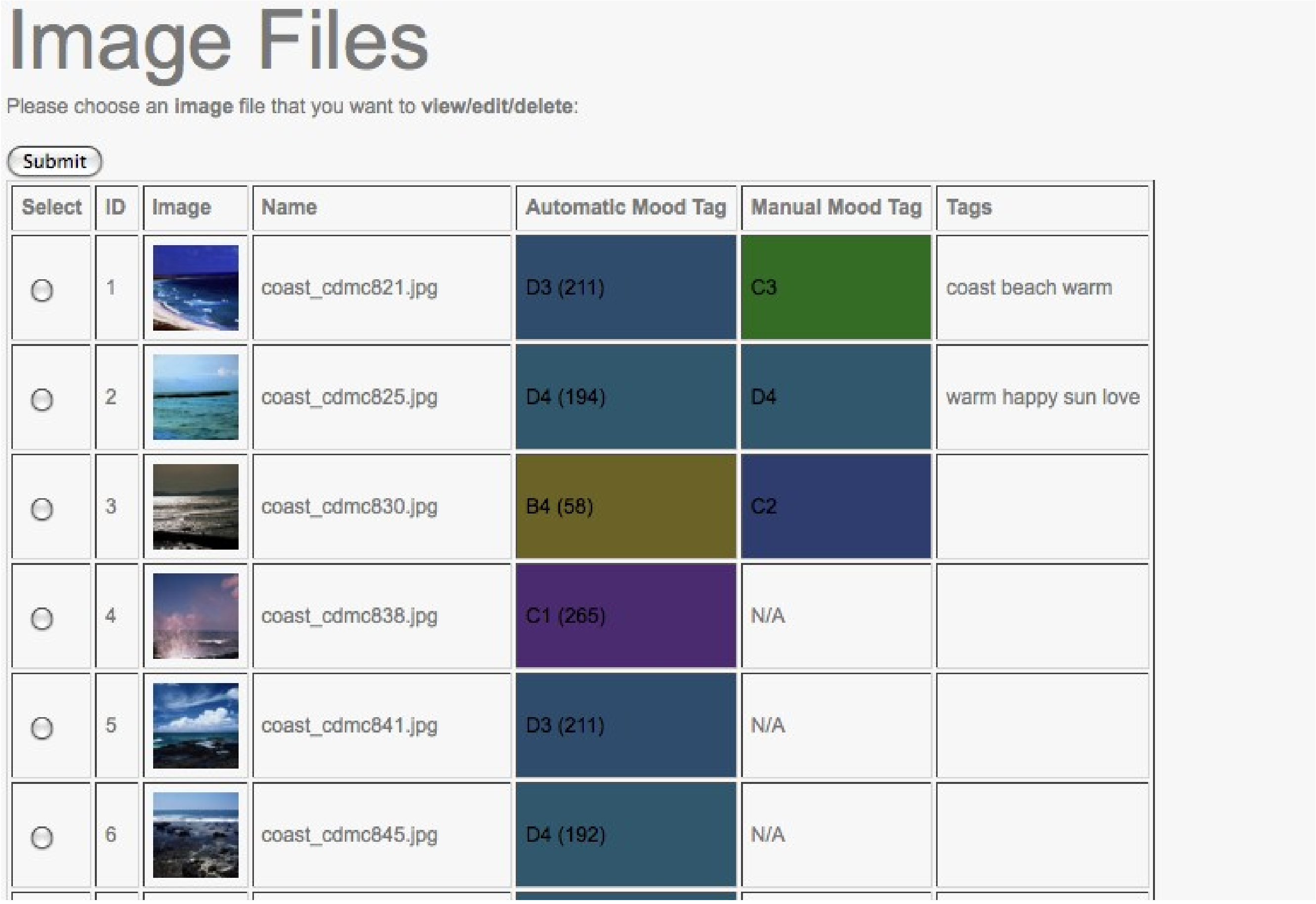


Image Files

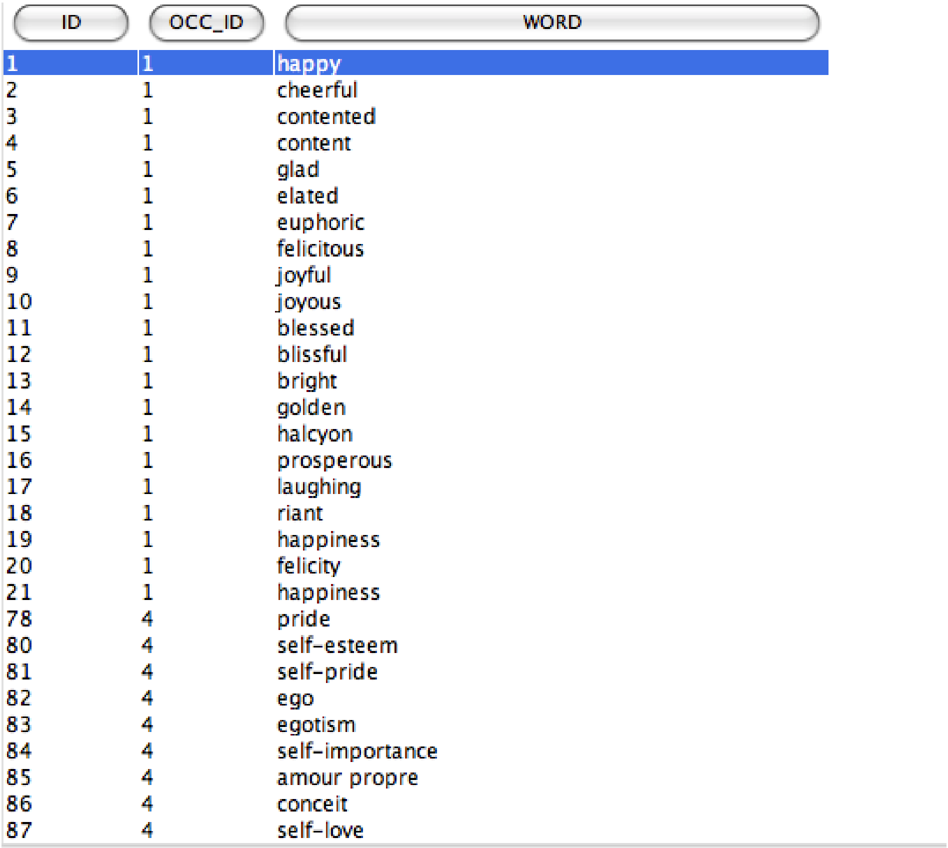
Please choose an **image** file that you want to **view/edit/delete**:

Select	ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
○	1		coast_cdmc821.jpg	D3 (211)	C3	coast beach warm
○	2		coast_cdmc825.jpg	D4 (194)	D4	warm happy sun love
○	3		coast_cdmc830.jpg	B4 (58)	C2	
○	4		coast_cdmc838.jpg	C1 (265)	N/A	
○	5		coast_cdmc841.jpg	D3 (211)	N/A	
○	6		coast_cdmc845.jpg	D4 (192)	N/A	

Figure 12: Image Files Data Set

4.2.3 Text

To anticipate the words with emotional values which tend to be adjectives that a user will input, a table filled with a small set of words. These values were grouped based on the 16 sectors. The Ekman word values were used as head values of the group. Then subgroups based on OCC words were created for each Ekman group to make mapping more granular for the 16 sector. For instance, the Ekman word “happy” is mapped to “A4” and will include words like “cheerful” and “joyful.” See Figure 13 in order to see how the OCC words are mapped to each sector.

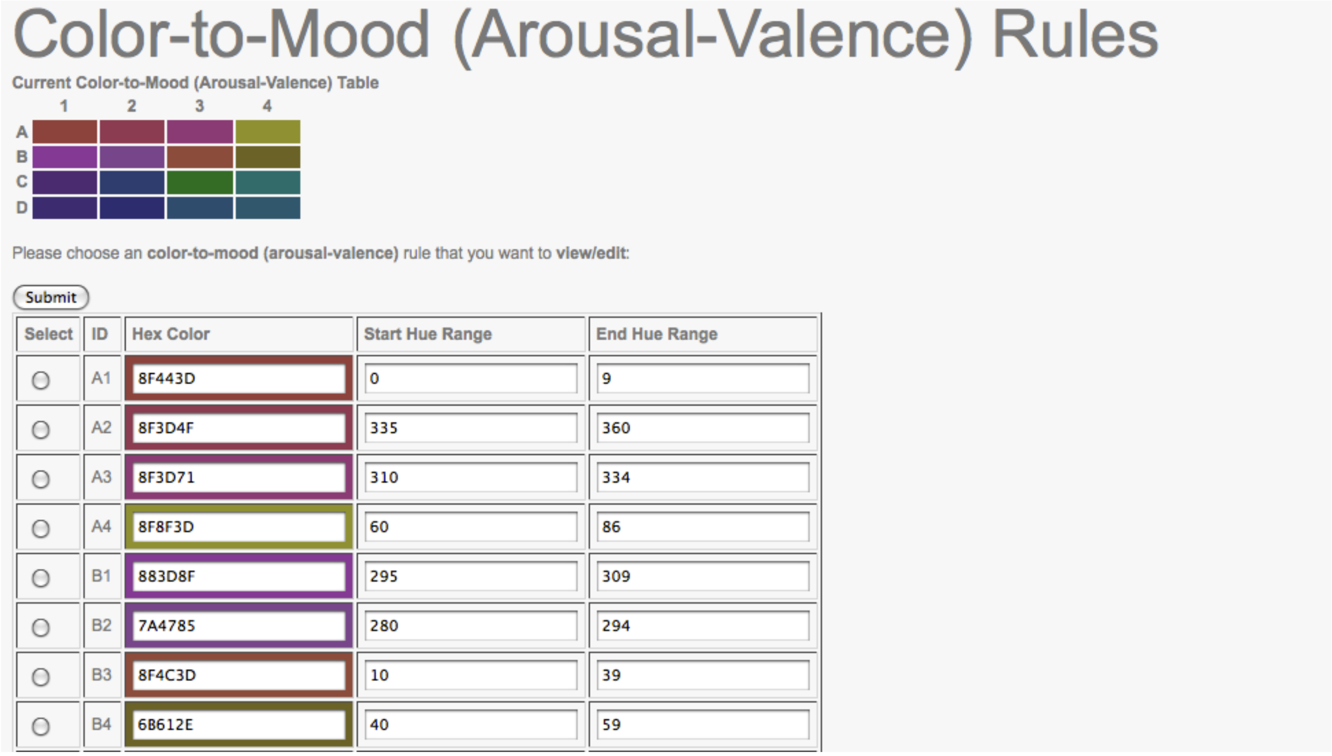


ID	OCC_ID	WORD
1	1	happy
2	1	cheerful
3	1	contented
4	1	content
5	1	glad
6	1	elated
7	1	euphoric
8	1	felicitous
9	1	joyful
10	1	joyous
11	1	blessed
12	1	blissful
13	1	bright
14	1	golden
15	1	halcyon
16	1	prosperous
17	1	laughing
18	1	riant
19	1	happiness
20	1	felicity
21	1	happiness
78	4	pride
80	4	self-esteem
81	4	self-pride
82	4	ego
83	4	egotism
84	4	self-importance
85	4	amour propre
86	4	conceit
87	4	self-love

Figure 13: Text Data Set

4.3 User Rule Customization

Prior to any matching involving images, user rule customization must be set up. This project provides a default set of rules that map certain range of colors to each sector. These rules are also called "color-to-mood" rules in this project. The hue wheel colors were broken into 16 bins to match each 16 sectors. For instance the orange hue range from 10 to 39 degrees will be mapped to B3 to represent a pleasantness. See Figure 14 for the interface that a user can use to change the color-to-mood rules.



Color-to-Mood (Arousal-Valence) Rules

Current Color-to-Mood (Arousal-Valence) Table

	1	2	3	4
A	■	■	■	■
B	■	■	■	■
C	■	■	■	■
D	■	■	■	■

Please choose an **color-to-mood (arousal-valence)** rule that you want to **view/edit**:

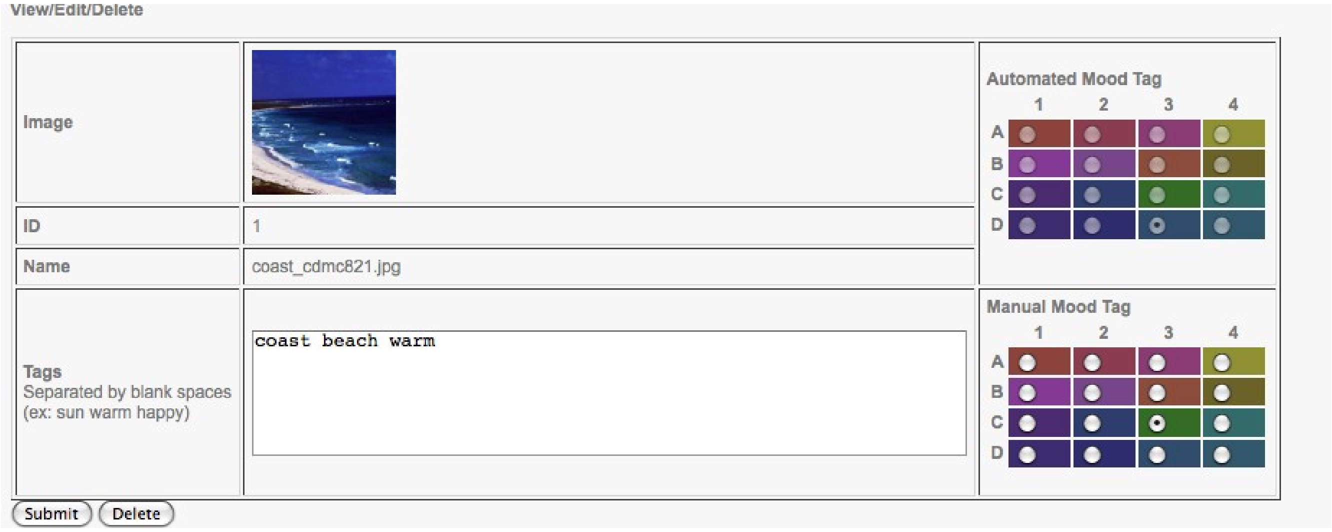
Select	ID	Hex Color	Start Hue Range	End Hue Range
○	A1	8F443D	0	9
○	A2	8F3D4F	335	360
○	A3	8F3D71	310	334
○	A4	8F8F3D	60	86
○	B1	883D8F	295	309
○	B2	7A4785	280	294
○	B3	8F4C3D	10	39
○	B4	6B612E	40	59

Figure 14: Interface to modify the color-to-mood rules

4.4 Tagging

4.4.1 Mood Tagging

Although the system can automatically annotate the audio and image files, there may be instances where the system can be incorrect (see the Test Scenarios description in the Results section), hence it is useful and intuitive to allow the user to annotate the mood of the file manually by using the same mood annotation. See Figure 15 to see the interface the user is allowed to use to annotate the mood of a file.



View/Edit/Delete

Image	
ID	1
Name	coast_cdmc821.jpg
Tags Separated by blank spaces (ex: sun warm happy)	coast beach warm

Automated Mood Tag

	1	2	3	4
A	○	○	○	○
B	○	○	○	○
C	○	○	○	○
D	○	○	●	○

Manual Mood Tag

	1	2	3	4
A	○	○	○	○
B	○	○	○	○
C	○	○	●	○
D	○	○	○	○

Figure 15: Interface to edit data sample

4.4.2 Text Tagging

Not all multimedia on the Internet has text tags, but tagging is become more important in searching them and being used more and more everyday. In the context of this project, tagging is not completely relied upon to find the initial matching results. Tags are used to help filter the initial matching result set or in some cases to broaden the result set as described later in the Medium Matching section. When including the tags in the filtering process, the tags are parsed into word tokens and then included in the SQL statement by using a LIKE predicate and each term is followed by an OR operator in order to be inclusive. The AND operator would exclude some possibly relevant results. Figure 15 shows the interface that allows the user to edit the tags.

4.5 Medium Mood Extraction

Before any inputted medium can be matched with another medium, the set of files for images and audio must be analyzed by extracting the moods and storing the values within the database for each file. Then when the match is done, the database is searched for those stored values and taken into account.

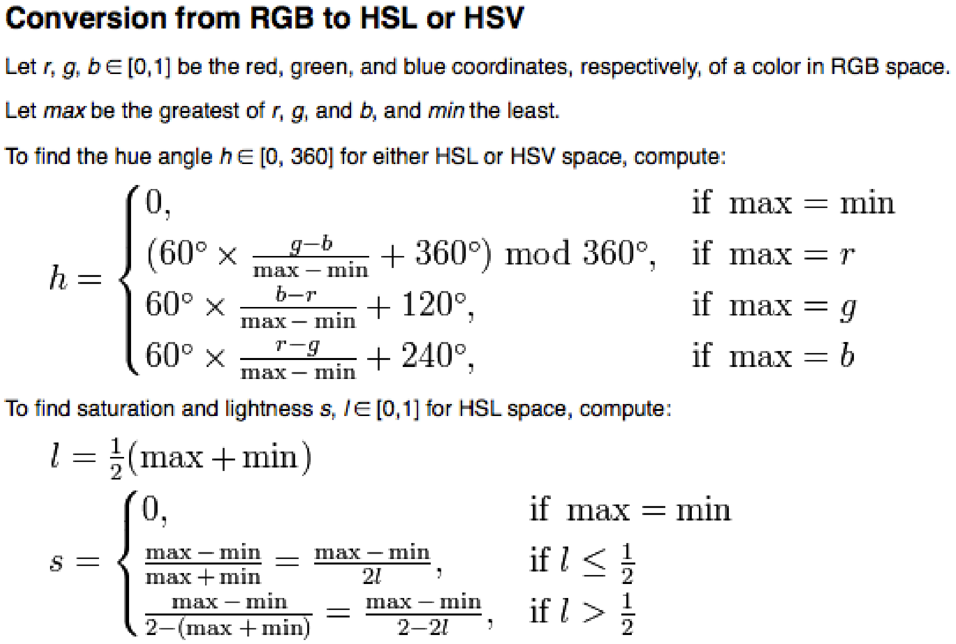
4.5.1 Audio Mood Extraction

Marsyas[30] is used and accepts .WAV and .AU format. All other formats can be converted to .WAV or .AU with the appropriate audio converter. Pre-processing must be completed. In order to do that, a training set of the audio files must be created. In order to that, of the 1,000 song clips available, each sector of the 16-sector model was populated with at least 2 songs in order to train each sector. Then the training set was fed through Marsyas and the following common suite of audio features were extracted (as pulled from the Marsyas code): Series, SoundFileSource, AudioSink, Stereo2Mono, TimbreFeatures, ShiftInput, Fanout, ZeroCrossings, Windowing, PowerSpectrum, PowerSpectrumNet1, STFT_features, Centroid, Rolloff, Flux, MFCC, Spectrum2Chroma, SCF, SFM, Filter, LPC, LSP, LPCC, TextureStats, Memory, Mean, StandardDeviation, Annotator, Classifier, ZeroRClassifier, GaussianClassifier, SVMClassifier, and Confidence. When training the set, the classifiers Gaussian and SVM were also trained. These two are commonly used in music emotion retrieval. Then the 100 songs were passed through Marsyas and classified based on the training set that was set to the 16 sector model. When Marsyas classified each song, it analyzed each second and predicted which sector that 1 second belongs to and gave a confidence level. Each sector was grouped and then added each group's confidence values and the dominant confidence indicated the dominant sector. These values were saved in a tab delimited file. The output files were then taken and inserted into the AUDIO_ANALYSIS table and the dominant sector would then be saved into the database, specifically in the AUDIO_FILES table so that it does need to be reprocessed again during matching.

4.5.2 Image Mood Extraction

The mood extraction is primarily done by analyzing the hue colors of the image. The code handled each pixel and converted the RGB to HSL (Hue, Saturation, and Lightness). These files were saved in a tab delimited file which will then be parsed and saved into the database, specifically in the

IMAGE_ANALYSIS table. The dominant hue value will be calculated dynamically when necessary by the database. The dominant value is then looked up in the color-to-mood rules and then saves the looked up value in the image file entry. This lookup is done when the system needs to retrieve the mood of the image automatically. Figure 16 shows the equations to convert RGB values to hue values.



Conversion from RGB to HSL or HSV

Let *r*, *g*, *b* ∈ [0,1] be the red, green, and blue coordinates, respectively, of a color in RGB space.

Let *max* be the greatest of *r*, *g*, and *b*, and *min* the least.

To find the hue angle *h* ∈ [0, 360] for either HSL or HSV space, compute:

$$h = \begin{cases} 0, & \text{if } \max = \min \\ (60^\circ \times \frac{g-b}{\max - \min} + 360^\circ) \bmod 360^\circ, & \text{if } \max = r \\ 60^\circ \times \frac{b-r}{\max - \min} + 120^\circ, & \text{if } \max = g \\ 60^\circ \times \frac{r-g}{\max - \min} + 240^\circ, & \text{if } \max = b \end{cases}$$

To find saturation and lightness *s*, *l* ∈ [0,1] for HSL space, compute:

$$l = \tfrac{1}{2}(\max + \min)$$

$$s = \begin{cases} 0, & \text{if } \max = \min \\ \frac{\max - \min}{\max + \min} = \frac{\max - \min}{2l}, & \text{if } l \leq \frac{1}{2} \\ \frac{\max - \min}{2-(\max + \min)} = \frac{\max - \min}{2-2l}, & \text{if } l > \frac{1}{2} \end{cases}$$

Figure 16: Conversion from RGB to HSL

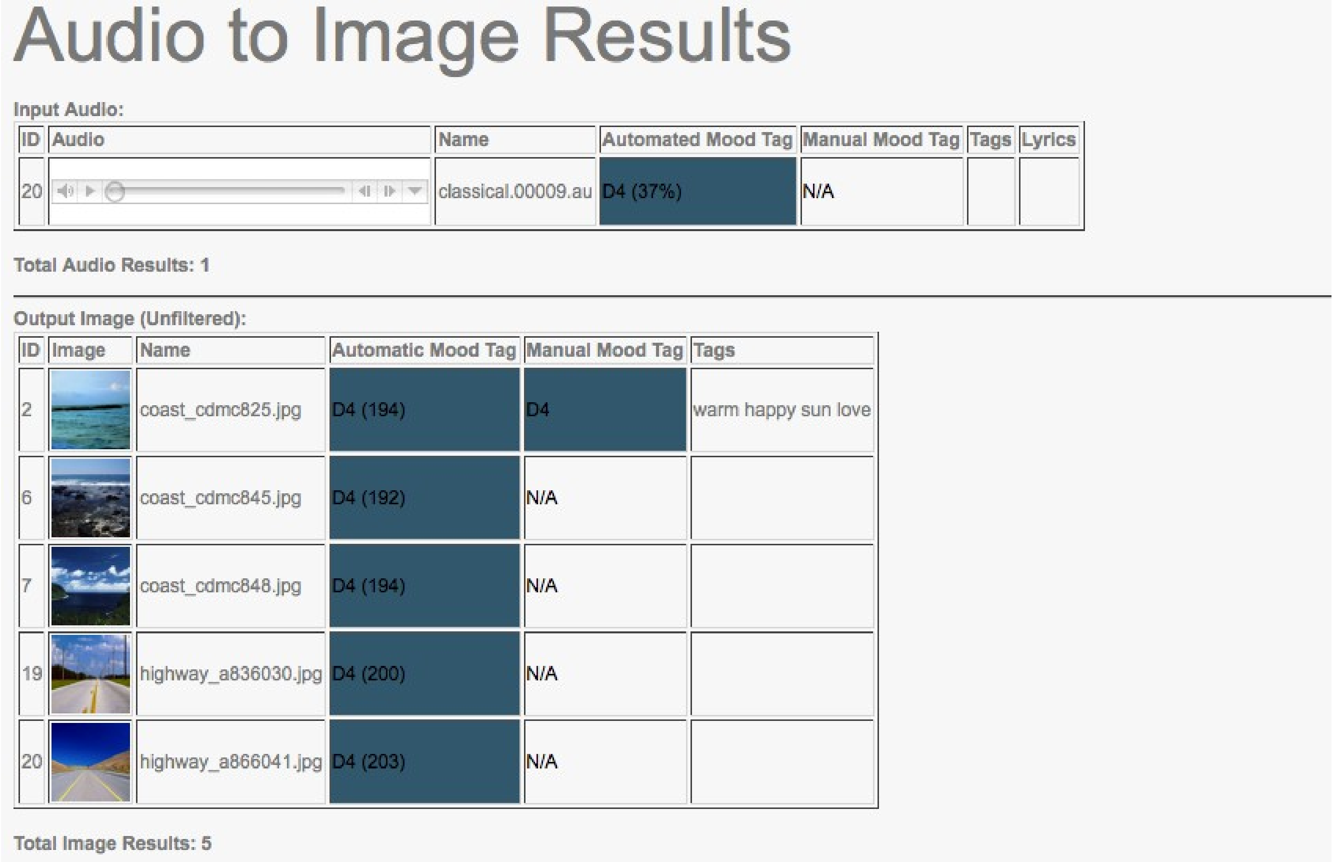
4.5.3 Text Mood Extraction

The mood extraction for text is straightforward. If text is provided as input for a match, the string is broken into word tokens. Text tags and lyrics are manually annotated and saved into the database as they were entered with blank spaces and break lines. When searching for certain text words, the project utilizes the database's inherent search predicate of LIKE '%<word>%'. The text acts mainly as a filter and in some matching cases, to help broaden results. The words with emotional value will be found in the SYNONYMS table. The words not in the SYNONYMS table, which are most likely nouns, will act to narrow the matching results by comparing the tags and lyrics.

4.7 Medium Matching

The project explores these six distinct medium matching processes in this section: audio to image, audio to text, image to audio, image to text, text to audio, and text to image.

4.7.1 Audio to Image



Audio to Image Results

Input Audio:

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
20		classical.00009.au	D4 (37%)	N/A		

Total Audio Results: 1

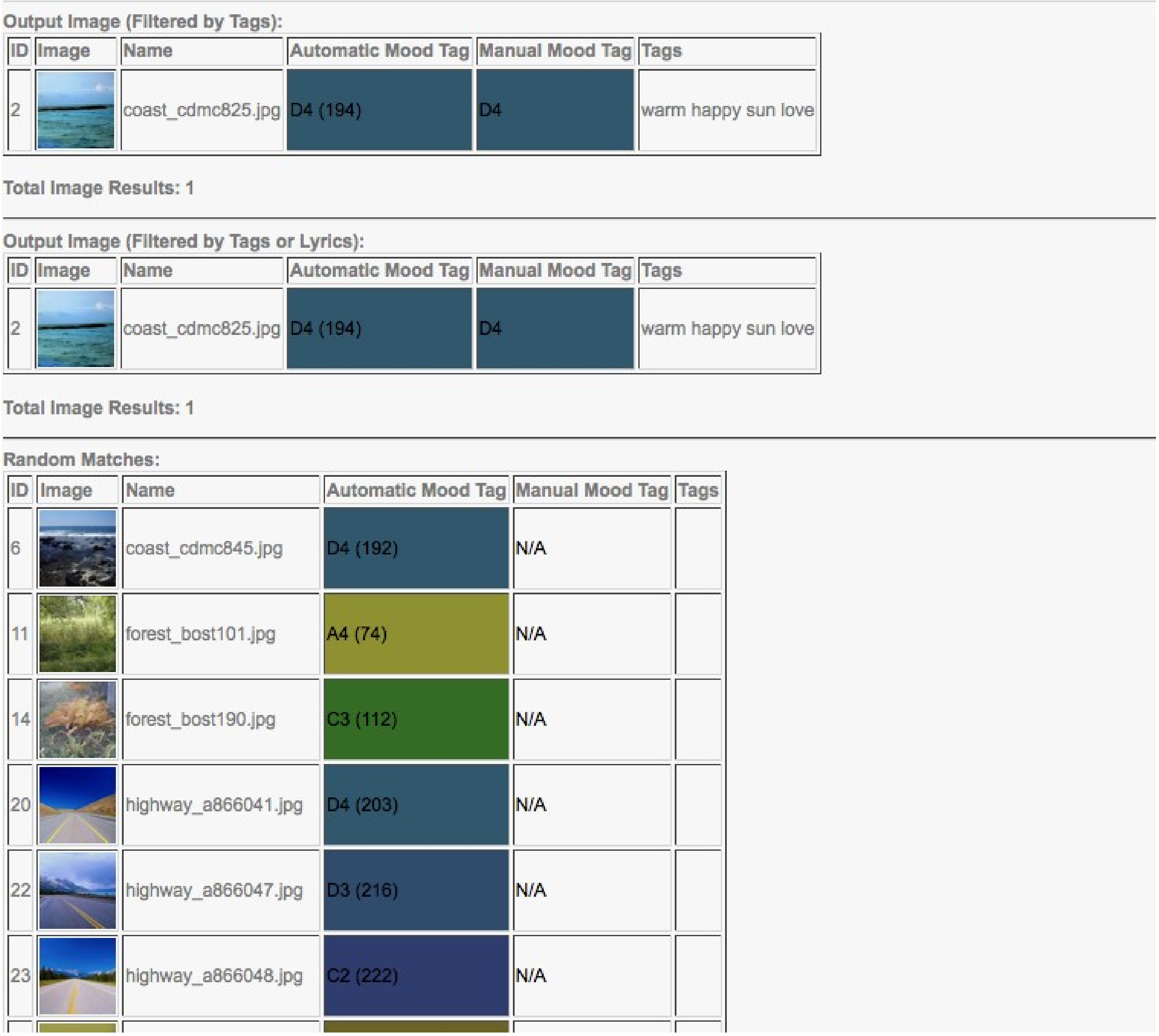
Output Image (Unfiltered):

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
2		coast_cdmc825.jpg	D4 (194)	D4	warm happy sun love
6		coast_cdmc845.jpg	D4 (192)	N/A	
7		coast_cdmc848.jpg	D4 (194)	N/A	
19		highway_a836030.jpg	D4 (200)	N/A	
20		highway_a866041.jpg	D4 (203)	N/A	

Total Image Results: 5

Figure 17: Audio to Image Input and CEM Output Sample

When matching audio to image, an audio file is chosen by the user. The audio file's mood is retrieved. The value is available either in the MOOD_TAG or the AUTO_CONTENT_MOOD column of the AUDIO_FILES table. If the MOOD_TAG is filled with a sector value then that value is used because it is manually annotated by a user which is assumed to be correct so it takes precedence over the AUTO_CONTENT_MOOD column. If the MOOD_TAG is not filled with a sector, the AUTO_CONTENT_MOOD column is then looked up which should contain a value previously processed by the Marsyas system. The IMAGE_FILES table is then updated because the AUTO_CONTENT_MOOD column needs to be updated with the latest sector due to any changes made by the user in the color-to-mood rules. For instance, the red color once represented by “A1” for anger could now be represented by “A3” for love. Once that column has been updated for all image files in the database, the MOOD_TAG is searched for the same value that was chosen by the audio file. If there is no value in the image file's MOOD_TAG, then it searches in the AUTO_CONTENT_MOOD column. If they match, then the image file is retrieved in the first set of results. The second of results takes the first set of results and filters them out based on any available tags found from the chosen audio file that are also found in the image file tags. The third set takes the second set of results and filters it out even further by examining any lyrics available from the audio file and searching for matching words in the tags of the image files. Figure 17 and 18 demonstrates an example of matching audio to image.



Output Image (Filtered by Tags):

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
2		coast_cdmc825.jpg	D4 (194)	D4	warm happy sun love

Total Image Results: 1

Output Image (Filtered by Tags or Lyrics):

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
2		coast_cdmc825.jpg	D4 (194)	D4	warm happy sun love

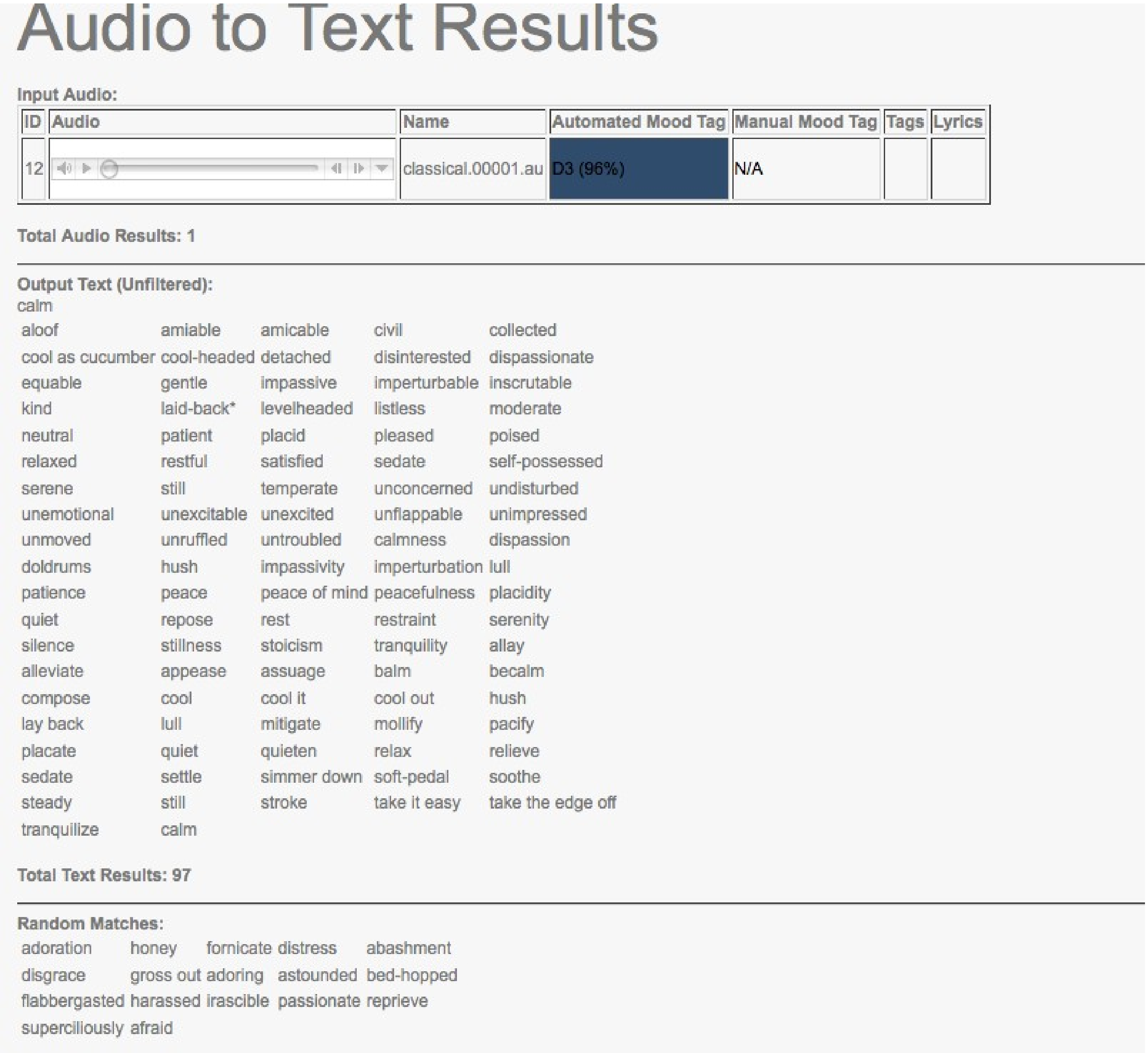
Total Image Results: 1

Random Matches:

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
6		coast_cdmc845.jpg	D4 (192)	N/A	
11		forest_bost101.jpg	A4 (74)	N/A	
14		forest_bost190.jpg	C3 (112)	N/A	
20		highway_a866041.jpg	D4 (203)	N/A	
22		highway_a866047.jpg	D3 (216)	N/A	
23		highway_a866048.jpg	C2 (222)	N/A	

Figure 18: Audio to Image CEMT and Random Output Sample

4.7.2 Audio to Text



Audio to Text Results

Input Audio:

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
12		classical.00001.au	D3 (96%)	N/A		

Total Audio Results: 1

Output Text (Unfiltered):

calm

aloof amiable amicable civil collected
cool as cucumber cool-headed detached disinterested dispassionate
equable gentle impassive imperturbable inscrutable
kind laid-back* levelheaded listless moderate
neutral patient placid pleased poised
relaxed restful satisfied sedate self-possessed
serene still temperate unconcerned undisturbed
unemotional unexcitable unexcited unflappable unimpressed
unmoved unruffled untroubled calmness dispassion
doldrums hush impassivity imperturbation lull
patience peace peace of mind peacefulness placidity
quiet repose rest restraint serenity
silence stillness stoicism tranquility allay
alleviate appease assuage balm becalm
compose cool cool it cool out hush
lay back lull mitigate mollify pacify
placate quiet quieten relax relieve
sedate settle simmer down soft-pedal soothe
steady still stroke take it easy take the edge off
tranquilize calm

Total Text Results: 97

Random Matches:

adoration honey fornicate distress abashment
disgrace gross out adoring astounded bed-hopped
flabbergasted harassed irascible passionate reprieve
superciliously afraid

Figure 19: Audio to Text Input and CEM and Random Output Sample

A user selects an audio file from the database to match to text, meaning a set of words will be displayed that will match the overall mood of the chosen audio file. The dominant arousal-valence sector of the chosen audio file must be retrieved initially. If the chosen audio file has a manually annotated mood tag, then it is already stored in the database and that is picked up since it takes precedence over an automatically generated value. If the manually annotated mood tag does not exist in the database, then it is automatically retrieved according to the preprocessed value stored in the database. When the arousal-valence sector has been determined, the database is then searched for the group of words that is mapped to that arousal-valence sector. That matching group of words is retrieved and displayed to the user. If the chosen audio file has text tags or lyrics, then that information may be used to broaden the set of words since the automated arousal-valence sector may be incorrect. Figure 19 demonstrates an example of matching audio to text.

4.7.3 Image to Audio

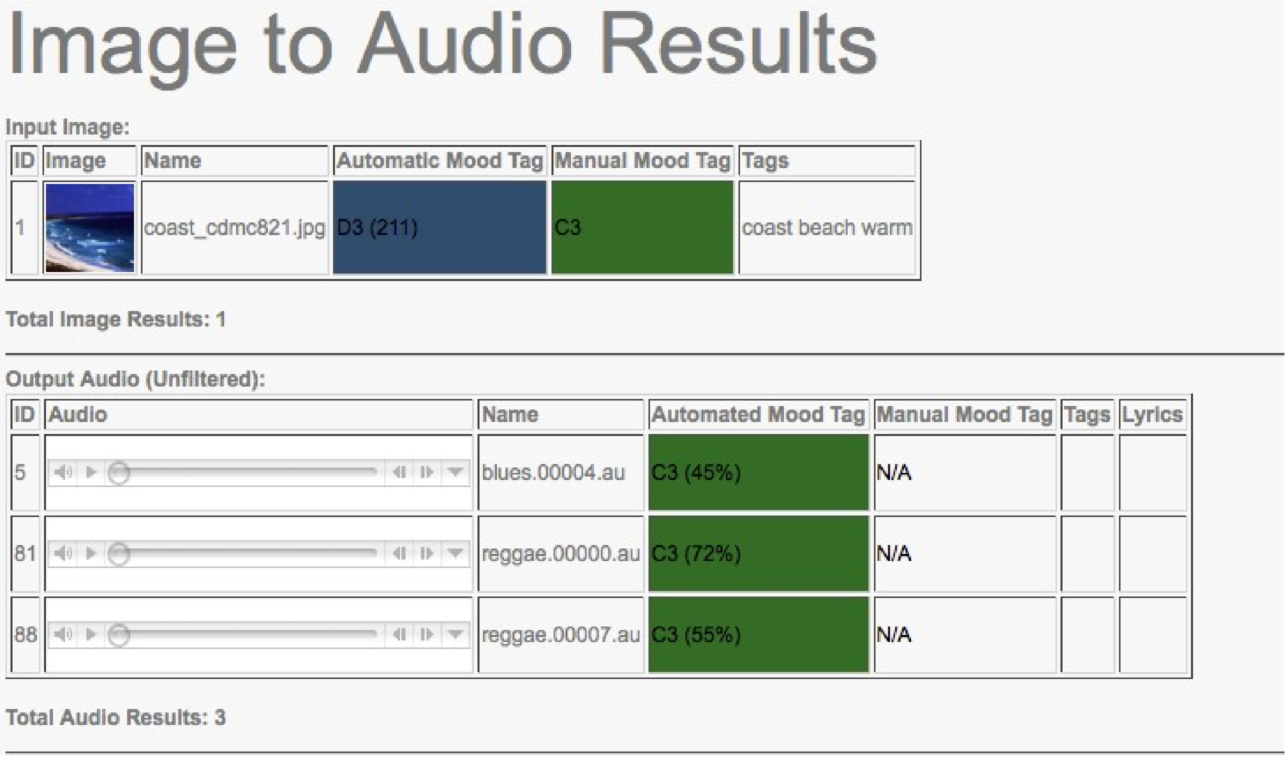


Image to Audio Results

Input Image:

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
1		coast_cdmc821.jpg	D3 (211)	C3	coast beach warm

Total Image Results: 1

Output Audio (Unfiltered):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
5		blues.00004.au	C3 (45%)	N/A		
81		reggae.00000.au	C3 (72%)	N/A		
88		reggae.00007.au	C3 (55%)	N/A		

Total Audio Results: 3

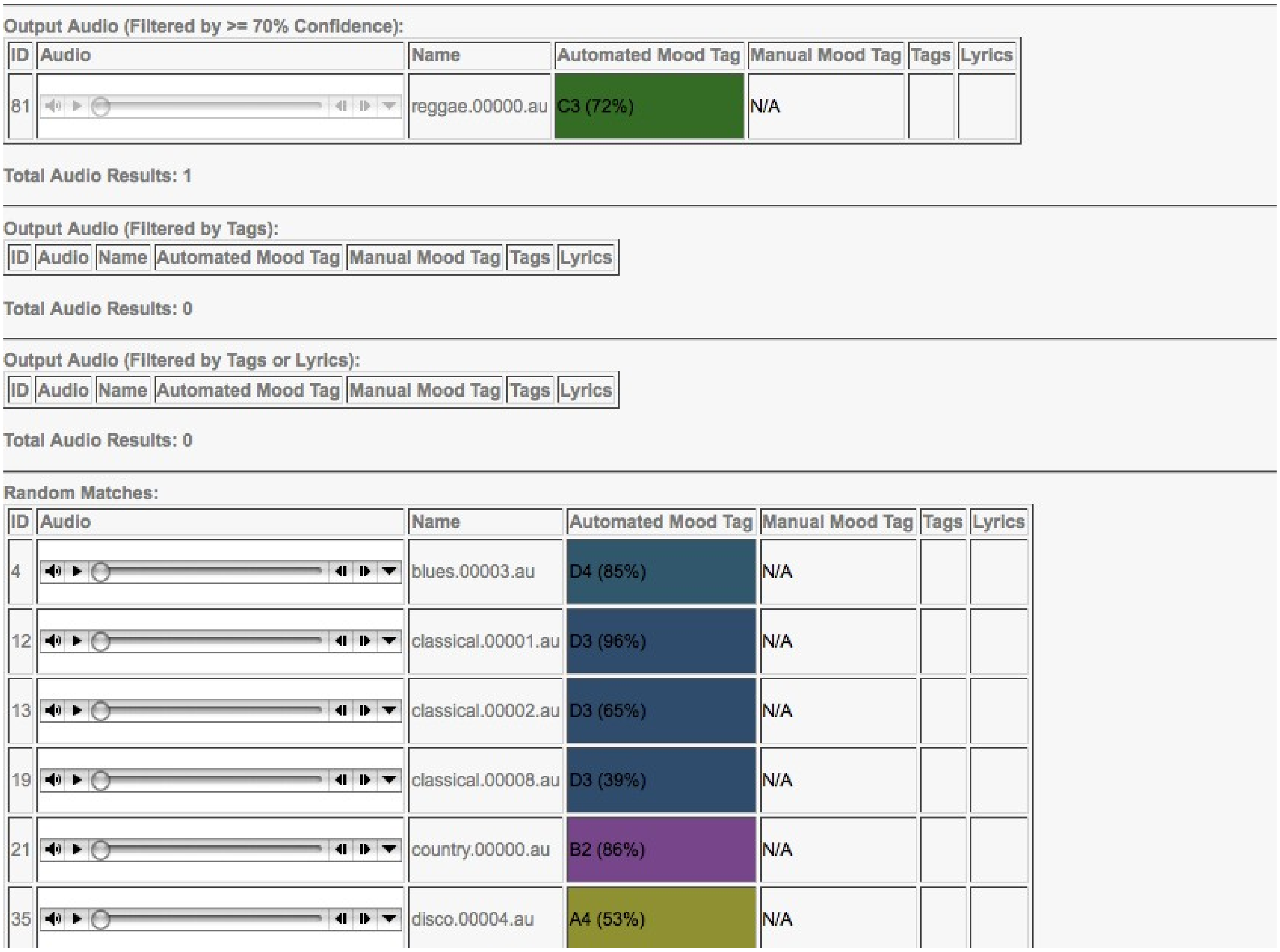
Figure 20: Image to Audio Input and CEM Output Sample

A user selects an image from the database to match to audio. Inputting one image may provide a set of images that fit the overall mood of the input. First the dominant arousal-valence sector must be retrieved from the chosen image. This is done by looking up the MOOD_TAG column in the

IMAGE_FILES table first since it contains a manually annotated mood tag and it takes precedence over the automatically retrieved value. If there is no value in MOOD_TAG, then the AUTO_CONTENT_MOOD column is looked up. The AUTO_CONTENT_MOOD column is then updated according to any available color-to-mood rules by finding the dominant hue color of the chosen image and then looking up the table containing the color-to-mood rules. It is updated at this point because prior to running this match, the user could have updated the color-to-mood rules.

Once the dominant arousal-valence sector has been determined, the set of audio files in AUDIO_FILES table with the matching arousal-valence sector is searched. No update is done for the AUTO_CONTENT_MOOD or MOOD_TAG columns in AUDIO_FILES table because there are no audio-to-mood rules. When searching for a matching sector in the audio file entry, the MOOD_TAG is first searched, then the AUTO_CONTENT_MOOD column because again, the manually annotated mood tag takes precedence. For example, suppose Audio File A has MOOD_TAG: B1, AUTO_CONTENT_MOOD: A1 and Audio File B has MOOD_TAG: NULL, AUTO_CONTENT_MOOD: A1, and if the chosen image's arousal-valence sector is determined to be A1 then Audio File B would be chosen and not Audio File A because Audio File A's true mood is B1 and not A1.

The first set of results will be done based on the mood content search. The next set of results will then be returned based on any image tags matching with the audio tags. Then the third, CEMT set will be returned based on any image tags that match with the audio tags or audio lyrics. Figure 20 and 21 demonstrate an example of matching image to audio.



Output Audio (Filtered by >= 70% Confidence):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
81		reggae.00000.au	C3 (72%)	N/A		

Total Audio Results: 1

Output Audio (Filtered by Tags):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics

Total Audio Results: 0

Output Audio (Filtered by Tags or Lyrics):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics

Total Audio Results: 0

Random Matches:

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
4		blues.00003.au	D4 (85%)	N/A		
12		classical.00001.au	D3 (96%)	N/A		
13		classical.00002.au	D3 (65%)	N/A		
19		classical.00008.au	D3 (39%)	N/A		
21		country.00000.au	B2 (86%)	N/A		
35		disco.00004.au	A4 (53%)	N/A		

Figure 21: Image to Audio CEMT and Random Output Sample

4.7.4 Image to Text

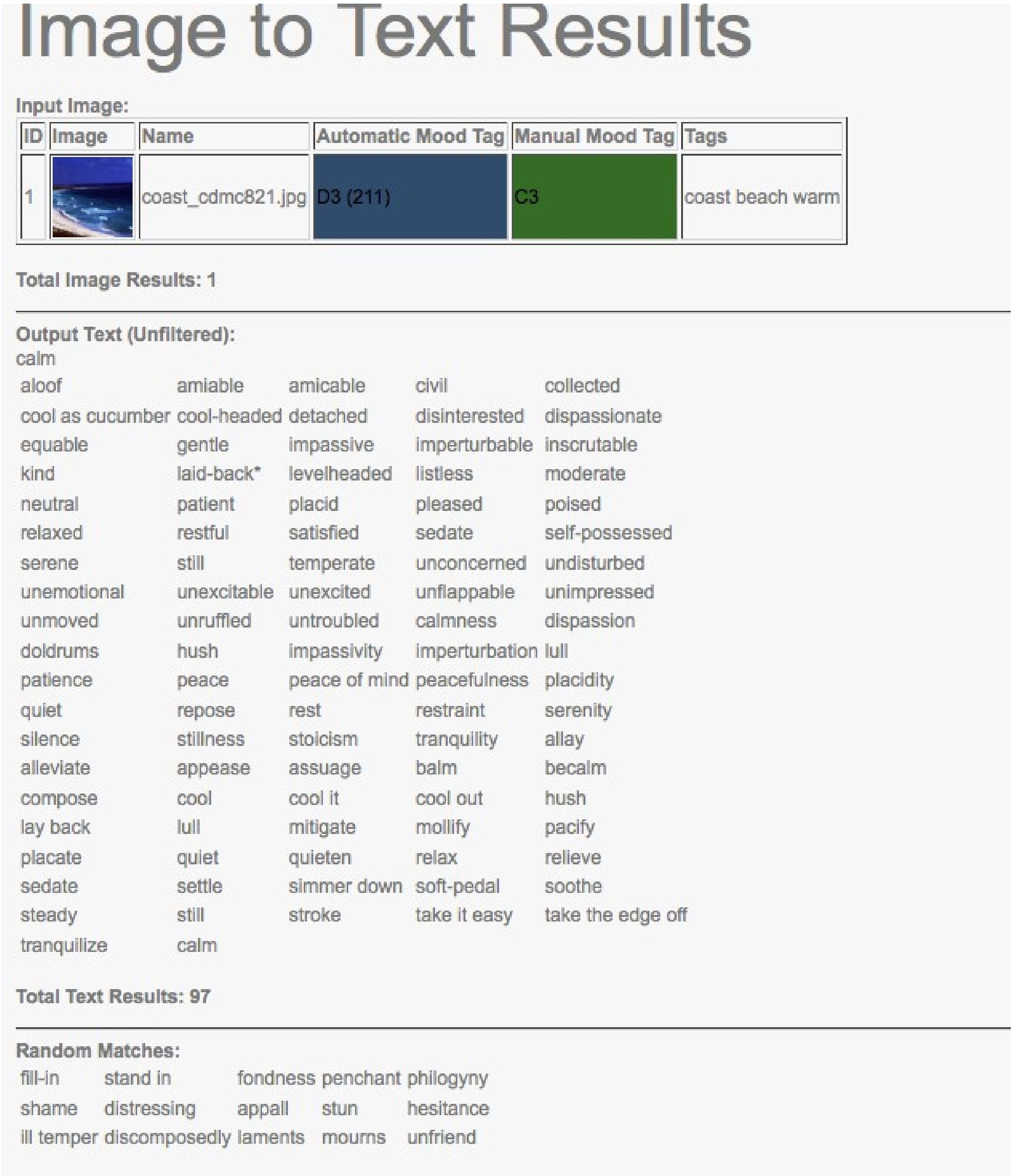


Image to Text Results

Input Image:

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
1		coast_cdmc821.jpg	D3 (211)	C3	coast beach warm

Total Image Results: 1

Output Text (Unfiltered):

```
calm
aloof              amiable       amicable      civil            collected
cool as cucumber   cool-headed   detached      disinterested    dispassionate
equable            gentle        impassive     imperturbable    inscrutable
kind               laid-back*    levelheaded   listless         moderate
neutral            patient       placid        pleased          poised
relaxed            restful       satisfied     sedate           self-possessed
serene             still         temperate     unconcerned      undisturbed
unemotional        unexcitable   unexcited     unflappable      unimpressed
unmoved            unruffled     untroubled    calmness         dispassion
doldrums           hush          impassivity   imperturbation   lull
patience           peace         peace of mind peacefulness     placidity
quiet              repose        rest          restraint        serenity
silence            stillness     stoicism      tranquility      allay
alleviate          appease       assuage       balm             becalm
compose            cool          cool it       cool out         hush
lay back           lull          mitigate      mollify          pacify
placate            quiet         quieten       relax            relieve
sedate             settle        simmer down   soft-pedal       soothe
steady             still         stroke        take it easy     take the edge off
tranquilize        calm
```

Total Text Results: 97

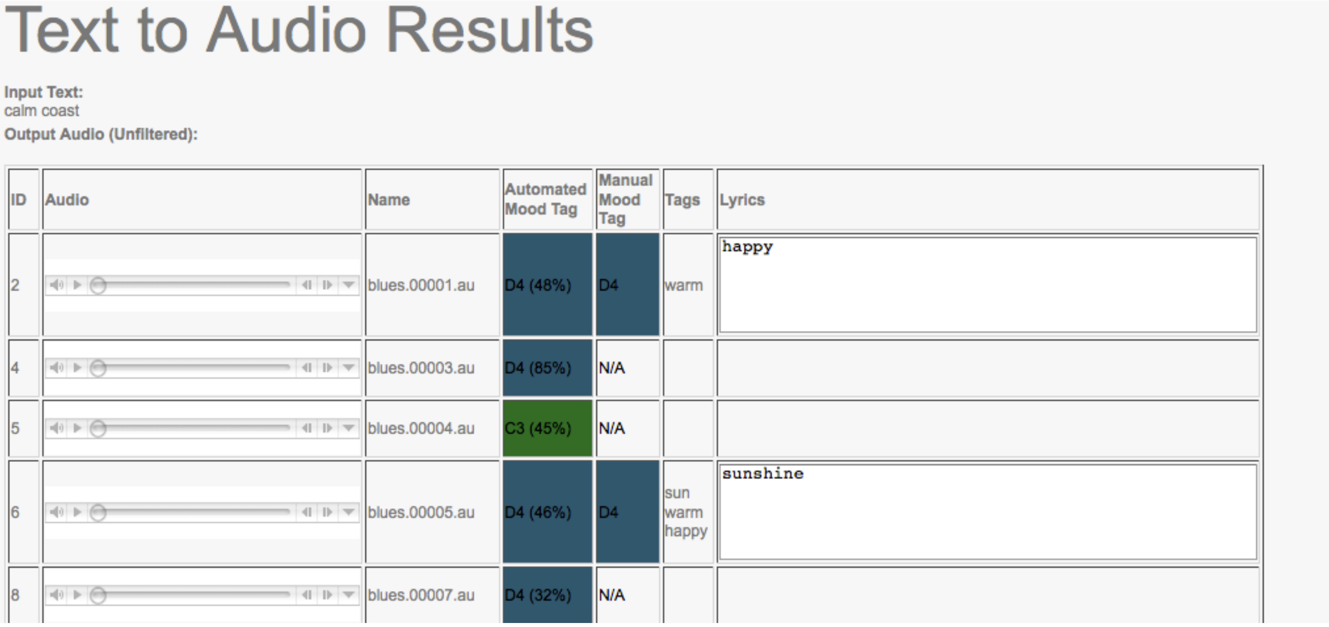
Random Matches:

```
fill-in      stand in        fondness  penchant  philogyny
shame        distressing     appall    stun      hesitance
ill temper   discomposedly   laments   mourns    unfriend
```

Figure 22: Image to Text Input and CEM and Random Output Sample

A user selects an image from the database to match to text, meaning a set of words will be displayed that will match the overall mood of the chosen image. This is a fairly straightforward process. The dominant arousal-valence sector of the chosen image must be retrieved first. If the chosen image has a manually annotated mood tag, then it is already stored in the database and that is picked up since it takes precedence over an automatically generated value. If the manually annotated mood tag does not exist in the database, then it is automatically retrieved according to the dominant hue color stored in the database and then the hue color is looked up in the color-to-mood rules available. When the arousal-valence sector has been determined, the database is then searched for the group of words that is mapped to that arousal-valence sector. That matching group of words is retrieved and displayed to the user. If the chosen image has text tags, then that information may be used to broaden the set of words since the automated arousal-valence sector may be incorrect. Figure 22 demonstrates an example of matching image to text.

4.7.5 Text to Audio



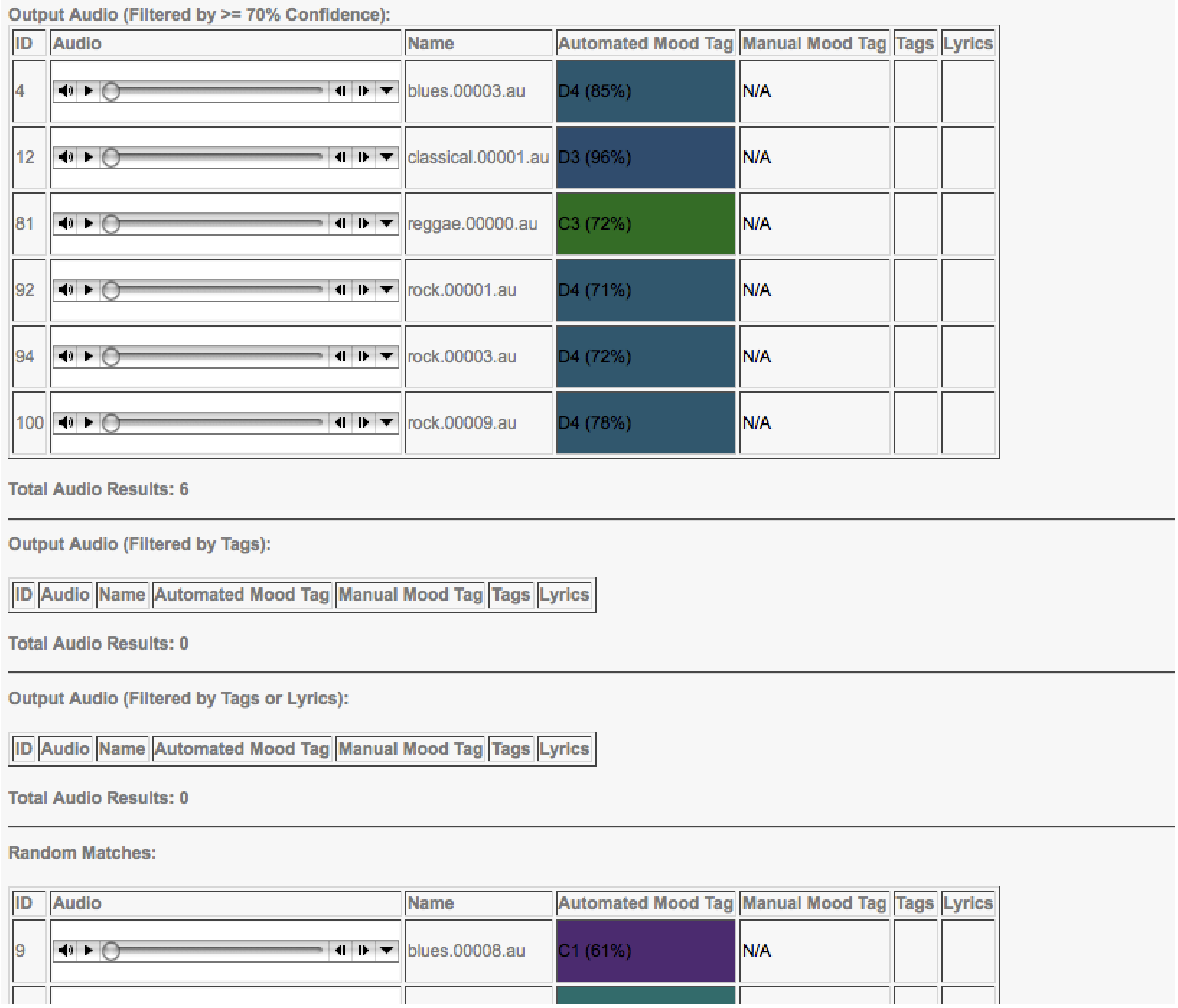
Text to Audio Results

Input Text:
calm coast
Output Audio (Unfiltered):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
2		blues.00001.au	D4 (48%)	D4	warm	happy
4		blues.00003.au	D4 (85%)	N/A		
5		blues.00004.au	C3 (45%)	N/A		
6		blues.00005.au	D4 (46%)	D4	sun warm happy	sunshine
8		blues.00007.au	D4 (32%)	N/A		

Figure 23: Text to Audio Input and CEM Output Sample

When matching text to audio, several database lookups are performed since the text values are already stored in the database and the audio files are already preprocessed and stored in the tables. There is no further need to re-process the audio files again because there are no audio-to-mood rules like there are with color-to-mood rules. User text input is parsed into word tokens. Those tokens are then searched for in the join of the OCC and the SYNONYMS tables. Once there are matches found, the corresponding arousal-valence sectors are stored. Then a search is done on the AUTO_CONTENT_MOOD column of the AUDIO_FILES table for the matching arousal-valence sectors. The MOOD_TAG column of the AUDIO_FILES table is actually first searched since it takes precedence over the automated AUTO_CONTENT_MOOD column because this concept is subjective, it is the user that is correct when assigning mood to the item. If the MOOD_TAG column is not filled with an arousal-valence sector, then the AUTO_CONTENT_MOOD column is relied upon by the search. The database then returns the matching results based on the dominant mood. This result set is then displayed to the user. Another set of results is displayed but has been CEMT by any matching text tags that was found in the input text. This CEMT result set should be smaller or equal to the initial mood content CEMT set since the audio files may have text tags or not. Then a third set of matching results will be CEMT based on any provided lyrics that match the input text. Again, this third set should be smaller or equal to the second set of matches since the audio files may have lyrics or not. Figure 23 and 24 demonstrate an example of matching text to audio.



Output Audio (Filtered by >= 70% Confidence):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
4		blues.00003.au	D4 (85%)	N/A		
12		classical.00001.au	D3 (96%)	N/A		
81		reggae.00000.au	C3 (72%)	N/A		
92		rock.00001.au	D4 (71%)	N/A		
94		rock.00003.au	D4 (72%)	N/A		
100		rock.00009.au	D4 (78%)	N/A		

Total Audio Results: 6

Output Audio (Filtered by Tags):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics

Total Audio Results: 0

Output Audio (Filtered by Tags or Lyrics):

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics

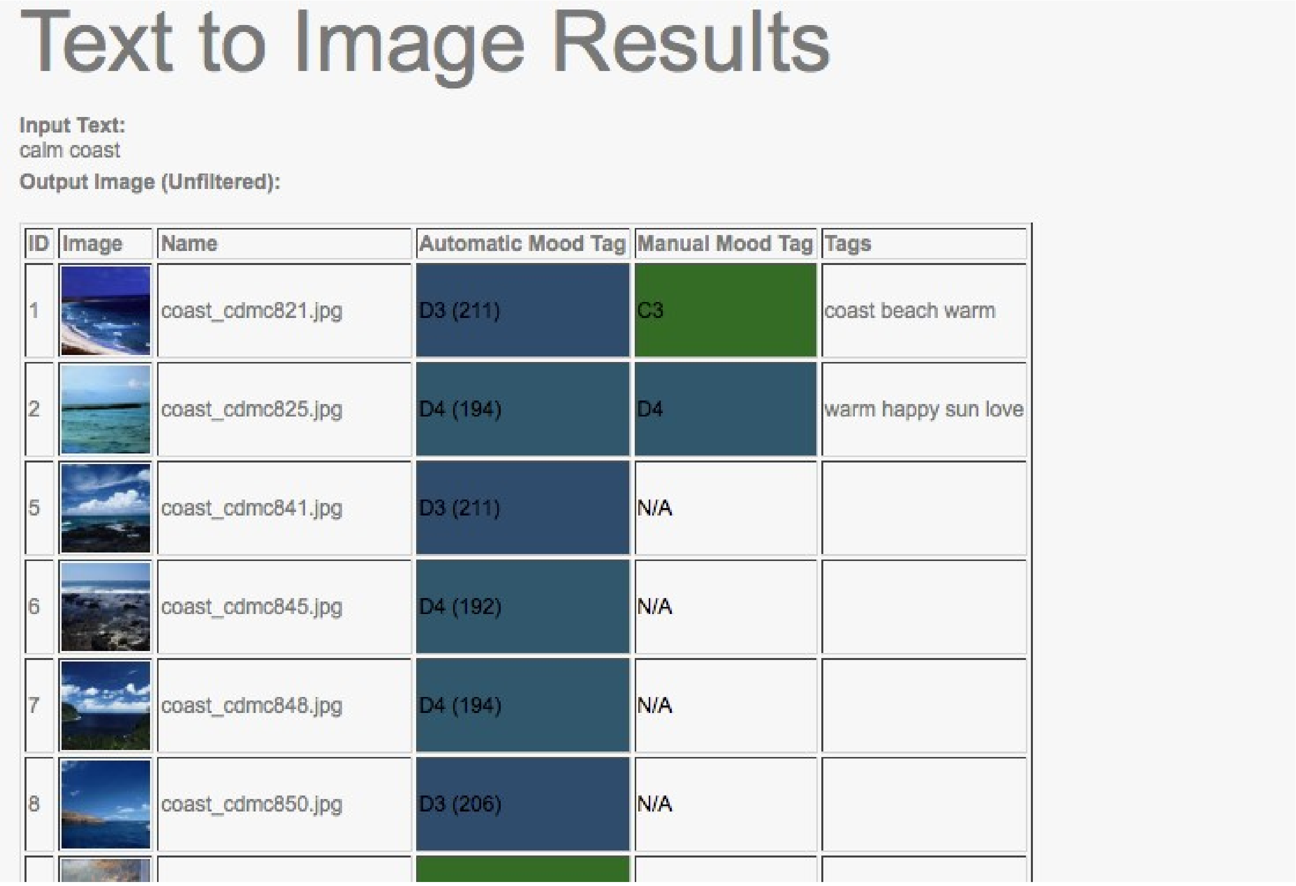
Total Audio Results: 0

Random Matches:

ID	Audio	Name	Automated Mood Tag	Manual Mood Tag	Tags	Lyrics
9		blues.00008.au	C1 (61%)	N/A		

Figure 24: Text to Audio CEMT and Random Output Sample

4.7.6 Text to Image



Text to Image Results

Input Text:
calm coast

Output Image (Unfiltered):

ID	Image	Name	Automatic Mood Tag	Manual Mood Tag	Tags
1		coast_cdmc821.jpg	D3 (211)	C3	coast beach warm
2		coast_cdmc825.jpg	D4 (194)	D4	warm happy sun love
5		coast_cdmc841.jpg	D3 (211)	N/A	
6		coast_cdmc845.jpg	D4 (192)	N/A	
7		coast_cdmc848.jpg	D4 (194)	N/A	
8		coast_cdmc850.jpg	D3 (206)	N/A	

Figure 25: Text to Image Input and CEM Output Sample

Matching text to image consists of updating the arousal-valence sector in the image file database entries according to the user color-to-mood rules, finding matching words in the database that have corresponding arousal-valence sectors, and then finding a match based on the matching arousal-valence sectors, then filtering out the image results based on any additional tags provided. When a blank space delimited string of text is provided to the system, the string is parsed into word tokens. Those tokens are then searched for in the join of the OCC and SYNONYMS tables. When a match is found in the SYNONYMS table, the corresponding OCC word is found as well which is mapped to an arousal-valence sector such as “A1” or “B2.” The arousal-valence sector is stored for later use. Then the AUTO_CONTENT_MOOD column in the IMAGE_FILES table is updated according to the user color-to-mood rules. The dominant hue value is found per image file and then correlated to the color-to-mood rule. Once the arousal-valence sectors are known for both text and images, then a search is done based on that matching arousal-valence sector. A search is done on the IMAGE_FILES table. The MOOD_TAG column of the IMAGE_FILES table is first searched since it takes precedence over the automated AUTO_CONTENT_MOOD column. If the MOOD_TAG column is not filled with an arousal-valence sector, then the AUTO_CONTENT_MOOD column is relied upon by the search. The database then returns the matching results based on the dominant mood. This result set is then displayed to the user. Another set of results is displayed but has been CEMT by any matching text tags that was found in the input text. This CEMT result set should be smaller or equal to the initial mood content CEMT set since the images may have text tags or not. If there are multiple matches when searching for matching words in the SYNONYMS table and different arousal-valence sectors are found, all arousal-valence sectors are included in the search subsequent database searches. Figure 25 demonstrates an example of matching text to image.

5.0 Results

5.1 Test Scenarios

This section will describe the types of test cases that are accepted by the system and will yield a match that is most likely to appease a user; also described are the test cases scenarios that would not yield an appropriate match. The training of the audio data set that is initially used is vital to the entire data set that is fed into the system that were not used to train the system. If the audio file data sets is not classified in the most appropriate arousal-valence sector, it is bound to be matched to the incorrect text and image file. Also if the mood of the music do not match the mood conveyed by the lyrics, then that will also create a mismatch from the perspective of the user. Also if the manually annotated mood tag, tags, and lyrics are purposely applied incorrectly to an audio file then that will also cause a mismatch. This system is designed to create the best match when the training set has been well-defined so as not to confuse the classifier, when the audio file is appropriately classified into the appropriate mood, arousal-valence sector, and the manual mood tag, the tags, and lyrics are appropriately labeled.

Because the system determines the dominant hue of a picture and determines the arousal-valence sector based on that trait, it is a critical point. If an image contains semantic information that is not conveyed through hue color, but rather a shape, for example, then a mismatch will most likely occur. Also if the brightness and saturation is crucial to the conveyance of a mood in a picture, it may possibly cause a mismatch since this system does not take saturation and brightness into consideration when mapping the dominant hue to an arousal-valence sector. If the hue color is green and is mapped to an arousal-valence sector is “calm,” and dark green and light green represent entirely different moods in a picture, then all pictures . Also pictures that have only one dominant, clear mood is ideal for matching. Ambiguous, ambivalent, or multiple moods are not considered during image processing but they will cause a mismatch in the user's opinion. Images that present a single, dominant mood based on a frequent hue range that falls within one mapped arousal-valence sector is the ideal image for the database in the system.

When matching to text, it is vital to map the group of text words to the appropriate arousal-valence sector(s); otherwise, a mismatch will occur. Also if a group contains contradicting feelings then it will retrieve all the words in that group and that will cause mismatches as well. It is best to create 16 or more groups of words with distinct emotions so that when the group of words are retrieved from the database, they do not present any contradictory emotions conveyed by the words. Hence, the scenarios this system works ideally with are when the text word groups are grouped and mapped appropriately.

Ultimately, the most crucial factor in this system is the mapping of the various values of each medium to the appropriate valence-arousal sector. The acceptable test cases described above are presented in the survey to the users; hence, it is expected that the system yield CEMT matches that are aesthetically and intuitively more preferred to the user than random matches.

5.2 Survey Results & Analysis

Due to the subjective nature of the concepts in this project, the testing of this system is conducted through a user survey. The user survey was presented in a web format and the results were saved to a database. The complete results are located in the Appendix A: Survey. Nine participants were given the same set of questions. The first question asked the user to rank his expertise in computer technology from 1 through 10, 1 being “Not Competent” and 10 being “Extremely competent.” The average competency was about 6 out of 10, and 3 out of 12 people ranked their competency as 5 or below, while 6 out of 12 people ranked their competency as 6 or higher. The users will further be referred on the name pattern of User #. The number is based on ascending order of their expertise. User 1 is the person with the lowest expertise score and User 9 is the user with the highest expertise score, see Figure 25.

The following set of questions were then broken into subsets regarding different activities. Within those subsets, there were even more subsets, each sub-subset pertained to a specific result set: the first being the CEM matches which are based solely on mood content and no text tags were considered, the second set considers the CEMT matches which are returned based on mood content and text tags, and lastly, the third set relates to the random matches; the size and contents of these matches are chosen randomly. Within those sub-subsets, the first question asks how many items were returned for the result set, then the user is prompted to specify the number of items that he thought matched the input, and finally he is then prompted to rate the match on a scale of 1 to 10, 1 being "No Match" and 10 being "Excellent Match." For example, a match that returns a large amount of results that only contain very few or no matches at all would be given a rating of 1. A score of 10 would be given for a match that is practically perfect in the user's opinion although there may be one or two slight mismatches in a large result set.

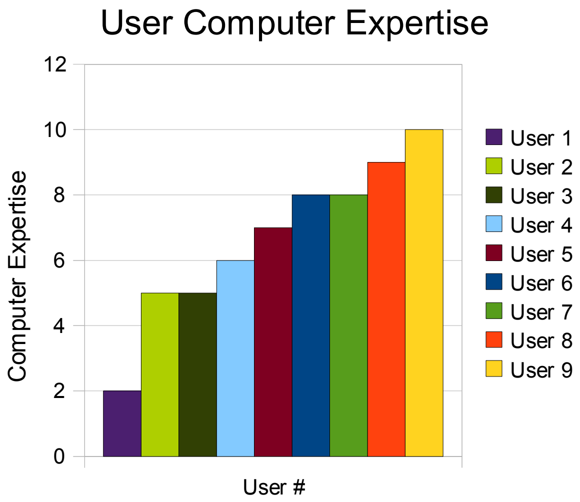


Figure 25: User computer expertise results

Within the set of survey questions, the first subset was aimed at walking the user through one particular medium matching: audio to image. This activity will return three different result sets as described above. The first CEM result set returned 5 items, on average the users said 5 items matched out of 5 (100%). They rated the overall matching experience to be 7 out of 10. The second CEMT result set returned 1 item, and on average the users said the 1 item matched the input, in fact all 9 users said the item matched, making it a perfect match. The users had an overall rating of 9 out of 10 for the CEMT results. The third random result set returned on average 11 items, on average the users said 5 matched (45.5%), but the overall experience rating is 4. The CEM and CEMT result sets faired better than the random set, but the CEMT set was rated higher and had a higher match rate than the CEM set which was expected. See figures 26, 27, 28, 29, and 30 for the results displayed visually.

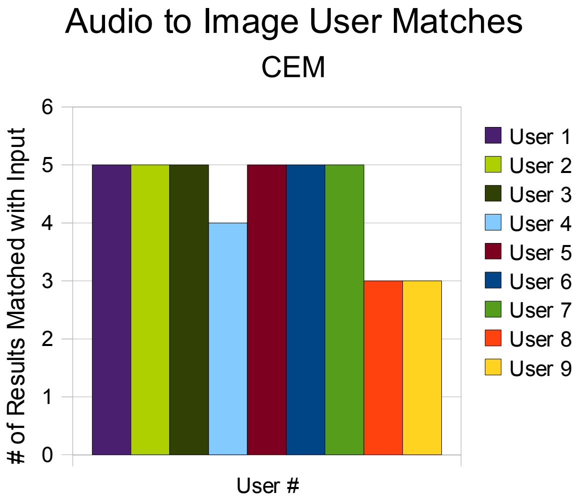


Figure 27: Audio to Image User Matches for CEM Set

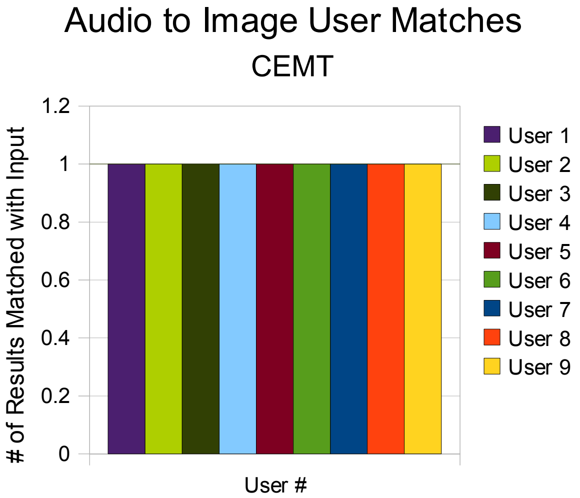


Figure 26: Audio to Image User Matches for CEMT Set

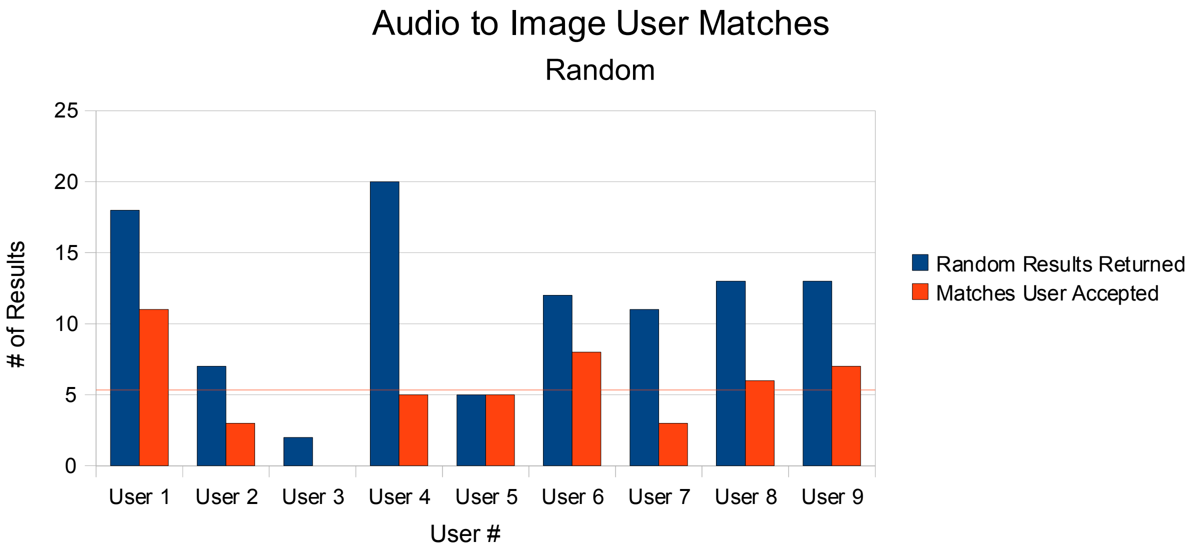


Figure 28: Audio to Image User Matches for Random Set

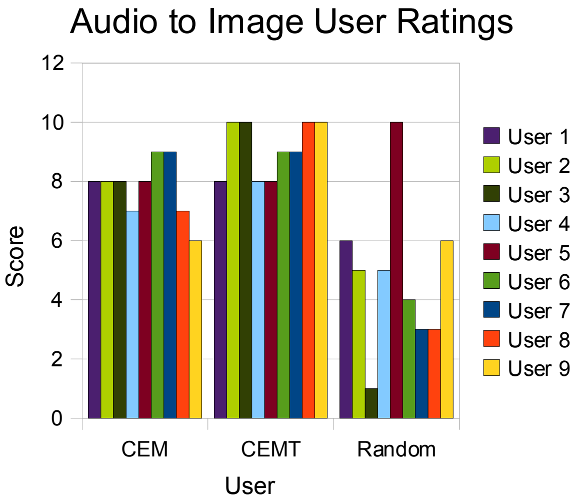


Figure 29: Audio to Image User Ratings Results

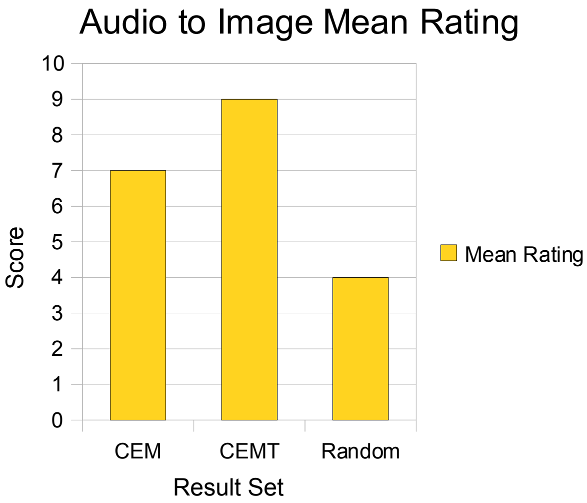


Figure 30: Audio to Image Mean Ratings Results

Within the set of survey questions, the second subset was aimed at walking the user through one particular medium matching: image to audio. This activity will return three different result sets as described above. The first CEM result set returned 3 items, on average the users said 2 items matched out of 3 (66.7%). They rated the overall matching experience to be 8 out of 10. The second CEMT result set returned 1 item, and on average the users said the 1 item matched the input, in fact all 9 users said the item matched, making it a perfect match again. The users had an overall rating of 9 out of 10 for the CEMT results. The third random result set returned on average 32 items, on average the users said 6 matched (18.6%), but the overall experience rating is 3. The CEM and CEMT result sets again faired better than the random set, but the CEMT set was rated higher and had a higher match rate than the CEM set which was expected. See figures 31, 32, 33, 34, and 35 for the results displayed visually.

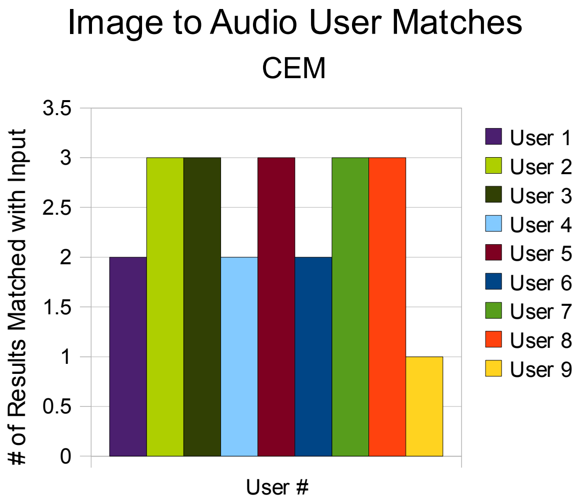


Figure 32: Image to Audio User Matches for CEM Set

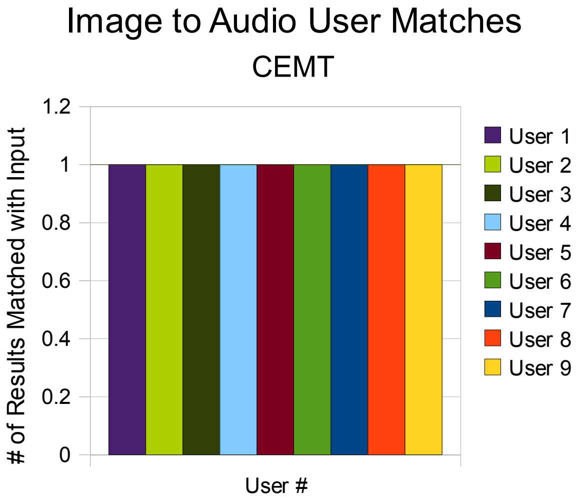


Figure 31: Image to Audio User Matches for CEMT Set

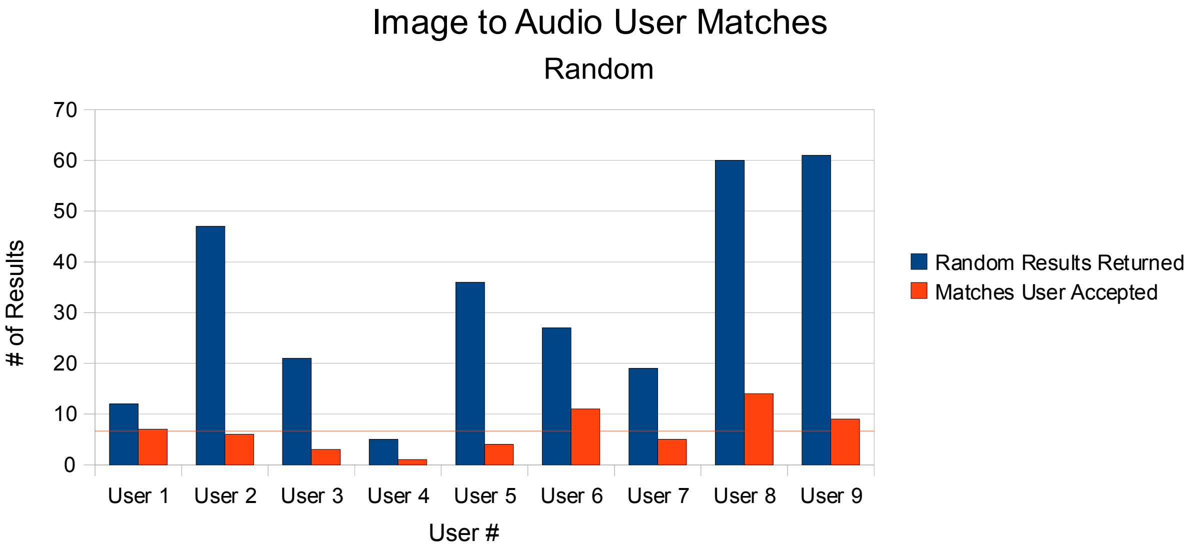


Figure 33: Image to Audio User Matches for Random Set

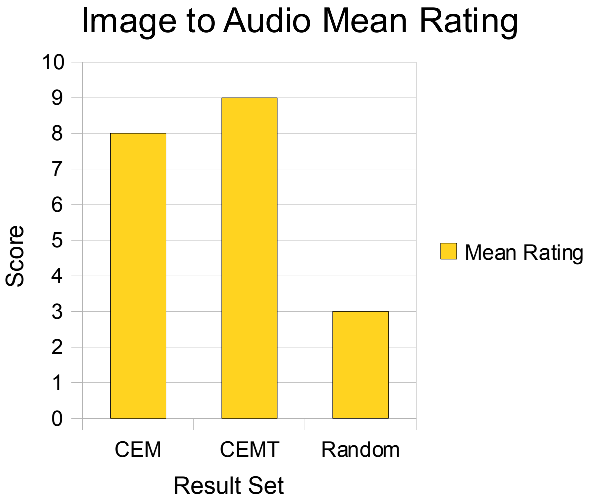


Figure 34: Image to Audio Mean Rating Results

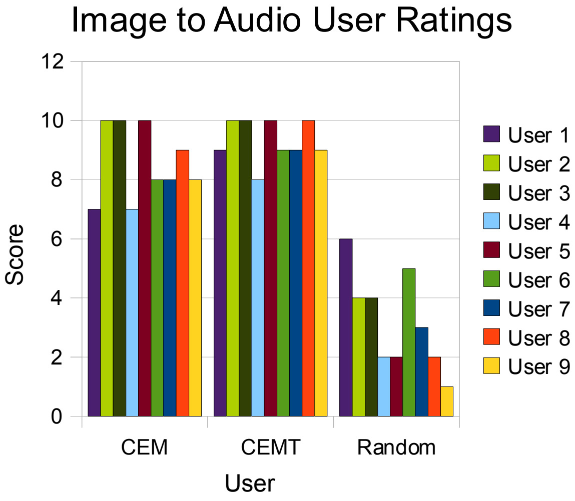


Figure 35: Image to Audio User Ratings Results

Within the set of survey questions, the third subset was aimed at walking the user through one particular medium matching: text to audio. This activity will return three different result sets as described above. The first CEM result set returned 34 items, on average the users said 15 items matched out of 34 (44%). They rated the overall matching experience to be 5 out of 10. The second CEMT result set returned 6 items, and on average the users said the 3 items matched the input (50%). The users had an overall rating of 6 out of 10 for the CEMT results. The third random result set returned on average 9 items, on average the users said 5 matched (55.6%), but the overall experience rating is 4. The CEM and CEMT result sets again faired better than the random set, but the CEMT set was rated higher and had a higher match rate than the CEM set which was expected. See figures 36, 37, 38, 39, and 40 for the results displayed visually.

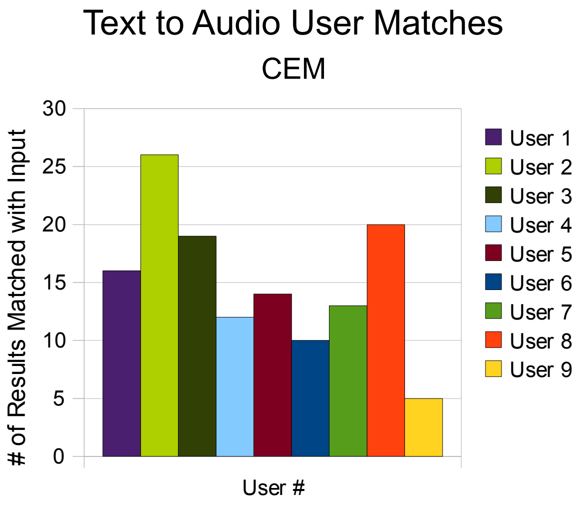


Figure 37: Text to Audio User Matches for CEM Set

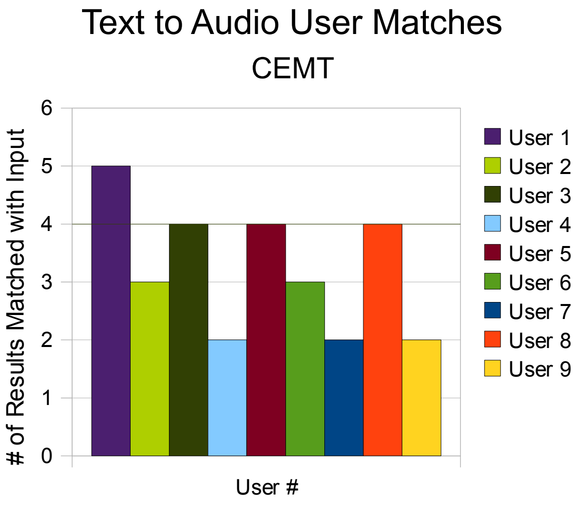


Figure 36: Text to Audio User Matches for CEMT Set

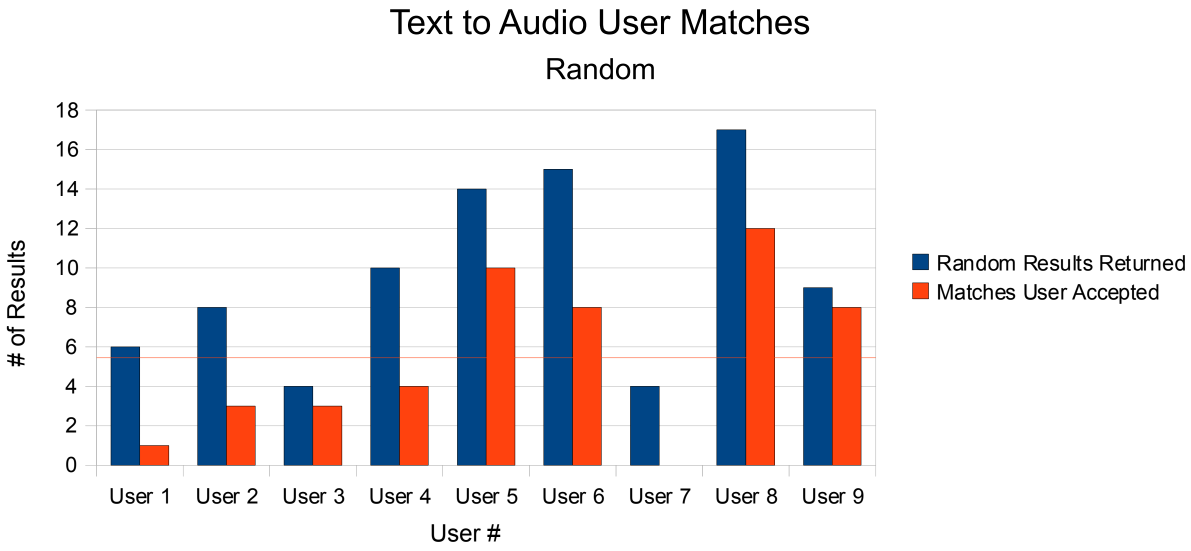


Figure 38: Text to Audio User Matches for Random Set

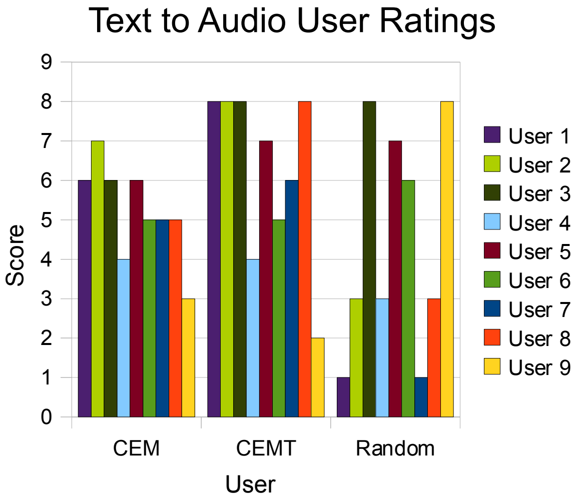


Figure 39: Text to Audio User Ratings Results

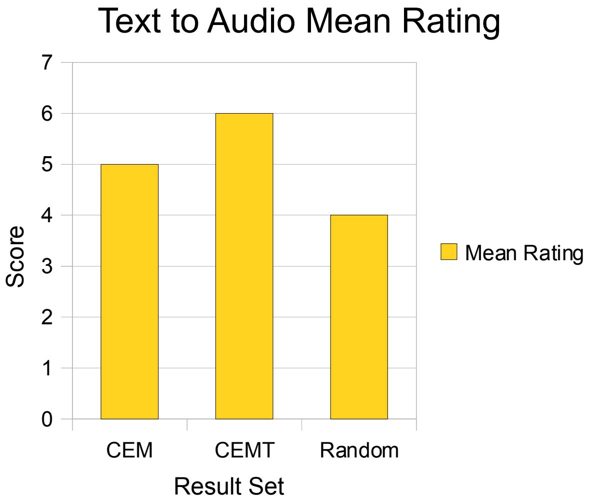


Figure 40: Text to Audio Mean Rating Results

Within the set of survey questions, the fourth subset was aimed at walking the user through one particular medium matching: text to image. This activity will return three different result sets as described above. The first CEM result set returned 21 items, on average the users said 12 items matched out of 21 (57.1%). They rated the overall matching experience to be 5 out of 10. The second CEMT result set returned 1 item, and on average the users said the 1 item matched the input, in fact all 9 users said the item matched, making it a perfect match again. The users had an overall rating of 10 out of 10 for the CEMT results. The third random result set returned on average 12 items, on average the users said 3 matched (25%), but the overall experience rating is 2. The CEM and CEMT result sets again faired better than the random set, but the CEMT set was rated higher and had a higher match rate than the CEM set which was expected. See figures 41, 42, 43, 44, and 45 for the results displayed visually.

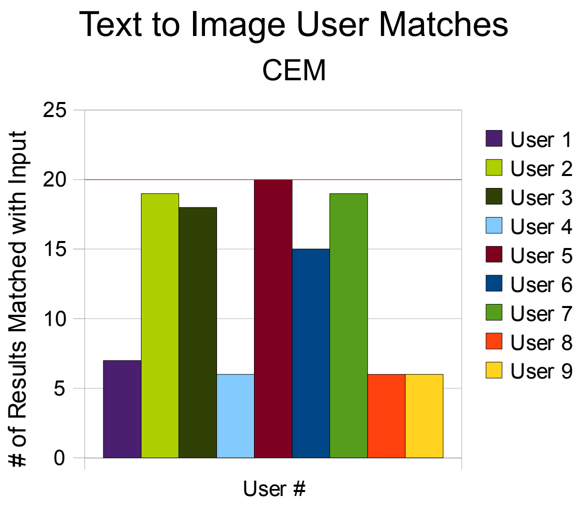


Figure 41: Text to Image User Matches for CEM Set

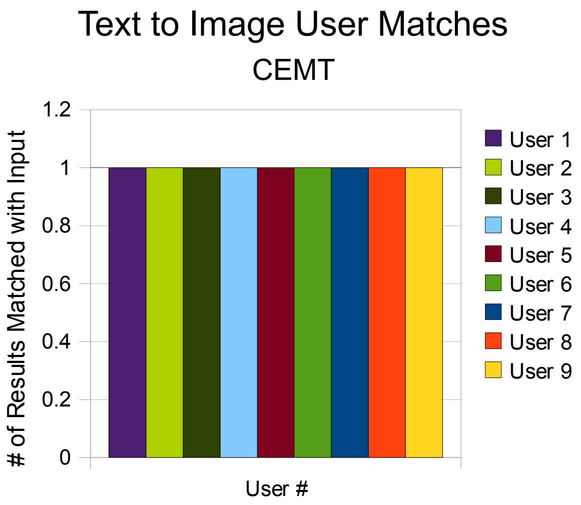


Figure 42: Text to Image User Matches for CEMT Set

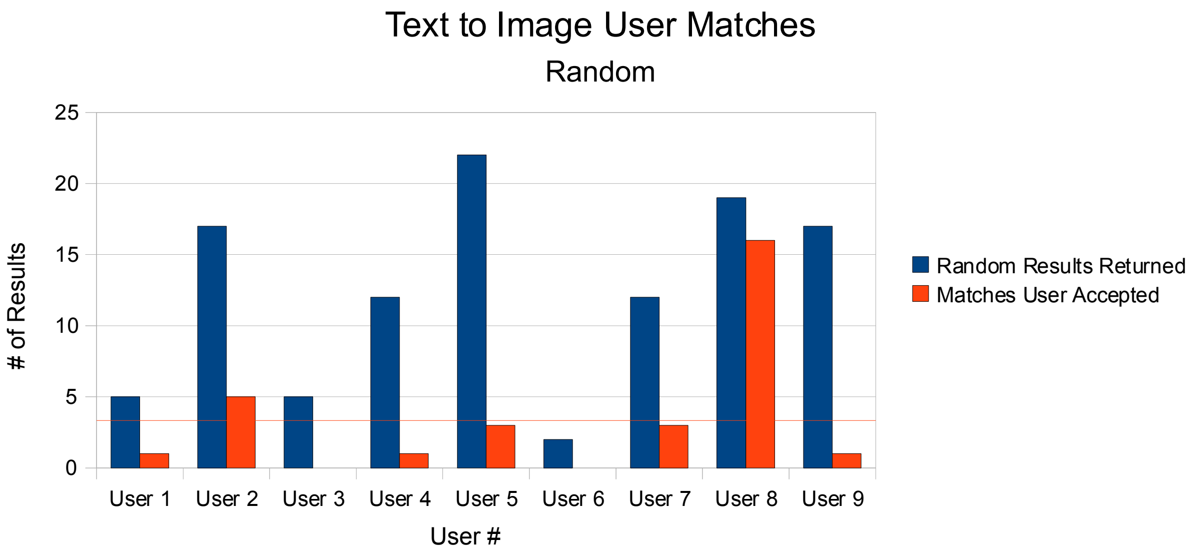


Figure 43: Text to Image User Matches for Random Set

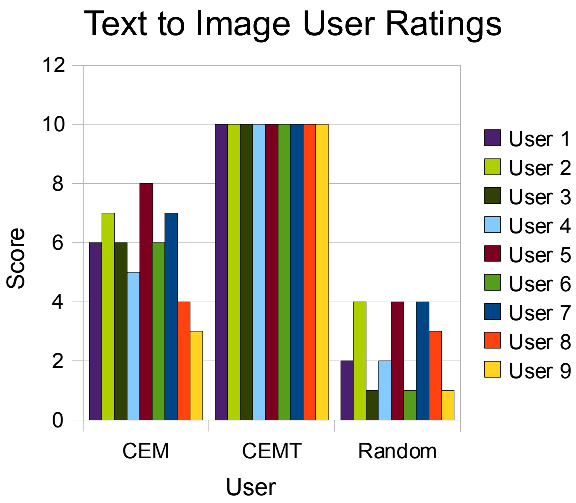


Figure 45: Text to Image User Ratings Results

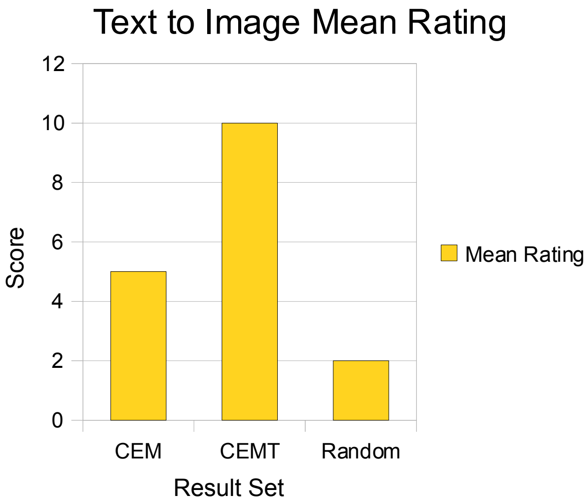


Figure 44: Text to Image Mean Rating Results

Within the set of survey questions, the fifth subset was aimed at walking the user through one particular medium matching: audio to text. This activity will return two different result sets as described above. The first CEM result set returned 97 items, on average the users said 79 items matched out of 97 (81.4%). They rated the overall matching experience to be 7 out of 10. The second random result set returned on average 12 items, on average the users said 1 matched, but the overall experience rating is 1 (8.3%). The CEM result set faired much better than the random set which was expected. See figures 46, 47, 48, and 49 for the results displayed visually.

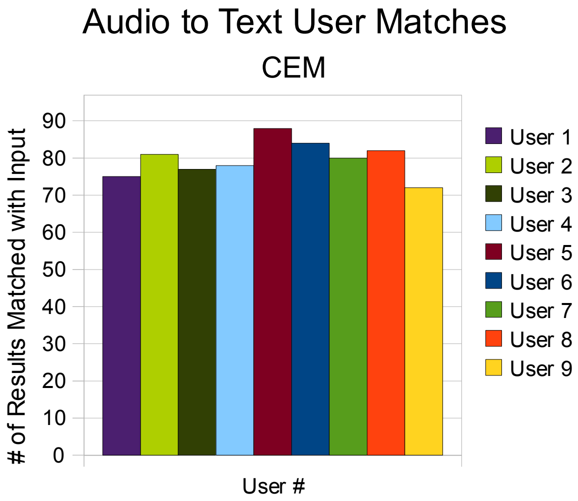


Figure 46: Audio to Text User Matches for CEM Set

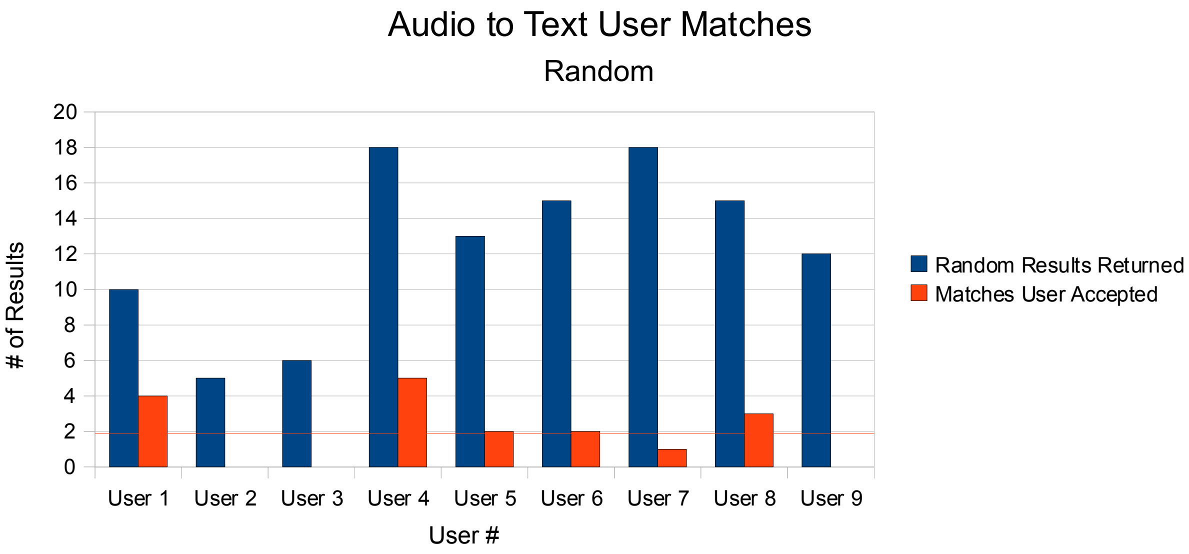


Figure 47: Audio to Text User Matches for Random Set

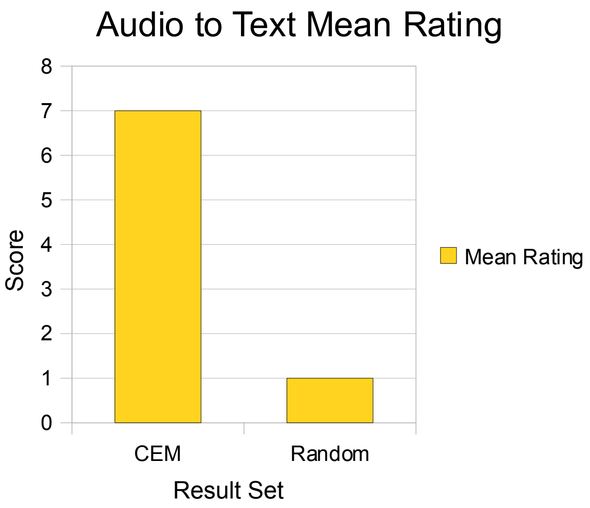


Figure 49: Audio to Text Mean Rating Results

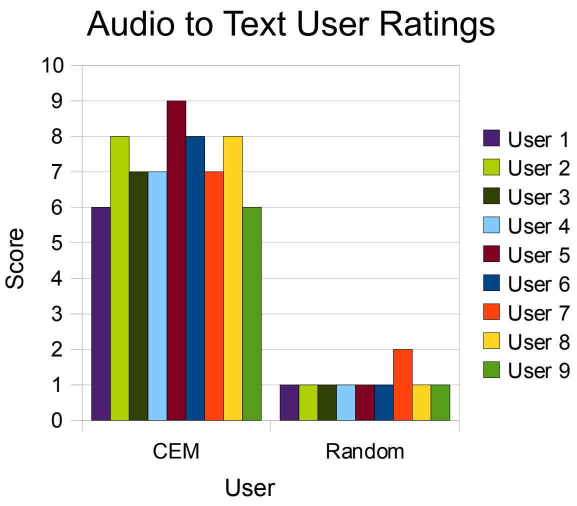


Figure 48: Audio to Text User Ratings Results

Within the set of survey questions, the sixth subset was aimed at walking the user through one particular medium matching: image to text. This activity will return two different result sets as described above. The first CEM result set returned 97 items, on average the users said 73 items matched out of 97 (75.3%). They rated the overall matching experience to be 6 out of 10. The second random result set returned on average 10 items, on average the users said 1 matched, but the overall experience rating is 1 (10%). The CEM result set faired much better than the random set which was expected. See figures 50, 51, 52, and 53 for the results displayed visually.

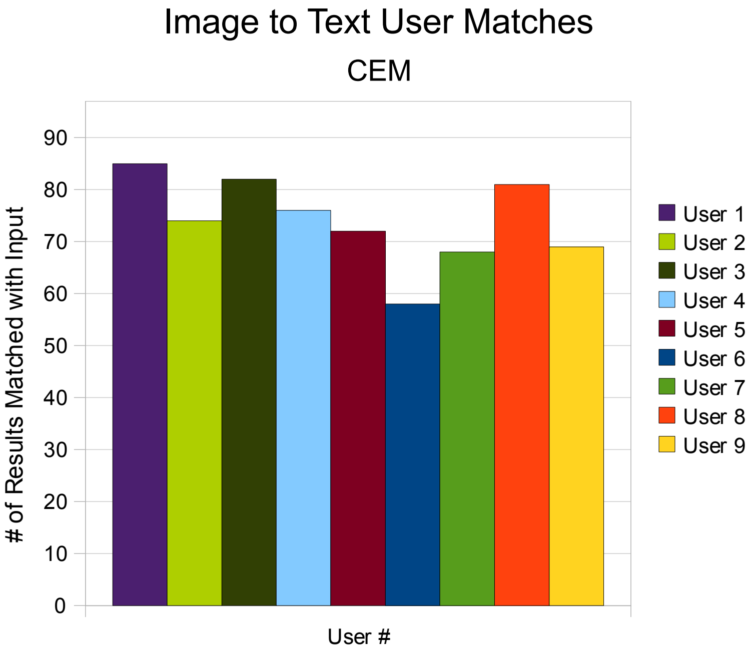


Figure 50: Image to Text User Matches for CEM Set

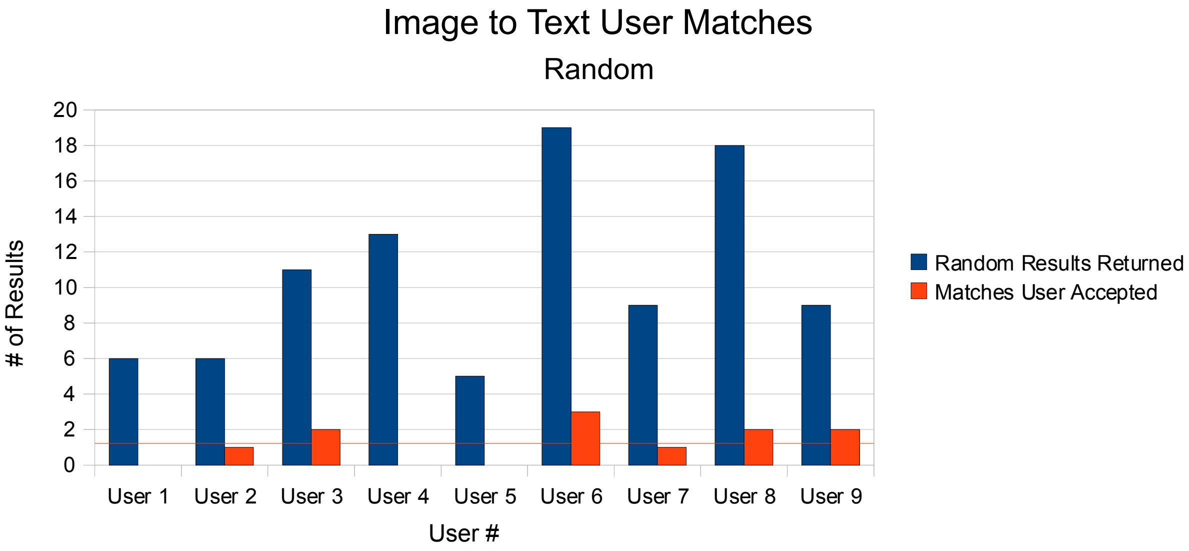


Figure 51: Image to Text User Matches for Random Set

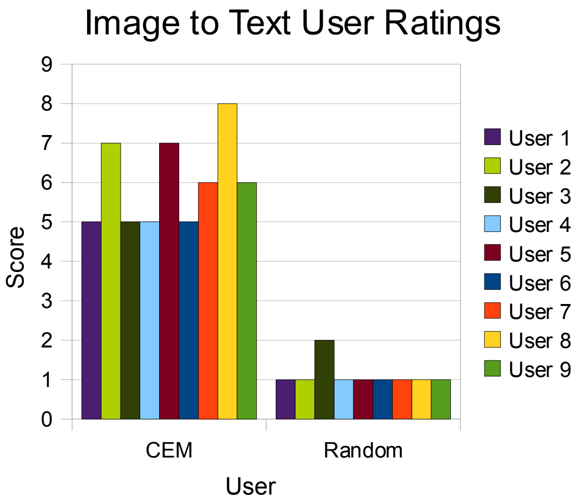


Figure 52: Image to Text User Ratings

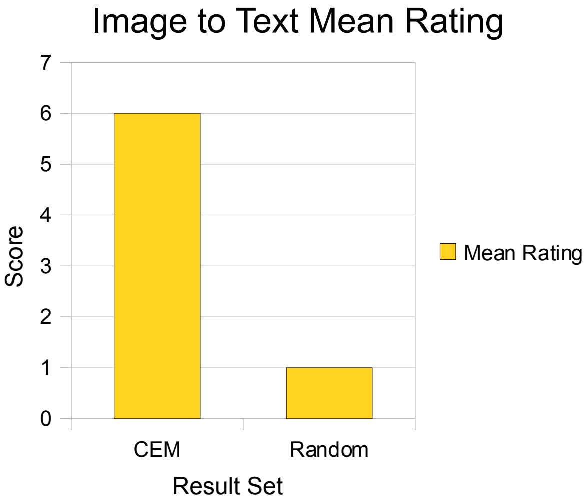


Figure 53: Image to Text Mean Rating

The users were then told to explore the rest of the system without any further instructions from the survey on how to run a match. The ninth and final subset of survey questions asked the which type of result set he preferred and how interesting the concept of matching distinct mediums is to him. All 9 users said they preferred the CEMT result set in Figure 54, no user preferred the CEM and random result sets. On average, the users gave the concept an 8 out of 10 in terms of interestingness, see Figure 55.

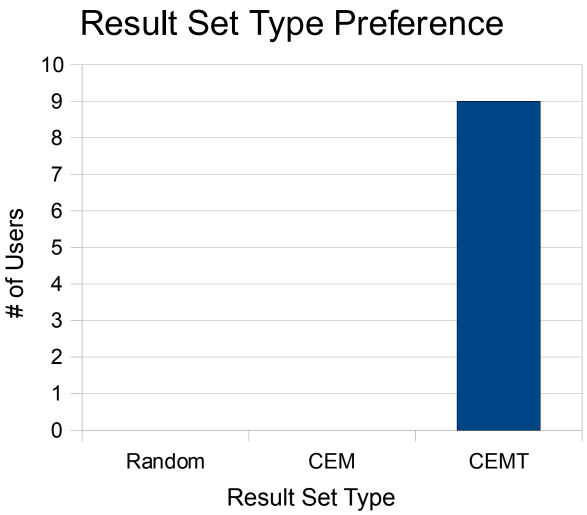


Figure 54: Result Set Type Preference

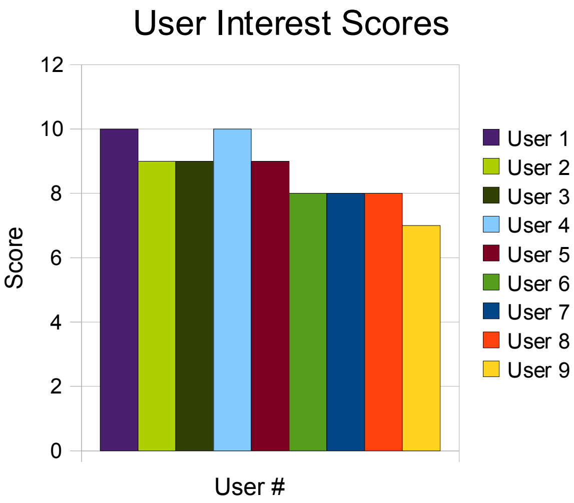


Figure 55: User Interest Scores

A previous study completed by Chen et al [5] showed that their audio to image match outperformed random selection by a small difference. See Figure 56 for their results. In CEM system's results, the delta between random and CEM is large and the delta between random and CEMT is also large.

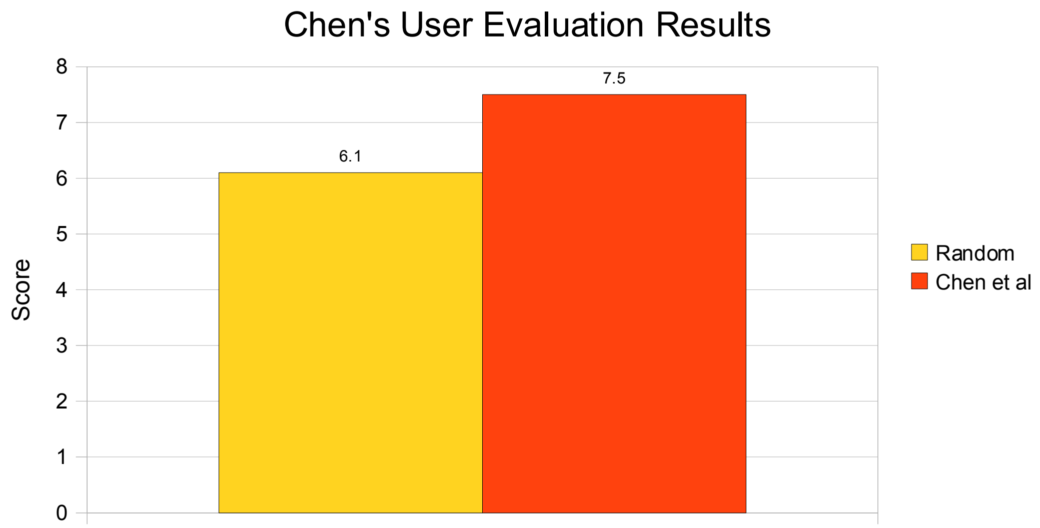


Figure 56: Chen et al[5]'s user results

In all of the 6 activities presented in the survey, the result set provided by the matching system is significantly better than random matching. Out of the four activities where CEMT result sets are available, these type of results are better than CEM and random result sets. Through the feedback of the users, it is known that this matching system is a better alternative for these types of test scenarios when compared to a system that depends on randomness. Also, the users think the concept of matching distinct mediums is fairly interesting, which shows that this has potential to be of use in the commercial sense.

6.0 Future Work

This project was completed by modifying and integrating several key components: audio analysis, image analysis, customizable mapping rules, text and tag filtering, and mood tag annotation and filtering. Most of the aforementioned components have its own field of study in which there are improvements discovered and invented regularly.

This project could expand to incorporate faster and improved mood extraction algorithms for image, audio, and text analysis and map the analysis results to the arousal-valence model in a similar manner as shown in this project. For example, a more selective training set could be created, or more or less audio features could be extracted depending upon the strength of the relationship between audio feature and emotion. For image analysis, shape, texture, and text extraction could be taken into account instead of just color since those traits play major roles in mood affectation. For text analysis, the database of words can be extended to include WordNet-Affect[28], a large database of words that are paired with synsets containing positivity, negativity, ambiguity, and neutrality values, and SentiWordNet[27], a large set of words that are associated with positive, negative, and objectivity values. The quantitative values could be incorporated when processing a final value that should be mapped on the arousal-valence model to provide more granularity than just mapping a single word to a sector.

A major component that can be added to this is the integration of video analysis, using similar methods along with additional considerations for audio and visual synchronization. The video would be broken up into audio and visual components and analyzed separately. Then an analysis would be done on the synchronization of the two components. Then the final results of the analyses would be mapped to the arousal-valence emotion model.

The user customizable rules could be expanded and restructured to accommodate more restrictions according to culture or whatever needs of the user. For example, the rules could allow for categorization according to culture, domains, a particular file, subgroups of files, and so forth. The project focused on color-to-mood rules but future work could allow audio-to-mood to be specified as well. The system could be expanded to allow many users and store their individual rules.

User-recommendation systems are popular and commonly utilized. These types of systems are used heavily in practical, e-commerce applications such as Amazon.com. If the user does not want to specify many rules, then they can specify a small set and then recommendations from other indirectly connected users with similar profiles can be used instead. Another method would be to explicitly add another user on a recommendation system and rely on the recommendations of those users.

The scope of the project could even expand to generate an entirely unique file based on the input. Each audio file can be considered to have a unique signature and so the generated output could match that unique signature with considerations of user input, so many different signatures could be generated from one single input. This would allow the advertising and marketing industry to market something in different ways for different targeted audiences. The integration and connection of distinct mediums based on moods has a significant amount of potential for innovation.

7.0 Conclusion

The field of mood extraction from mediums such as images, text, and audio have become increasingly useful in understanding human beings and how they relate to those mediums. Within this affect extraction field, there are many sub-fields for each medium because there are many ways of interpreting the mood of each medium, and within those sub-fields are even more fields focusing, elaborating, and improving upon certain techniques for mood extraction. To my knowledge, little has been done about connecting those techniques by utilizing a middle ground since affect extraction of a single medium is such weighty subject in and of itself.

This project demonstrated how it is possible to integrate several distinct mediums—audio, image, and text—by utilizing a common emotional, arousal-valence model by mapping certain zones and filtering the matching results by using any available textual information such as tags or lyrics which improved the overall matches in the user's opinion. Also, the accuracy of the matches was improved even further by supplementing the system with a user-centric, customizable set of parameters. Although the technique has room for expanding the mappings by increasing the amount of zones and other improvements, survey participants felt the concept was intriguing and useful for a variety of activities such as automatic retrieval of mediums when creating a family slideshow or e-commerce applications.

References

[1] Agarwal, A. and Meyer, A. 2009. Beyond usability: evaluating emotional response as an integral part of the user experience. In Proceedings of the 27th international Conference Extended Abstracts on Human Factors in Computing Systems (Boston, MA, USA, April 04 - 09, 2009). CHI EA '09. ACM, New York, NY, 2919-2930.

[2] Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. 2008. Can all tags be used for search?. In Proceeding of the 17th ACM Conference on information and Knowledge Management (Napa Valley, California, USA, October 26 - 30, 2008). CIKM '08. ACM, New York, NY, 193-202.

[3] Bischoff, K., Firan, C. S., and Paiu, R. 2009. Deriving music theme annotations from user tags. In Proceedings of the 18th international Conference on World Wide Web (Madrid, Spain, April 20 - 24, 2009). WWW '09. ACM, New York, NY, 1193-1194.

[4] Cai, R., Zhang, C., Wang, C., Zhang, L., and Ma, W. 2007. MusicSense: contextual music recommendation using emotional allocation modeling. In Proceedings of the 15th international Conference on Multimedia (Augsburg, Germany, September 25 - 29, 2007). MULTIMEDIA '07. ACM, New York, NY, 553-556.

[5] Chen, C.H., Weng, M.F., Jeng, S.K., and Chuang, Y.Y. Emotion-Based Music Visualization Using Photos. LNCS, 4903:358–368, 2008.

[6] Chen, L., Wright, P., and Nejdl, W. 2009. Improving music genre classification using collaborative tagging data. In Proceedings of the Second ACM international Conference on Web Search and Data Mining (Barcelona, Spain, February 09 - 12, 2009). R. Baeza-Yates, P. Boldi, B. Ribeiro-Neto, and B. B. Cambazoglu, Eds. WSDM '09. ACM, New York, NY, 84-93.

[7] Choi, Y. K., Choi, C., and Park, C. C. 2004. Image retrieval using color component analysis. In Proceedings of the 2004 international Symposium on information and Communication Technologies (Las Vegas, Nevada, June 16 - 18, 2004). ACM International Conference Proceeding Series, vol. 90. Trinity College Dublin, 214-219.

[8] Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2 (Apr. 2008), 1-60.

[9] Dunker, P., Nowak, S., Begau, A., and Lanz, C. 2008. Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach. In Proceeding of the 1st ACM international Conference on Multimedia information Retrieval (Vancouver, British Columbia, Canada, October 30 - 31, 2008). MIR '08. ACM, New York, NY, 97-104.

[10] Feng, Y., Zhuang, Y., and Pan, Y. 2003. Music Information Retrieval by Detecting Mood via Computational Media Aesthetics. In Proceedings of the 2003 IEEE/WIC international Conference on Web intelligence (October 13 - 17, 2003). Web Intelligence. IEEE Computer Society, Washington, DC, 235.

[11] [Emotion personality] Gebhard, P. 2005. ALMA: a layered model of affect. In Proceedings of the Fourth international Joint Conference on Autonomous Agents and Multiagent Systems (The Netherlands, July 25 - 29, 2005). AAMAS '05. ACM, New York, NY, 29-36.

[12] Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. 2008. The language of emotion in short blog texts. In Proceedings of the ACM 2008 Conference on Computer Supported

Cooperative Work (San Diego, CA, USA, November 08 - 12, 2008). CSCW '08. ACM, New York, NY, 299-302.

[13] Hanjalic, A., "Extracting moods from pictures and sounds: towards truly personalized TV," Signal Processing Magazine, IEEE , vol.23, no.2, pp.90-100, March 2006.

[14] Hou, X. and Zhang, L. 2007. Color conceptualization. In Proceedings of the 15th international Conference on Multimedia (Augsburg, Germany, September 25 - 29, 2007). MULTIMEDIA '07. ACM, New York, NY, 265-268.

[15] Kwon, J., Kim, H., Yoo, M., and Lee, I. 2009. Generating affective music icons in the emotion plane. In Proceedings of the 27th international Conference Extended Abstracts on Human Factors in Computing Systems (Boston, MA, USA, April 04 - 09, 2009). CHI EA '09. ACM, New York, NY, 3389-3394.

[16] Luo, J.; Hanjalic, A.; Tian, Q.; Jaimes, A., "Integration of Context and Content for Multimedia Management: An Introduction to the Special Issue," Multimedia, IEEE Transactions on , vol.11, no.2, pp.193-195, Feb. 2009

[17] Rho, S., Han, B., and Hwang, E. 2009. SVR-based music mood classification and context-based music recommendation. In Proceedings of the Seventeen ACM international Conference on Multimedia (Beijing, China, October 19 - 24, 2009). MM '09. ACM, New York, NY, 713-716.

[18] Saari, T., Ravaja, N., Laarni, J., Turpeinen, M., and Kallinen, K. 2004. Psychologically targeted persuasive advertising and product information in e-commerce. In Proceedings of the 6th international Conference on Electronic Commerce (Delft, The Netherlands, October 25 - 27, 2004). M. Janssen, H. G. Sol, and R. W. Wagenaar, Eds. ICEC '04, vol. 60. ACM, New York, NY, 245-254.

[19] Sebe, N. and Tian, Q. 2007. Personalized multimedia retrieval: the new trend?. In Proceedings of the international Workshop on Workshop on Multimedia information Retrieval (Augsburg, Bavaria, Germany, September 24 - 29, 2007). MIR '07. ACM, New York, NY, 299-306.

[20] Strapparava, C. and Mihalcea, R. 2008. Learning to identify emotions in text. In Proceedings of the 2008 ACM Symposium on Applied Computing (Fortaleza, Ceara, Brazil, March 16 - 20, 2008). SAC '08. ACM, New York, NY, 1556-1560.

[21] Wang, Y., Kan, M., Nwe, T. L., Shenoy, A., and Yin, J. 2004. LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics. In Proceedings of the 12th Annual ACM international Conference on Multimedia (New York, NY, USA, October 10 - 16, 2004). MULTIMEDIA '04. ACM, New York, NY, 212-219.

[22] Westerman, S. J. and Kaur, S. 2007. Supporting creative product/commercial design with computer-based image retrieval. In Proceedings of the 14th European Conference on Cognitive Ergonomics: invent! Explore! (London, United Kingdom, August 28 - 31, 2007). ECCE '07, vol. 250. ACM, New York, NY, 75-81.

[23] Xia, Y., Wang, L., Wong, K., and Xu, M. 2008. Sentiment vector space model for lyric-based song sentiment classification. In Proceedings of the 46th Annual Meeting of the Association For Computational Linguistics on Human Language Technologies: Short Papers (Columbus, Ohio, June 16 - 17, 2008). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 133-136.

[24] Yang, Y., Liu, C., and Chen, H. H. 2006. Music emotion classification: a fuzzy approach. In Proceedings of the 14th Annual ACM international Conference on Multimedia (Santa Barbara, CA, USA, October 23 - 27, 2006). MULTIMEDIA '06. ACM, New York, NY, 81-84.

[25] Yang, Y., Su, Y., Lin, Y., and Chen, H. H. 2007. Music emotion recognition: the role of individuality. In Proceedings of the international Workshop on Human-Centered Multimedia (Augsburg, Bavaria, Germany, September 28 - 28, 2007). HCM '07. ACM, New York, NY, 13-22.

[26] http://searchengineland.com/bing-2-0-unveiled-visual-search-25703

[27] http://sentiwordnet.isti.cnr.it

[28] http://wndomains.itc.it/wnaffect.html

[29] http://en.wikipedia.org/wiki/HSL_and_HSV#Conversion_from_RGB_to_HSL_overview

[30] http://marsyas.sness.net/

[31] http://www.codeproject.com/KB/GDI-plus/HSLColorSpace/hue.jpg

[32] http://www.crayonroom.com/moody.php

[33] http://www.spezify.com

[34] http://www.gazopa.com