

NAIST-IS-MT1551126

Master's Thesis

EmoBGM : Emotion-aware BGM recommendation for creating impressive slideshow

Konan N'djabli Cedric Ange

March 8, 2017

Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Konan N'djabli Cedric Ange

Thesis Committee:

Professor Keiichi Yasumoto	(Supervisor)
Professor Kazushi Ikeda	(Co-Supervisor)
Associate Professor Yutaka Arakawa	(Co-supervisor)
Assistant Professor Hirohiko Suwa	(Co-supervisor)
Assistant Professor Manato Fujimoto	(Co-supervisor)

EmoBGM : Emotion-aware BGM recommendation for creating impressive slideshow*

Konan N'djabli Cedric Ange

Abstract

Music is largely used in ceremonies, in movies or for therapeutic purposes. Its abilities to influence humans mood are highly important in our lives. However, not every music suits to every event. A certain music, selected based on the tempo, the beat, the instrument used, the lyrics etc. is more suitable for a given event than another one. This fact is easily visible while creating a slideshow from a personal photo as a memory for an event. The match of the images and the background music accompanying it must be perfect to keep intact the memory of the scene lived. This thesis presents a study about estimating the emotions conveyed in clips of background music (BGM) to be used in an automatic slideshow creation system. The system we aimed to develop, automatically tags each given pieces of background music with the main emotion it conveys, in order to recommend the most suitable music clip to the slideshow creators, based on the main emotions of embedded photos. As the first step of our research, we developed a machine learning model based on audio features such as MFCCs, LPC, Beat sum etc. to estimate the emotions conveyed in a music clip and achieved 88% classification accuracy with cross-validation technique. The second part of our work involved developing a web application using Microsoft Emotion API to determine the emotions in photos, so the system can find the best candidate music for each photo in the slideshow. In our evaluation experiment, 16 users rated the recommended background music for 3 set of photos using a 5-point Likert scale

*Master's Thesis, Graduate School of Information Science,
Nara Institute of Science and Technology, NAIST-IS-MT1551126, March 8, 2017.

and we achieved an average rate of 4.1, 3.6 and 3.0 for the photo sets 1, 2, and 3 respectively.

Keywords:

Emotion estimation, Sentiment, Background music assignment, Photo slideshow, Machine learning, Random forest, image and music matching

Contents

Contents	iii
List of Figures	v
List of Tables	vi
1 Introduction	1
2 Related work and challenges	3
2.1 Music’s emotion recognition approaches	3
2.1.1 Combining audio and lyrics	3
2.1.2 Combining audio and tags	4
2.1.3 Combining audio and images	5
2.2 Existing commercial system	5
2.3 Challenges	6
2.3.1 Understanding human feelings	6
2.3.2 Influence of Music’s feature on the emotion conveyed	6
3 Proposed method	7
3.1 Background Music’s emotion model creation	8
3.1.1 Tagging task (A)	9
3.1.2 Audio features extraction (B)	12
3.1.3 EmoBGM (C)	14
3.2 Implementation of the evaluation application	15
3.2.1 Microsoft Emotion Api Evaluation (D)	16
3.2.2 Best background music selection	17

4	Evaluation	20
4.1	Evaluation scenario	20
4.2	Evaluation results	22
5	Discussion	23
5.1	Discussion on the experimental result	23
5.2	Discussion	24
5.2.1	Approaches	24
5.2.2	System enhancement	24
	Subclasses	24
	Google Cloud Vision	25
6	Conclusion	26
	References	30
	Publication List	33

List of Figures

3.1	Proposed method	7
3.2	Summary of music model creation process	8
3.3	Plutchik's wheel of emotions	9
3.4	Tagging task application	13
3.5	Example of result given by Microsoft Emotion API	16
3.6	Microsoft Emotion Api evaluation app	17
3.7	EmoBGM evaluation App	19
4.1	Evaluation set 1	20
4.2	Evaluation set 2	21
4.3	Evaluation set 3	21
4.4	Average rating (on a scale of 5) for how appropriate the recommended BGM is for the set of pictures	22
5.1	Screenshot of google vision api output	25

List of Tables

2.1	Common acoustic feature types for emotion classification.	3
3.1	Ekman's list of emotion	10
3.2	Parrot's list of emotion	11
3.3	List of emotion tags	12
3.4	Comparison of result of classification using different algorithm . .	15
3.5	Microsoft Emotion API Evaluation task result	18

Chapter 1

Introduction

Music is ubiquitous in everyday life. It is in the background in movies, supermarkets, elevators etc. as we go about our everyday tasks. Some studies show that music can provoke both positive and negative behaviors [1] [11] and promote attention and wakefulness when performing tasks. It can also affect how we perceive the emotions observed in an image or video. For instance, people viewing photographs of smiling faces while listening to happy music will rate the emotions observed in the photos as happier than those viewing the photos without background music. This ability of music to enhance the emotional impact of a photo or a scene has led to its prevalence in movies, and more and more people are creating slideshows of their favorite photos set to background music as mementos.

In order to preserve the emotional memories of the events in photos, we need to embed music that conveys a similar emotion to the one depicted in the images. However, how can we accurately choose this suitable background music for the perfect video memory? This is the main question we aim to answer in this study.

Various researchers have focused on both assigning emotions to pieces of music and extracting emotions from media files. The used techniques have ranged from textual analysis [19], audio patterns and sound frequency analysis [8] to image processing combined with machine learning, and have resulted in high levels of classification accuracy. However, the studies do not address the problem of matching images and background music based on emotions conveyed. We build upon this work by developing a system that combines both emotion extraction

from images and emotion assignment to music, to recommend appropriate music clips for a slideshow of pictures. Our work focuses on two factors: (i) how to assign background music clips despite the absence of lyrics or user’s annotation and (ii) user evaluation of the system.

We started our work with a tagging task where 250 participants tagged 50 music clips from a set of 14 different emotions. We developed EmoBGM, an emotion estimation model for background music, using audio features (MFCC, LPC, Beat sum etc.) that were extracted with JAudio [10] and the emotion previously tagged. We then conducted a subjective evaluation of the resulting model. 16 users rated, using a 5-likert scale, the recommended background music for 3 sets of photos we provided. The background music database, embedded in the evaluation system, had been analyzed by our model (EmoBGM) and tagged automatically with emotion tags. The users rated the recommendations with an average of 4.1, 3.6 and 3.0 for the photo sets 1, 2, and 3 respectively, indicating a good level of appropriateness.

The rest of this thesis is arranged as follows. Chapter 2 presents some related researches followed by a description of the challenges of our study. Chapter 3 then presents the methodology used in this study. After explaining the evaluation part in chapter 4, we discuss the results in chapter 5 before concluding this thesis in Chapter 6.

Chapter 2

Related work and challenges

2.1 Music’s emotion recognition approaches

Various approaches have been used to determine the main emotion provoked by a piece of music [7]. They ranged from the usage of acoustic features [4] [16] [15] to Audio-content based method [14] [18]. An overview of the most common acoustic features used for mood recognition is given in Table 2.1.

Table 2.1: Common acoustic feature types for emotion classification.

Type	Features
Dynamics	RMS Energy
Timbre	MFCCs, spectral shape, spectral contrast
Harmony	Roughness, harmonic change, key clarity, majoriness
Register	Chromagram, chroma centroid and deviation
Rhythm	Rhythm strength, regularity, tempo, beat histograms
Articulation	Event density, attack slope, attack time

The most promising approaches are the one combining two kind of features.

2.1.1 Combining audio and lyrics

Y. Yang et al. combined lyrics and audio features using a database of 1240 Chinese pop songs [20]. The songs were hand-labeled according to one of the

four valence-arousal quadrants of the RussellThayer model, and acoustic features (MFCC and spectral descriptors) were computed for a 30-second segment drawn from the middle of each song. Lyrics (assumably from the entire song) were text-analyzed using the BOW approach. This work compared three different methods of combining audio and text features, obtaining the best results by first using audio and text separately to classify arousal and valence, respectively, using SVMs and then merging the results to determine a full V-A classification. On an 80/20 training-testing split of the data with 1000- fold cross-validation, acoustic features alone resulted in a baseline of 46.6% correct classification, while combining audio and lyrics yielded 57.1% accuracy (a relative performance increase of 21%)

Schuller et al. investigated using audio features, lyrics, and metadata to automatically label music in a discretized version of the V-A space [17]. For a database of 2648 pop songs, each was rated by four listeners who selected one of 5 discrete labels. Their classification task, is ultimately reduced to two independent three-class problems. Their best performing system made use of feature selection algorithms and label filtering, achieving 64.1% and 60.9% on valence and arousal, respectively.

2.1.2 Combining audio and tags

Music-IR researchers have also focused on multimodal approaches incorporating tags and low-level audio features for classification tasks.

Bischoff et al. combine social tag information and audio content-based analysis specifically for the task of emotion recognition [2]. For each of 4,737 songs, they collect social tags from Last.fm and generate a 240-dimensional audio feature vectors (including MFCCs, chroma features, and other spectral features). They then train a naive Bayes classifier for the social tags and an SVM for the audio feature vectors, combining them using a simple weighted combination approach. In one experiment, this approach is used to predict one of the five MIREX mood clusters. In the second experiment, the approach is used to predict one of the four quadrants of the V-A mood space. Ground truth for each track is based on a manually-determined (ad-hoc) mapping between one of 178 mood tags to a MIREX mood cluster and to a V-A quadrant. In both experiments, the multi-modal approach demonstrates better performance than either the tag-based and

audio-based approaches alone.

2.1.3 Combining audio and images

Analysis of audio features of music in combination with associated images (album covers, artist photos, etc.) for Music-IR tasks has very recently sparked research interest. Dunker et al. investigated methods to combine music and image domains for mood classification [3]. Their work investigates both classifying audio paired with images in a multimodal media tagging approach as well as trying to pair audio and images that have been tagged separately. More recently, Libeks and Turnbull analyze promotional photographs of artists using content-based image annotation [9]. Although they label these artists with genre tags (provided by Last.fm), it would be straightforward to adapt their approach to use emotion tags.

The studies mentioned above tackled some challenges we faced in our work. However, the main difference with our work lies in the fact that we focus on background music which does not have lyrics. Therefore, we have to focus only on core audio features such as MFCC-OSD (Mel-Frequency Cepstral Coefficients Overall Standard Deviation), Area method of moments of MFCCs OSD and Area Method of Moments of MFCCs Overall Average. We extracted these features using the feature extractor JAudio [10]. In addition, as we are targeting recommendation of background music for personal photos, we want to accurately guess the sentiment without relying on comments or annotations to limit the user burden.

2.2 Existing commercial system

Some applications that are currently available in app markets (Google play or Apple App store) try to achieve a similar aim to ours by creating automatic slideshows from personal photos accompanied by background music. Such applications include Google photos* which curate user's photos into beautiful videos accompanied by a generic piece of instrumental music that might go along with any kind of photos. They generally provide the option to change the background

*<https://play.google.com/store/apps/details?id=com.google.android.apps.photos>

music in case of a mismatch but this may bring some burden to the user who is looking for a good video rapidly. In addition, the use of the same generic song for various videos may remove the emotional memories which the user wishes to keep, and may also bore the user. We enhance the services and options provided by such applications through considering emotion.

2.3 Challenges

2.3.1 Understanding human feelings

Due to the extreme heterogeneity of human sentiment, it is particularly difficult to guess what a person feels while listening to a certain music. For example, the person's feeling can be influenced by his cultural background and his current state of mind. However, an ethnomusicology study of people with no exposure to Western music (or culture), showed that Mafa natives of Cameroon categorized music examples into three categories of emotion in the same way as Westerners; demonstrating the universality of emotion in music [6]. Also, taking the example of movies, we notice that the director always manages to select suitable music to accompany each scene, and therefore it is possible for humans to easily recognize what music suits what types of scenes. We exploit this insight in our research.

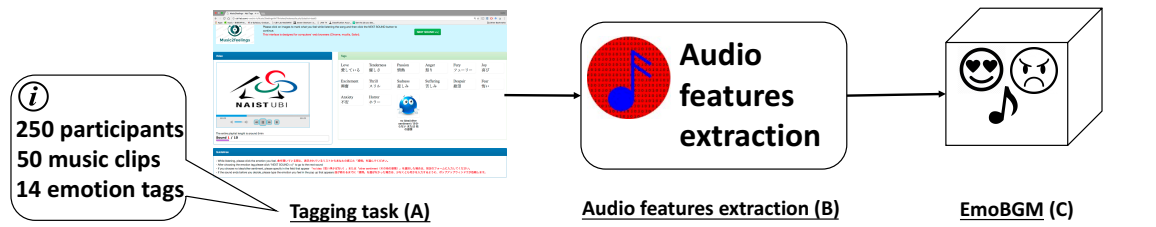
2.3.2 Influence of Music's feature on the emotion conveyed

Some other problems are: what features to use in order to make the most accurate model possible? how to determine those features and what features, among the hundreds of music features available, influence the most the emotion in the music? We also face the challenge to determine how to match images and music in our final application.

Chapter 3

Proposed method

Background Music's emotion model creation



Final system architecture

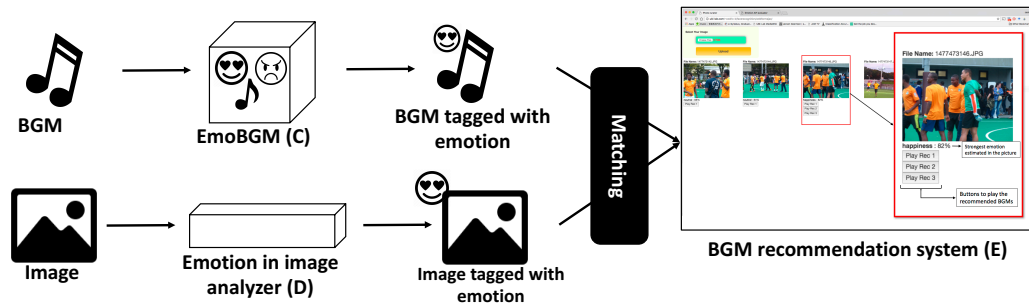


Figure 3.1: Proposed method

In order to achieve automatic background music assignment to a slideshow of images, the following issues need to be addressed: (i) tagging each piece of BGM music with an emotion tag and (ii) determining the main emotion conveyed in a scene. In our research, we focus on developing a prediction model for the former and use already available tools to perform the latter. Figure 3.1 shows the method

we used to solve our problem.

3.1 Background Music's emotion model creation

The first part of our proposed method dealt with the method to create a background music emotion model. We conducted an experiment where users tagged a set of background music with a list of emotions that we provided. Then, we extracted audio features from the music files and used them to create a model via machine learning. Figure 3.2 presents a detailed summary of the background music model creation process.

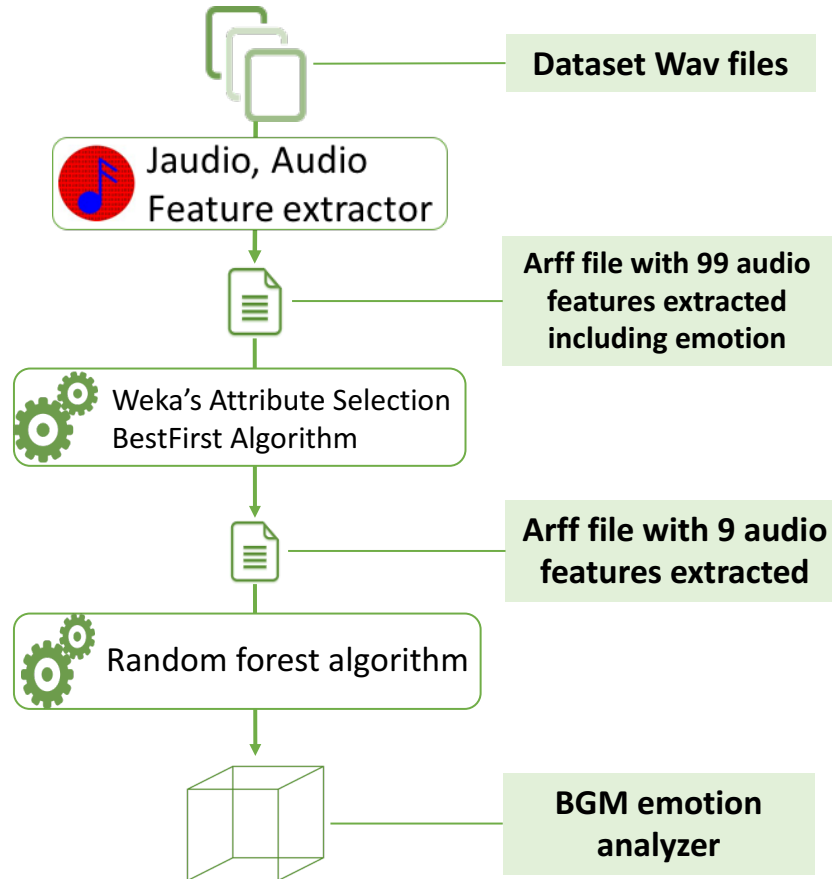


Figure 3.2: Summary of music model creation process

3.1.1 Tagging task (A)

To tag a music piece with an emotion, we first needed to determine the main emotions people feel when listening to various kinds of music scores.

Several researches classified the human emotion into categories. In 1980, Robert Plutchik developed a theory saying that the eight (8) basic human's emotions are: Fear, Anger, Sadness, Joy, Disgust, Trust, Anticipation and Surprise [13]. The wheel resulting of his research shows what is the result of the combination of two basic emotions. For example, according to Plutchik, a mix of disgust and anger make a person feels contempt as shown in figure 3.3

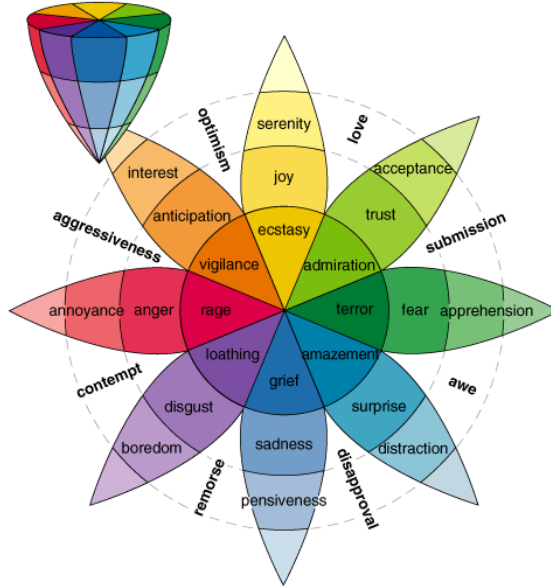


Figure 3.3: Plutchik's wheel of emotions

Paul Ekman released in 1972 a list of basic emotion from a transcultural research on a Papua New Guinea tribe. He concluded that 6 basic emotions (Sadness, Anger, Joy, Fear, Disgust, Surprise) are either biological or belongs to all humans without distinctions. In 1987 Ekman extended the list including a range of positive and negative emotions that are not all encoded in facial muscles [5]. Table 3.1 shows the Ekman's list of emotion.

Table 3.1: Ekman’s list of emotion

List of emotions released in 1972	List of emotions added in 1987
Sadness	Amusement
Anger	Contempt
Joy	Contentment
Fear	Embarrassment
Disgust	Excitement
Surprise	Guilt
	Pride in achievement
	Relief
	Satisfaction
	Sensory pleasure
	Shame

Parrot [12] divided them into 6 primary emotions (Love, joy, surprise, anger, sadness and fear). These primary emotions are then split into secondary and tertiary emotions, which results in a list of 100+ emotions. The full list of Parrott’s classification of emotion is shown in Table 3.2. This extensive list provides a comprehensive and easily understandable structure of human emotion. In an initial survey, we found that people could more easily tag music when provided with the comprehensive list than when provided with the basic emotion classes alone. We, therefore, used Parrot’s list for our work.

Table 3.2: Parrot's list of emotion

Primary emotion	Secondary emotion	Tertiary emotions
Love	Affection	Adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Arousal, desire, lust, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, zest, excitement, thrill, exhilaration
	Contentment	Contentment, pleasure
	Pride	Pride, triumph
	Optimism	Eagerness, hope, optimism
	Relief	Relief
Surprise	Surprise	Amazement, surprise, astonishment
Anger	Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness
	Exasperation	Exasperation, frustration
	Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment
	Disgust	Disgust, revulsion, contempt
	Envy	Envy, jealousy
	Torment	Torment
Sadness	Suffering	Agony, suffering, hurt, anguish
	Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, disappointment, displeasure
	Shame	Guilt, shame, regret, remorse
	Neglect	Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, embarrassment, humiliation, insult
	Sympathy	Pity, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread

Presenting users with 100+ emotions as tags can lead to confusion and make the task tedious. We, therefore, sought to identify the most frequently felt emotions and remove the others from the list. We asked 6 native English speakers in our university to listen to a music database and highlight the emotions they thought best expressed the feelings they felt while listening to each clip. The result of this task gave us 14 emotion tags listed in table 3.3 that we used in our research.

Table 3.3: List of emotion tags

Primary emotions	Secondary emotions
Love	Love, Tenderness, Passion
Anger	Anger, Fury
Happiness	Joy, Excitement , thrill
Sadness	Sadness, Suffering,Despair
Fear	Fear, Anxiety, Horror

The sound dataset we used in our research is composed of 50 music files split into 5 set of 10. They are background music of famous movies of different genres: Comedy, Action, Animation, Drama etc. We cut the background music with the rule that one background music must cover only one scene, with no lyrics and should last for less than 40 secs.

We used Amazon Mechanical Turk to recruit 250 turkers that will help us to tag the background music dataset. Via a web application that we developed (Figure 3.4), we asked the turkers to listen to background music in their set and to select the appropriate emotions from the 14 emotions listed. Each music clip could be tagged with a maximum of 3 emotions. The music looped until they choose at least one tag and validated their choice by clicking the "Next Song" button. We targeted 50 tags per song (i.e. 50 workers).

3.1.2 Audio features extraction (B)

After the tagging task, we extracted the acoustic features (MFCC, LPC, Beat sum etc.) of each music file using JAudio.

We used only acoustic features for 2 main reasons. First, to allow users to



Figure 3.4: Tagging task application

include instrumental music pieces including their own compositions which lack textual information and second because prior research has shown that acoustic information alone does provide a high accuracy of recognition of emotion in music pieces [14].

Some of the features among the 99 extracted are described below:

- Mel frequency cepstral coefficients: MFCC is calculated based on the short time Fourier transform (STFT). First, log-amplitude of the magnitude spectrum is taken, followed by grouping and smoothing the fast Fourier transform (FFT) bins according to the perceptually motivated Mel-frequency scaling. Jaudio provided all of the 13 coefficients, among which 4 were found to produce the best classification result.
- Spectral flux: Spectral flux is defined as the squared difference between the normalized magnitudes of successive windows:

$$F_i = \sum_{n=1}^n (N_t[n] - N_{t-1}[n])^2 \quad (3.1)$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitudes of the Fourier transform at the current frame t and the previous frame $t-1$, respectively. The spectral flux represents the amount of local spectral change.

- **Spectral Centroid:** Spectral Centroid is the center of gravity of the magnitude spectrum of the STFT. Here, $M_i[n]$ denotes the magnitude of the Fourier transform at frequency bin n and frame i . The centroid is used to measure the spectral shape. A higher value of the centroid indicates brighter textures with greater frequency. The spectral centroid is calculated as follows:

$$C_i = \frac{\sum_{n=0}^n n M_i[n]}{\sum_{n=0}^n M_i[n]} \quad (3.2)$$

- **Beat sum:** This feature is measured as the sum of all entries in the beat histogram. It is a very good measure of the importance of regular beats in a signal.

In our final recommendation system, the resulting classification model is used to tag a library of candidate songs with emotions. We use Microsoft Emotion API to detect the emotions in images uploaded by a user and the system recommends songs from the library whose tags match those in the uploaded images. Then depending on the emotion perceived by both systems we match them and succeed to recommend some BGM for a slideshow curation.

3.1.3 EmoBGM (C)

We used WEKA to input the acoustic features extracted and built a classification model to automatically tag music files with an emotion. We created a model using different sets of features and different algorithms (random forest, J48, support vector machine). Table 3.4 summarizes the result we obtained. The best classification accuracy is reached when using random forest algorithm with the 9 best features computed by the attribute selection BestFirst algorithm.

We observe that randomly choosing 20 features among the provided ones gives relatively good classification accuracy between 80% and 86% depending on the algorithm used. This good result needs to be put into perspective when we see the result given when we choose 9 random features. This means that we need to find a better way to find out which features can provide us a good classification accuracy. We use the attribute selection BestFirst algorithm to get the 9 best features from the 99 features. With these features, we reach 88% of accuracy with

support vector machine algorithm as well as with random forest. To make our final decision, we create and evaluate our model with 5 random features among the 9 best provided. The result shows that random forest keeps its accuracy while support vector machine drops to 80%. We finally choose random forest algorithm for our model in this research.

Table 3.4: Comparison of result of classification using different algorithm

	J48	SVM	Random forest
20 random features	86%	80%	86%
9 random features	74%	68%	80%
5 among best features	84%	80%	88%
9 best features	84%	88%	88%

3.2 Implementation of the evaluation application

The second part of our work dealt with the evaluation of the resulted model. The main question we want to answer is: Is our model capable of recommending a background music to a set of photos conveying a certain emotion? We performed a subjective evaluation to verify whether the users found the recommendation of background music based on emotion useful and if the result provided suits them.

To do this, we built a web application which embeds a set of background music that our model analyzed and classified per category with a confidence percentage. Since we want to recommend a background music to a set of images based on emotion, we need an automatic image’s emotion analyzer to output the main emotion conveyed in a given image. Then, based on emotion and confidence estimation, our application recommends a suitable background music.

As automatic image emotion analysis system, we used Microsoft Emotion API that we first evaluate to confirm its ability to guess emotion in images. We followed the process explained below.

3.2.1 Microsoft Emotion Api Evaluation (D)

Microsoft Emotion API is a service provided by Microsoft cognitive services. From a photo, this service is able to analyze the faces and give an estimation of the emotion of the person in the image. It classifies the facial emotion into 8 categories (Anger, contempt, Happiness, Disgust, Neutral, Sadness, Surprise and Fear). Figure 3.5 shows an example of the API's output from their test platform*.

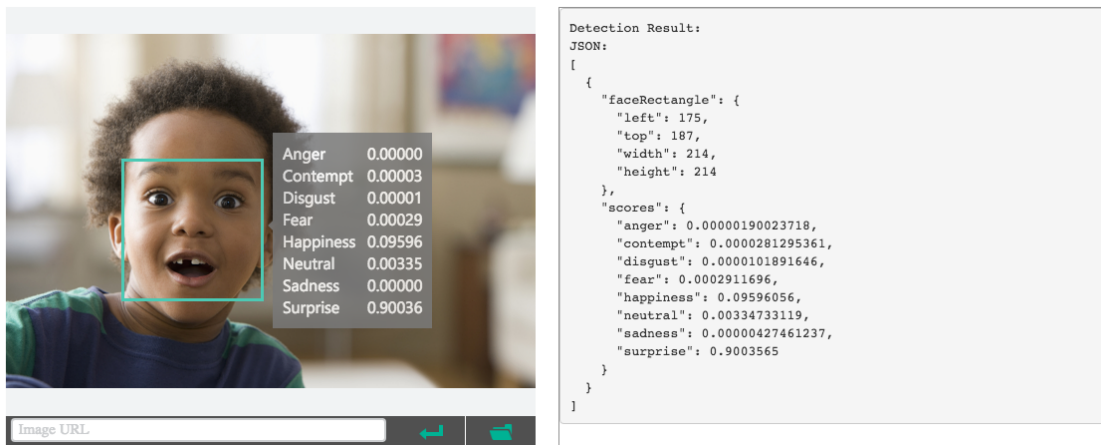


Figure 3.5: Example of result given by Microsoft Emotion API

To evaluate the prediction of Microsoft emotion API, we built another web application (Figure 3.6), where the user had to choose the emotion he thinks the main character in the photo is feeling. Microsoft Emotion Api analyses all the faces in the photo but we take into account only the main character's facial expression, which is the one with the biggest face detected. Among the 10 candidate photos, some are from our lab photo library and some are from the Internet.

*<https://www.microsoft.com/cognitive-services/en-us/emotion-api>



Figure 3.6: Microsoft Emotion Api evaluation app

24 people participated in this task. 18 people out of 24 were present when some of the photos were taken so they knew the emotion felt by the people at that time, a fact we used as ground truth. We asked the participants to focus on the facial expressions and not on the whole scene since Microsoft emotion API only uses facial expressions. The results, summarized in table 3.5, show that Microsoft Emotion API is very accurate, guessing in 8 cases out of 10 the major emotion in the image provided.

3.2.2 Best background music selection

Our application embeds some candidate background music with their emotion estimation given by our model. To get the recommended background music, the user uploads his photos. The system sends a request to Microsoft emotion API and gets the emotion estimation of each photo. Using this estimation, we calculate the matching score of each candidate BGM following this equation:

$$Score_{BGM(i)} = |S_{img(i)} - P_{bgm(i)}| \quad (3.3)$$

Table 3.5: Microsoft Emotion API Evaluation task result

Image	Microsoft emotion API Guess	Human Guess
Image 1	Surprise 71%	Surprise 67%
Image 2	Neutral 63%	Happiness 50% Neutral 33%
Image 3	Anger 99%	Anger 99%
Image 4	Happiness 100%	Happiness 99%
Image 5	Disgust 55%; Anger 34%	Disgust 41% Sadness 33%
Image 6	Sadness 100%	Sadness 67%
Image 7	Happiness 99%	Happiness 67%
Image 8	Surprise 51%; Fear 47%	Fear 63%
Image 9	Happiness 99%	Happiness 99%
Image 10	Neutral 99%	Neutral 45%; Contempt 36%

where $S_{img(i)}$ is the percentage of the strongest sentiment detected in the image and $P_{bgm(i)}$ is the percentage of the strongest sentiment of the candidate background music.

As shown in Figure 3.7, which is the screenshot of the final system, the 3 background music with the highest scores are selected and recommended via a play button displayed below each photo. For the photos tagged as neutral, we recommended a random background music downloaded from the free music archive[†].

[†]<http://www.freemusicarchive.org/>



Figure 3.7: EmoBGM evaluation App

Chapter 4

Evaluation

4.1 Evaluation scenario

To check whether our system succeeded to match the background music with the emotions depicted in pictures, we tested it on 3 different sets of photos, with each set containing 5 photos. The photos in the first set were a wedding scene showing the bridal couple smiling directly into the camera.

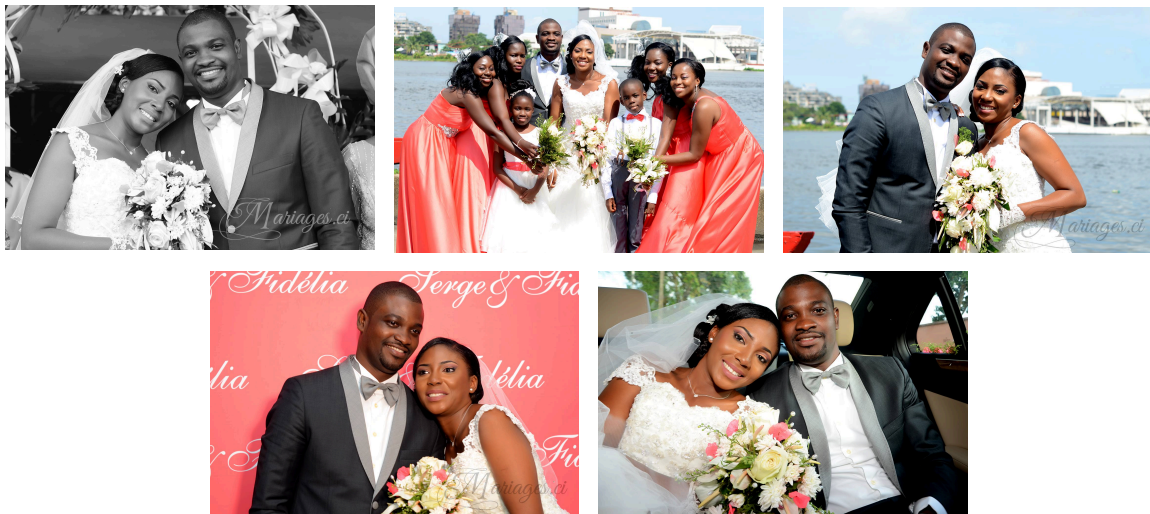


Figure 4.1: Evaluation set 1

The second set was of people on a sightseeing trip in Kanazawa, Japan. Some

photos showing the people interacting with the surroundings and each other while others showed them smiling into the camera at a restaurant.

The third set showed people talking and playing in a soccer game. None of the people pictured were looking into the camera.



Figure 4.2: Evaluation set 2



Figure 4.3: Evaluation set 3

16 people participated in the evaluation. Each participant completed 3 tasks: First, they answered the following question: Do you think that a system that analyses the emotion in the photos and suggests the music accordingly is necessary? Next, they rated the recommended music(s) for each set of photos using a 5-point Likert scale and gave a comment on their feelings on the association between the music and photo set (i.e. reason behind the rating given). Finally, they answered the question: What information or features are important to provide better matching results?

4.2 Evaluation results

For the first question, 100% of the participants agreed that a system for recommending background music according to the emotion is necessary and will save a lot of time while creating a slideshow. Figure 4.4 shows the average user rating on how appropriate the recommended music clips are based on the emotions observed in the photos (task 2 in evaluation). Out of a maximum of 5, users gave ratings of 4.1, 3.6 and 3 for photo sets 1, 2 and 3 respectively. The key phrases from the comments given on the reasons behind the ratings are summarized below. We discuss them in more detail in the discussion section.

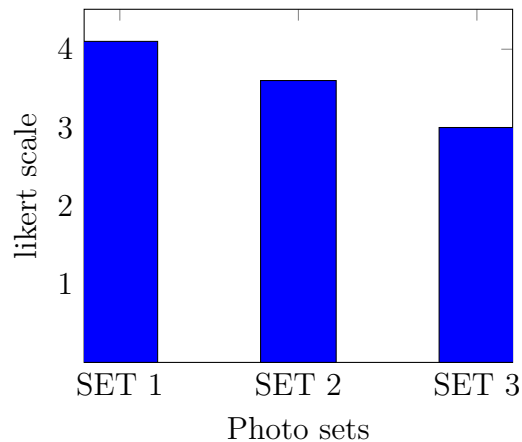


Figure 4.4: Average rating (on a scale of 5) for how appropriate the recommended BGM is for the set of pictures

Chapter 5

Discussion

5.1 Discussion on the experimental result

We observed a decrease in rate from 4.1 for the first set (wedding ceremony) to 3 for the third set (soccer game). Indeed, the users were very excited about the first set and appreciated the accuracy of the recommended music which, according to them, matches with the emotion conveyed in the images.

Then their appreciation rate dropped because of the difference in the emotional impact of each set. The photo set 1, depicting a wedding ceremony and rated as happy by the image emotion analyzing system, caused a peak of excitation within the users. Hence the rate was relatively high in the first set. The second photo set, presenting a lab trip, aroused less excitement among users especially as the pictures do not necessarily show the face of the people. The rate of the third set is relatively low. This is due to the fact that few of our users play soccer or team sports. It was, therefore, difficult for them to understand the emotion of the people present in the photos and decide whether a background music is suitable to express what the person in the photo was feeling.

5.2 Discussion

5.2.1 Approaches

Many approaches could have been used to automatically assign background music to a set of photos.

We could mine through social network services for slideshows accompanied by bgm, and use this data for machine learning. However, this would result in a seemingly infinite amount of data (photos, music and videos), therefore infinite number of features and differences, that would lower the accuracy of any model.

We preferred to use emotion to match music and images because it is a common component of both of them. We can categorize music in a finite category of emotion and images as well, thanks to context recognition and facial expression. Finally, we aimed to create impressive slideshows which means a slideshow which preserves the memory, and we believe that using emotion is the best approach to reach our goal.

5.2.2 System enhancement

It emerges from these analysis that, besides the major emotion conveyed in a photo, the context recognition of the scene is also very important to match accurately the images and the background music.

Subclasses

From the results and comments we obtained, we are able to confirm that our method succeeded to accurately estimate the emotion in a piece of background music. However, to increase the accuracy and a better recommendation for slideshows, we must extend the classification to subclasses such as parties or sports events for the happiness category.

Google Cloud Vision

In order to tackle the problem of context recognition in photos, we investigated Google Cloud Vision* . With this API, it is possible to get facial emotion as well as labels that illustrate the picture and help to understand the context of the photo.

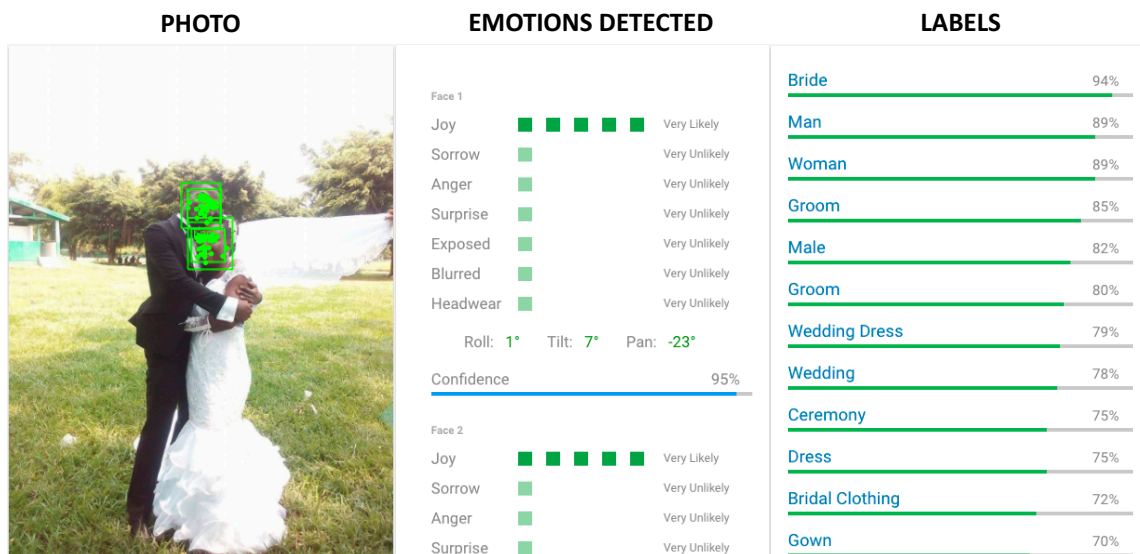


Figure 5.1: Screenshot of google vision api output

Figure 5.1 is an example of what Google vision API give as result. As input we uploaded a wedding photo, showing the groom and the bride smiling. Google vision API detected the two faces and estimated that the two faces are expressing joy with 95% of confidence. The labels outputted are also very accurate describing a wedding ceremony. Using this tool will help us to understand easily the context in the photo and recommend a better background music.

*<https://cloud.google.com/vision/>

Chapter 6

Conclusion

Finding out the emotion someone feels out from a media file is a very difficult task. Nevertheless, the human senses succeed somehow to distinguish the different emotions conveyed in photos or music. We exploited this insight to create a machine learning model capable of determining which emotion is conveyed by a piece of background music. The goal of our proposed system is to be able to recommend a suitable background music for slideshows based on the main emotion conveyed in the photos. Despite the lack of additional information such as lyrics or annotations, which could give us a clue for our emotion recognition model, we succeeded to reach 88% of accuracy with cross validation techniques, and an average rate of 4.1, 3.6 and 3.0 for the photo sets 1, 2, and 3 respectively of the subjective evaluation we conducted, demonstrating the effectiveness of our proposed system.

For future work, we plan to improve our music model by adding the possibility to label the background music with a genre and possible suitable event automatically. This will permit to our system a better recommendation even inside categories. Our image emotion analyzer also can be improved by using Google Cloud Vision * and photos' EXIF information. Besides detecting faces and classifying them in emotion category, Google cloud vision detects individual objects in the photo and is able to provide the relevant words (labels) that can help to understand the context of the photo as well as the GPS data and date info etc. that can be obtained from the EXIF files of the photos. Next, we will

*<https://cloud.google.com/vision/>

define the rules to choose the suitable background music for a whole set of related photos, taking into account the possible changes of emotion detected from a photo to the next one and manage the case when there is more than one emotion detected in a photo.

Acknowledgements

Upon the completion of this thesis, I would like to express my deepest gratitude to all those people who have made this dissertation possible.

First and foremost, i would like to thank Professor Keiichi Yasumoto, my advisor, for accepting me in his prestigious research lab and allowing me to work under his supervision. For two and half years in the Ubiquitous Computing Systems Laboratory (Ubi Lab), he taught me everything i know today about research and team work and i especially thank him for his help on my writings and for carefully reading and commenting on countless revisions of all my manuscripts. Being strict and kind at the same time makes him the most respectful professor i have ever met.

Special thanks to Professor Kazushi Ikeda, his advices and comments since the seminar 2 and on my thesis defense gave me a new perspective of my research that drives me to the future.

I would also like to thank Professor Yutaka Arakawa for all his advices and guidances. I feel very lucky to have the chance to be around this professor full of brilliant ideas. He transmitted me his passion for research and the courage to keep moving forward despite the bad results i sometimes encountered. From the bottom of my heart thank you sensei.

I also thank Assistant Professor Hirohiko Suwa for his reliable support. Despite his frequent travels, he always found time in his busy schedule to advice me via Skype or e-mail. I feel very blessed to have the chance to work with him. His valuable suggestions, questions, and support helped me a lot to improve my

research especially in the machine learning area.

Thanks to Assistant Professor Fujimoto Manato for being more than a professor, a elder brother. his ability to understand the students and always encourage them made the ambiance in the laboratory, once of the most awesome i ever encountered in a working area.

I also thank all of the Ubi lab members who benefit me with their wisdom, encouragement, and friendship. These two years with you, in this warm laboratory, will remain forever engraved in my mind.

Special Thanks to PhD. Doudou Fall, PhD. Nelson Kibinge, Babatunde Ojetunde, Edith Luhanga and all the African community in NAIST, who made my stay smooth and pleasant. They are the big brothers from mother Africa, who always supported me in many things including my research.

Thank from the bottom of my heart, to my two buddies William-Fabrice Brou and Elder Akpa. You were and are still there whenever, wherever I needed you. Thanks brothers.

To my parents and family who did everything for me, who support me and believe in me.

To my dear wife Rolande KONAN.
I love you all.

References

- [1] Charles S Areni and David Kim. The influence of background music on shopping behavior: classical versus top-forty music in a wine store. *NA-Advances in Consumer Research Volume 20*, 1993.
- [2] Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. Music mood and theme classification-a hybrid approach. In *ISMIR*, pages 657–662, 2009.
- [3] Peter Dunker, Stefanie Nowak, André Begau, and Cornelia Lanz. Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 97–104. ACM, 2008.
- [4] Tuomas Eerola, Olivier Lartillot, and Petri Toivainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *ISMIR*, pages 621–626, 2009.
- [5] Paul Ekman, Wallace V Friesen, Maureen O’Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.
- [6] Thomas Fritz, Sebastian Jentschke, Nathalie Gosselin, Daniela Sammler, Isabelle Peretz, Robert Turner, Angela D Friederici, and Stefan Koelsch. Universal recognition of three basic emotions in music. *Current biology*, 19(7):573–576, 2009.

- [7] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.
- [8] Tao Li and Mitsunori Ogiwara. Detecting emotion in music. In *ISMIR*, volume 3, pages 239–240, 2003.
- [9] Janis Libeks and Douglas Turnbull. Exploring artist image using content-based analysis of promotional photos. In *Proc. of the Int. Computer Music Conf*, 2010.
- [10] Cory McKay, Ichiro Fujinaga, and Philippe Depalle. jaudio: A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval*, pages 600–3, 2005.
- [11] Ronald E Milliman. The influence of background music on the behavior of restaurant patrons. *Journal of consumer research*, 13(2):286–289, 1986.
- [12] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.
- [13] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.
- [14] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.
- [15] Erik M Schmidt and Youngmoo E Kim. Prediction of time-varying musical mood distributions from audio. In *ISMIR*, pages 465–470, 2010.
- [16] Erik M Schmidt, Douglas Turnbull, and Youngmoo E Kim. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the international conference on Multimedia information retrieval*, pages 267–274. ACM, 2010.

- [17] Björn Schuller, Johannes Dorfner, and Gerhard Rigoll. Determination of nonprototypical valence and arousal in popular music: features and performances. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):735854, 2010.
- [18] Janto Skowronek, Martin F McKinney, and Steven Van De Par. A demonstrator for automatic music mood estimation. In *ISMIR*, pages 345–346, 2007.
- [19] Dan Yang and Won-Sook Lee. Music emotion identification from lyrics. In *Multimedia, 2009. ISM’09. 11th IEEE International Symposium on*, pages 624–629. IEEE, 2009.
- [20] Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. Toward multi-modal music emotion classification. In *Pacific-Rim Conference on Multimedia*, pages 70–79. Springer, 2008.

Publication List

- [1] EmoBGM: estimating sound's emotion for creating slideshows with suitable BGM, Percom Worshop, 2017.
- [2] Sentiment Analysis of BGM Toward Automatic BGM Selection Based on Emotion, Mobisys Poster session, 2016.