Emotion Based Music Player

Final year project (Comp30040)

Student: Umer Iftikhar
Supervisor: Dr. Tim Morris

The University Of Manchester
The School of Computer Science
Abstract

Emotion Based Music Player

The study of music and emotions suggests that there is a psychological relationship between a person’s emotional state and the type of music they listen to.

The purpose of this project is to understand and analyze various algorithms for an emotion recognition system. This project is split into two halves; extracting metadata from songs to determine their genre and using machine learning to determine the emotion of the user.

The project centers heavily towards machine learning and data mining, where different techniques are used to determine the emotion of the user and analyzes the implications of using each technique. Using the determined emotion, the final music library of the user is sorted out.

This report begins with a review of the existing feature extractors and the current techniques used for emotion recognition before discussing the development approach and design of the system.

Having completed the application, the performance of each of the different techniques is reviewed. A comparison and analysis of the different feature extractors used is made. These results are then compared against real world results. Finally I discuss further research opportunities which could improve the application.

Student: Umer Iftikhar
Supervisor: Dr. Tim Morris
# Table of Contents

Introduction ........................................................................................................................................ 6  
1.1 Motivation and Project Overview: .......................................................................................... 6  
1.2 Aims and Objectives: .............................................................................................................. 7  
  1.2.1 Emotion Recognition: ......................................................................................................... 7  
  1.2.2 Music Analysis: .................................................................................................................. 7  
1.3 Technologies Used: .................................................................................................................. 7  
1.4 Report Overview: ...................................................................................................................... 8  
Background and Past Work ............................................................................................................ 9  
2.1 Past Work: .............................................................................................................................. 9  
  2.1.1 Feature Extractors: .......................................................................................................... 9  
  2.1.2 Classifiers and Prediction: ................................................................................................ 9  
  2.1.3 Facial Emotion Recognition: ............................................................................................ 9  
System Design ............................................................................................................................... 11  
3.1 System Overview: ................................................................................................................... 11  
3.2 Data Training: ........................................................................................................................ 12  
3.3 Live Learning Algorithm and User upload: ........................................................................... 13  
3.4 Development Approach: ......................................................................................................... 14  
Emotion Recognition System: ...................................................................................................... 15  
4.1 Face Location Determination: .............................................................................................. 15  
  4.1.1 Viola Jones Face Detection: ............................................................................................. 16  
  4.1.2 Local Binary Patterns: .................................................................................................... 16  
4.2 Regions of Interest: ............................................................................................................... 17  
4.3 Prediction and Feature Extraction: ......................................................................................... 19  
  4.3.1 SIFT: ............................................................................................................................... 20  
  4.3.2 SURF ............................................................................................................................. 21  
  4.3.3 Bag Of Words Model: ..................................................................................................... 22  
  4.3.4 BRIEF: ........................................................................................................................... 23  
  4.3.5 Harris Edge Detection: .................................................................................................... 23  
  4.3.6 Principal Component Analysis: ......................................................................................... 24  
Music Classification System .......................................................................................................... 26  
Testing and Evaluation ................................................................................................................. 28
List of Tables

Table 1 Metadata stored in an MP3 File ........................................................................27
Table 2 Frontal and sideways comparison ....................................................................29
Table 3 Frontal and sideways accuracy ..........................................................................29
Table 4 SIFT Confusion Matrix - Support Vector Machines .........................................30
Table 5 SURF Confusion Matrix - Support Vector Machines ........................................31
Table 6 PCA with BRIEF Confusion Matrix - Support Vector Machines .....................31
Table 7 PCA with Harris Edge Detector Confusion Matrix - Support Vector Machines ....32
Table 8 SIFT Confusion Matrix - Naive Bayes ................................................................34
Table 9 SURF Confusion Matrix - Naive Bayes .............................................................35
Table 10 PCA with BRIEF Confusion Matrix - Naive Bayes ...........................................35
Table 11 PCA with Harris Edge Confusion Matrix - Naive Bayes ...................................35
Introduction

Facial expressions are a great indicator of the state of a mind for a person. Indeed the most natural way to express emotions is through facial expressions. Humans tend to link the music they listen to, to the emotion they are feeling. The song playlists though are, at times too large to sort out automatically. It would be helpful if the music player was “smart enough” to sort out the music based on the current state of emotion the person is feeling.

The project sets out to use various techniques for an emotion recognition system, analyzing the impacts of different techniques used. Thereafter it sorts out a person’s playlist based on the predetermined emotion of the user. There are two main parts of the program; determining the emotion of the user using a webcam and then sorting out the current playlist based on that emotion.

The application is developed in such a way that it can analyze the image properties and determine the mood of the user. The application also includes the facility of sorting songs based on mp3 file properties so that they can be added into appropriate playlists according to the mood.

1.1 Motivation and Project Overview:

Computer Systems based on an effective user interaction could play a vital role in the next generation of computer vision systems. Facial gestures and emotions give clues about a person’s emotions. Face emotions have been used in areas such security, entertainment and Human Machine Interface (HMI).

Generally people have a large number of songs in their databases. The project is an attempt to minimize the effort taken in sorting out these playlists. Generally people randomly select a song from their playlists and the given song might not be appropriate for the current mood of the user and might disappoint them. Moreover there are very few softwares which actively sort music based on the mood of the user. The project aims to provide better enjoyment to the music lovers in music listening.

In the model, the following moods are included:

1. Happy
2. Sad
3. Angry

The system involves the image processing and facial detection processes. The input to the model is still images of user which are further processed to determine the mood of user. Using this
emotion, the playlist of the user is sorted. Several techniques are used for emotion detection, the techniques are compared and evaluated based on different criteria.

**1.2 Aims and Objectives:**

The key objectives of this project can be split into two parts, the recognition of the emotion of the user and music analysis. The project is centered and focused more towards different approaches to emotion recognition and the impact of each technique used. The emotion recognition stage is heavily based on image processing and machine learning. The music analysis is done by reading the MP3 metadata of a music file.

1.2.1 Emotion Recognition:
The key aim of this section is to implement and analyze various techniques to extract features and classify the emotion of a person. The image processing step requires turning the image to grayscale and resizing it. This is followed by extracting multiple features using different techniques and adapting different classifiers to determine the mood of the user. Using these different methods and techniques, an analysis is made to determine the best solution for the emotion recognition problem based on my project.

1.2.2 Music Analysis:
Using the bit stream from mp3 files, we extract metadata to determine the required information for each particular song. Using the determined emotion, create a playlist of songs for the user.

**1.3 Technologies Used:**

The system was written in C++ with a Model View Controller (MVC) approach. The use of MVC enforced that the user interface, the models and the controller are separate hence improving cohesion and decreasing coupling.

The chief libraries used by the system are OpenCV and TagLib.

OpenCV was used for the image processing and machine learning section of the project. It had tools to perform feature extraction and classification. Using the frameworks provided by OpenCV the system was efficiently able to perform feature extraction and classification providing some very highly accurate results as depicted in 6 Testing and Evaluation.

To access the Mp3 files as a bit stream and read the metadata tags ID3v1 and ID3v2, the library Taglib was used. It provided access to reading in a song library and extracting metadata, used for classification. To supply the system with either the training data of pictures from sample subjects, or the user’s song playlist, a comma-separated values (CSV) file was used. This was to upload any data to the system.
1.4 Report Overview:

1 provides the introduction and motivation for the project.
2 details on past works and a little background information required to understand the system. 3 gives an analysis of the system design and some specific key features of the system. 4 details the various techniques and approaches taken for the machine learning and classification of human emotions and how the system has implemented them.
5 discusses the problem of extracting metadata from an Mp3 file source and the solution proposed to it by the system. 6 produces an evaluation on the various algorithms implemented for emotion classification based on thorough testing done by the system. The conclusion in 7 comprises of targets achieved in comparison to the aims of the project, limitations of the project and proposes further research areas.
Background and Past Work

As outlined in chapter 1 the project is divided into two major parts: Image processing, which is used for recognizing human emotions; and a mp3 music file analysis used to extract song information and classify songs.

2.1 Past Work:
The current most popular emotion based music player is StereoMood. It lacks capabilities in the sense that the user needs to type in what he is feeling, rather than using computer vision to determine his emotion.

To solve the problem of emotion recognition a lot of work has been done in the past. To extract and determine the emotion of a user, we need to extract features from an image and use them against a trained data set to classify the input and determine the emotion.

2.1.1 Feature Extractors:
A feature extractor is an application which extracts important points in an image. Different works have been done in the field of Computer vision for feature extractors, the most prominent ones being Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF). Each of these has different impacts on classifying the emotion of the user. I initially work with a new technique called Binary Robust Independent Elementary Features (BRIEF) before moving onto techniques such as SURF and SIFT.

2.1.2 Classifiers and Prediction:
After extracting features from an image set of training and testing data, a feature classifier is needed to sort out and classify the testing data with relevance to the training data. A Support Vector Machine (SVM) is the most predominantly used classifier to tackle the emotion recognition problem. For experimental purposes I use an SVM and a Naive Bayes Classifier.

2.1.3 Facial Emotion Recognition:
Several approaches have been proposed to classify human affective states. The features used are typically based on displacements of specific points or spatial locations of particular points; this technique is known as Facial Action Coding System (FACS).

In an approach taken by Liu et al in [3], he presents an algorithm for classification of brain electrical signals in human emotions. This algorithm was based on the model of fractal dimension. He proposed a bi dimensional Valence - Arousal approach, where by the six emotions are divided into different categories.
Black et al. [1] explored the use of local parameterized models of image motion for recovering and recognizing the non-rigid and articulated motion of human faces. They used these parametric models to extract the shape of the mouth, eyes and eyebrows. They achieved a high success rate of 95% to classify Happy, 90% to classify Anger and a 100% success rate to classify the Sad emotions. On the other hand the approach used by Yacoob and Davis [2] in which facial expressions are recognized in image sequences using statistical properties of the optical flow with only very weak models of facial shape.

In this project several approaches are considered, including a Principal Component Analysis (PCA) approach, using multiple Facial Action Units and different feature extractors with clustering approaches. Each of these approaches are used with different classifiers to determine the emotion of the user with the accuracy tested on a data set of 40 subjects each.
System Design

This chapter covers design of the system, including an overview of the architecture and descriptions of the key components.

3.1 System Overview:

The system consists of two major sections: Image analysis; and music file analysis. The program heavily focuses on image analysis using various training samples to train and predict data. It consists of multiple classifiers, various feature extractors and an evaluation function. Past works and research heavily influenced the approach considered towards development for this project.

The key features include; a live learning algorithm for associating each music genre with an emotion based on the preference of the user, recording pictures of the user based on a predefined time scale to capture changes in emotion and organizing music playlists in a lexicographic order in the end to output a music playlist.

Towards development, a test driven approach is taken along with various other design patterns. All images used for testing and training are converted from RGB to Grayscale and resized to the same dimensions. Once image analysis is finished, the mp3 files of the users are analyzed as a bitstream where the ID3 tags are extracted. These contain information about the songs including titles, genre etc. Once this data is extracted, the songs are classified using the predetermined emotion to create a music playlist.
3.2 Data Training:

To determine the current emotion of the user 2 things are required, an image from the user and a predetermined training dataset.

The training phase only occurs the first time in the development process and once the required data had been extracted and the classifier has been trained, the data is stored as an XML File. This file was later used to predict and classify against, rather than training the dataset on every single instance the program was run.

A dataset of 50 images per emotion are used for training purposes. The database used for training was the open source Yale Face database B [6]. The emotions being considered in the project are happy, sad, angry and neutral. For each image in the dataset, the images are resized to a 260 x 360 pixels. Once the images have been resized, they are smoothed and converted to grayscale in order to reduce the “noise” and extract only important details.

After performing this preprocessing of images, a feature extraction algorithm is run over each of the images. Amongst the various feature extractors available, the system makes use of SIFT, SURF or BRIEF. By using these various feature extractors, different key points are located and fed to a classifier.

Several other techniques are used to improve final accuracy of the system. These are image preprocessing steps which occur before feature extraction. These techniques include only working with a few regions of interested within the face, using clustering techniques and a dimensionality reduction technique known as Principal Component Analysis (PCA).

The techniques to improve accuracy depend strongly on the choice of the feature extractor, as the results in 7 show. Clustering techniques such as the Bag of Words model, and K means clustering, proposed and used in the system do not work effectively with binary features extracted using the BRIEF feature extractor. Whereas the PCA approach has been effective with the Harris Edge Detector rather than the SURF or SIFT feature extractors.

The extracted data is fed to the classifier of choice - SVM or Naive Bayes and the classifier is then trained on the data. This data is stored in an XML file to be used later.
3.3 Live Learning Algorithm and User upload:

One key feature of the system is the live learning algorithm to classify music genres. After discussions with potential users, research suggested that each person associates a music genre to a different emotion based on their own preferences. As the system starts up, it outputs a list of all the 38 possible music genres. The user has is to sort out which genre’s he or she listens to based on the particular emotions. These preferences are stored and used later in the system.

The elementary step to emotion recognition from a live camera feed is to first accurately locate and extract the face of a person. The predominantly used technique for this task is to use a Cascade Classifier: a cascade of boosted classifiers working with particular features.

The two Cascade classifiers the system uses are Haar Cascade Classifier and Local Binary Pattern (LBP) Cascade Classifier. Both work completely differently with LBP being slightly more efficient and faster than the Haar classifier. As soon as the camera feed opens up, the classifier locates the face based on primarily difference of pixel intensities. Once the face is located, the user has to select the option to save the image once he is happy with the emotion he is portraying.

The saved image from the user goes through the exact image preprocessing steps that the training data goes through. These steps are done to maintain the consistency of the system.

Once the image has gone through these required steps, the key points and features are extracted using the chosen feature extractor. The data points about extracted as cartesian X Y coordinates and stored in a matrix. This data is fed to the classifier of choice and it is compared against the training data. Data Classification is done and the emotion of the user is predicted.
3.4 Development Approach:

The entire system was designed with a test driven development approach. Final test cases of people portraying particular emotions was predetermined. A set of 160 images was used to test the system, 40 images per each emotion being to test the system thoroughly.

Through research, the Model View Controller (MVC) approach is considered to be the ideal design pattern because of its ability to separate concerns between the data, the user and the algorithms [5]. MVC concentrates its operations with 3 distinct areas:

1. The model, which represents the data that the system handles.
2. The view, which is concerned with controlling the user interface and the presentation of data.
3. The controller, which handles the algorithms and database to develop the system.

The project was developed in an iterative manner, using sprint cycles; key pieces of functionality identified in the planning process were designed, implemented, deployed and tested first. Testing occurred at the end of each cycle on the new implemented features, before further additions.
Emotion Recognition System:

The system for emotion recognition is divided into 3 parts: face location determination, feature extraction and emotion classification. The project and system heavily focuses on the latter two parts, the feature extraction and classification.

Once the face location has determined and the face extracted, the system moves onto extracting regions of interest, in particular, the eyes, eyebrows and mouth. This stage is followed by a feature extraction stage, where various algorithms are used to implement multiple techniques. Their results are evaluated and presented in the later Testing and Evaluations. This extracted data is fed to a classifier which determines the emotion of the user.

4.1 Face Location Determination:

The system uses two particular algorithms for face detection, Haar features using the Viola-Jones Technique and Local Binary Patterns. The report does not delve too deeply into these, as they are not central to the project.

Figure 4 overview of the emotion recognition system
4.1.1 Viola Jones Face Detection:

The algorithm is a robust real time learning algorithm which uses special features called Haar-like features to do object detection. The algorithm has 4 stages:
1. Haar Features Selection
2. Creating Integral Image
3. Adaboost Training algorithm
4. Cascaded Classifiers

![Figure 5 Typical Haar-like features used by Viola Jones](image)

4.1.2 Local Binary Patterns:

This algorithm has its roots in the 2D texture analysis of images. The basic idea is to summarize the local structure of an image by comparing each pixel with its neighbor. The proposed idea is to take a pixel in the center and threshold it against its neighbors. If the value is less than its neighbor, denote it with a 0 or else denote it with a 1. Hence you end with a binary value for each pixel, known as the Local Binary Patterns (LBP).

A useful extension to the original operator is the so-called uniform patterns[8]. These patterns improved the efficiency of the system by reducing the length of the feature vector and hence implementing a simple rotation invariant descriptor. The system makes use of these uniform patterns to determine face locations from a camera feed.
The LBP approach was a much faster approach to locate the faces, although it compromised slightly on accuracy rates. But the gains from the speed vastly outweighed the very miniscule loss of accuracy.

### 4.2 Regions of Interest:

Using the local binary patterns again, this time trained to extract certain regions of the face, the eyes, mouth and eyebrows were extracted. These were the regions of interest from where the features would be later extracted from.
Figure 9 Eyes and Forehead Region of Interest Extraction.
The figure below shows the different facial landmarks that are provided to a classifier for training and testing purposes.

![Facial Landmarks detected by Harris Corner detector](image)

*Figure 10 Facial Landmarks detected by Harris Corner detector*

Using the facial landmarks shown in figure 10, the following measurements are used for training various classifiers as stated in section 4.3.5, Harris Edge Detector.

- Eye Width = \((\text{Right Eye Width} + \text{Left Eye Width}) / 2\)
  \[= \frac{(d1 - d2) + (c2 - c1)}{2}\]
- Eye Height = \((\text{Right Eye Height} + \text{Left Eye Height}) / 2\)
  \[= \frac{(d4 - d3) + (c4 - c3)}{2}\]
- Mouth width = \(f2 - f1\)
- Mouth Height = \(f4 - f3\)
- Distance between eyes = \(d2 - c2\)

### 4.3 Prediction and Feature Extraction:

In order to predict the current emotion of the user, different feature extractors have been used to effectively locate important key points and distinguishing features from particular face regions.
Based upon research discussed in 2, the system makes use of a few different feature extractors and draws conclusions, based on speed and accuracy to determine which the most successful feature extractor is. The feature extractors and techniques used in the system are

- Scale Invariant Feature Transform - SIFT
- Speeded Up Robust Features - SURF
- Binary Robust Independent Elementary Features - BRIEF
- Principal Component Analysis - PCA
- Harris Corner Edge Detection

This chapter introduces the basic ideas and concepts behind each of these feature extractors and how they are implemented by the system. 5 discusses the relative impacts on testing accuracy of using these different techniques.

4.3.1 SIFT:
SIFT is amongst the founding feature extraction algorithms. It is highly efficient being invariant to image rotation, scaling and partially invariant to camera illumination. SIFT has its roots in three elements, the SIFT descriptor, the key point location, and the orientation. The match for each key point is found by identifying its nearest neighbor, which is defined as the key point within a minimum distance threshold. To ensure a correct match Lowe [3] suggests rejecting all matches in which the distance ratio is greater than a threshold.
Automatically locating landmark points from an arbitrary view facial image is very challenging [7]. [Figure 12]. To overcome this problem, the system makes use of a dense SIFT feature description to describe facial images [Figure 13]. To be precise, the system divides the entire face into patches and extracts a 128 bit SIFT feature vector from each region [8]. These features are then used in the training and testing of images [9].

![Facial landmarks extracted from SIFT](image1)

**Figure 12** Facial landmarks extracted from SIFT

![Facial region divided into patches and each patch produces a SIFT vector.](image2)

**Figure 13** Facial region divided into patches and each patch produces a SIFT vector.

### 4.3.2 SURF

The properties of SURF lie within scale invariance, strong robustness and a strong distinction between feature points. In comparison to SIFT the SURF operator has greatly improved computer speeds [12]. The algorithm consists of 4 main parts:

1. Generating an Integral Image
2. Fast Hessian detector (interest point detection)
3. Descriptor orientation assignment (optional),
4. Descriptor generation.

The key aspect of SURF algorithm is the use of an intermediate image, the Integral image. This integral image is subsequently used by all parts of the algorithm later on.
The image is convolved with the squares, rather than the Gaussian average, as convolving of an image with the square is much faster if an integral image is used [12].

The integral image is defined as:

\[ S(x, y) = \sum_{i=0}^{i=x} \sum_{j=0}^{j=y} I(i,j) \]

Due to its high accuracy rates, the SURF algorithm has its roots based in the determinant of the Hessian Matrix. Using the Integral Image, we calculate the Hessian matrix, as a function of both space \( x = (x; y) \) and scale \( \sigma \). Given a point \( x = (x, y) \) in an image \( I \), the Hessian matrix \( H(x, \sigma) \) in \( x \) at scale \( \sigma \), is defined as follows [12]:

\[
H(x, \sigma) = \begin{bmatrix}
L_{xx}(x, \sigma) & \cdots & L_{xy}(x, \sigma) \\
\vdots & \ddots & \vdots \\
L_{xy}(x, \sigma) & \cdots & L_{yy}(x, \sigma)
\end{bmatrix}
\]

Where, \( L_{xx}(x, \sigma) \) is the convolution of second order derivative \( \frac{\partial I}{\partial x^2} g(\sigma) \) with the image in the point \( x, y \) similarly with \( L_{xy}(x, \sigma) \) and \( L_{yy}(x, \sigma) \).

Both these algorithms have been used with a clustering approach known as a bag of words model [13]. This is a clustering approach which has been used in the system to improve the accuracy of the classifiers.

### 4.3.3 Bag Of Words Model:

The Bag Of Words (BoW) model is based on an order less collection of image features and does not take into account spatial information [13]. The steps for a Bag Of Words model using either SIFT or SURF are:

1. Extract features from the entire training set.
2. Cluster those features into a vocabulary \( V \); get \( K \) distinct cluster centers.
3. Encode each training image as a histogram of the number of times each vocabulary element shows up in the image. Each image is then represented by a length-\( K \) vector.
4. Train the classifier.
5. When given a test image, extract the features. Now represent the test image as a histogram of the number of times each cluster center from \( V \) was closest to a feature in the test image. This is a length \( K \) vector again.

This approach vastly improved the accuracy of the classifiers in the case for SURF and SIFT. Given a Bag Of Words model uses Euclidean distances to cluster feature points, it was not feasible to use it with BRIEF as BRIEF uses a vector of binary values. Hence calculating the Euclidean distance...
for these vectors does not give the right results. Therefore with BRIEF either an approach using PCA was considered or just using this feature extractor on its own. Even on its own the BRIEF feature extractor gave a relatively decent accuracy rate, as shown in 7.

4.3.4 BRIEF:

The BRIEF feature extractor works on the principles of Local Binary Patterns, as explained above in section 4.3.2. The only difference applied to this particular approach was to use this feature extractor with a FAST Pyramid scheme. The algorithm uses FAST in pyramids to detect stable key points. It then selects the strongest features using the Harris Corner Edge Detection response. Using first order-moments, it then determines their orientation computes the descriptors using BRIEF (where the coordinates of the random point pairs are rotated based on the measured orientation). This approach is known as Oriented Fast and BRIEF, or simply as ORB.

4.3.5 Harris Edge Detection:

The Harris Edge and corner detection basically finds the difference in intensity for a displacement of \((u,v)\) in all directions [4]. This is expressed as below:

\[
E(u, v) = \sum_{x,y} w(x, y) \left( I(x + u, y + v) - I(x, y) \right)^2
\]

The window function can either be a rectangular or a Gaussian window which gives weights to the pixels underneath. I will not delve into the mathematical details of this approach. To get all the details on this corner and edge detector, please read the paper [4].

This corner detector is used in almost all feature extractors. The system explicitly makes use of this particular edge detector to manually find facial landmarks shown in figure 10. These landmarks are then in turn used to calculate specific features. Some of these features are:

1. Eye width
2. Eye Height
3. Mouth Width
4. Mouth Height
5. Distance between the Eyes

These dimensions are calculated over a range of different subjects. These subjects are of different ethnicities, ages and gender. These feature points and calculated dimensions are fed to the classifiers.
The final technique used by the system to improve the accuracy and classify emotions is known as the Principal Component Analysis or PCA.

### 4.3.6 Principal Component Analysis:

PCA is a renowned technique used in pattern recognition for dimensionality reduction [15]. As these patterns contain redundant information, mapping them into feature vectors can reduce or even completely get rid of the redundancy all the while preserving most of the intrinsic information content of the pattern itself. A face image, with size $N \times N$ in 2-dimension can also be considered as one dimensional vector of dimension $N^2$ [16]. Given facial images share the same prominent characteristics, and being similar in overall configuration, they will not be distributed randomly in our image space. Therefore they can be represented in a relatively low dimensional image sub space.

PCA has its roots in identifying underlying trends in a data set and therefore the main idea is to find vectors that best represent the distribution of faces within our image space. These vectors define the subspace of face images, which we call “face space” [16]. These vectors are known as Eigen faces because they are the eigenvectors of the covariance matrix of the original face image. The number of PCA components selected for the system was 90, as it gave a reasonably high accuracy rate without compromising the speed of the system.

The proposed steps for the approach involving the use of PCA for the system are as follows:

1. The images from the training Dataset are utilized to create a low dimensional face space. This is achieved by performing PCA on the training images and selecting the principal components with the greater Eigen. The number of principal components selected by the system being 90 and the number of Eigen being 1 less than the training database size; 39 per each emotion.
2. Create the projected versions of the selected images.
3. Project the test images on to the face space. All the test images are represented in terms of the selected principal components.
4. Generate the mean or average face for each particular emotion from the training dataset.

Once the average face or mean face has been extracted, the system uses two ways to determine the emotion.

Firstly, in the most basic approach it uses various feature extractors on the Eigen faces rather than actual images to do the training and testing of data. This approach worked well with BRIEF and binary features. SIFT and SURF performed extremely badly in comparison.

The other approach, the one which produced some of the highest accuracy rates for the system
was to compare the mean face against the test subject. In order to determine the intensity of the particular expression its Euclidean distance from the mean of the projected neutral images is calculated. This distance is compared against a range of predefined distances for each emotion and hence the emotion is classified.

Another robust and exhaustive approach considered was to compare the Euclidean distance of the test data against all images in the training dataset. The image with the lowest Euclidean distance was considered to be the emotion the test image was portraying.

1. Images from the training Dataset are utilized to create a low dimensional face space. This is achieved by performing PCA on the training images and selecting the principal components with the greater Eigen.
2. Create the projected versions of the selected images.
3. Project the test images on to the face space. All the test images are represented in terms of the selected principal component.
4. Generate the mean or average face for each particular emotion from the training dataset.
Music Classification System

The other core aspect of the system is the Automatic Music Genre Classification (AMGCC) problem. Music genres, in essence are categorical labels created by human experts in order to identify and classify the style and manner of the music. To extract music information from a given music MP3 file, two approaches were considered.

The first approach considered involved the use of machine learning to extract key features such as Mel Frequency Cepstral Coefficients (MFCC), pitch and harmony, Using these three vital features, classifying music based on it’s genre would be possible. For such an approach the C++ library Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) was used. The approach above was deemed to be unfeasible as the project gathered speed where only after a long development time, the MFCC was successfully extracted. That alone in itself was not enough to classify music genres.

With the limited time constraints of the project, a more simplistic approach was used. The system takes into account only MP3 files. Music and MP3 files always contain metadata in the last bits of the bit stream of the MP3 file.

MP3 Metadata:

Music meta information such title, artist genre etc are stored in tags knows as ID3 tags. These tags allows information such as the title, album, track number, and other information about the file to be stored in the file itself. There are two versions of these tags: ID3v1 and ID3v1

The ID3v1 is a 128 bytes long tag. The Strings are either space- or zero-padded. The Layout for the ID3V1 tag is as follows:

<table>
<thead>
<tr>
<th>FIELD</th>
<th>LENGTH (bits)</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Header</td>
<td>3</td>
<td>Tag</td>
</tr>
<tr>
<td>Title</td>
<td>30</td>
<td>30 characters for the title</td>
</tr>
<tr>
<td>Artist</td>
<td>30</td>
<td>30 characters for the artist name</td>
</tr>
<tr>
<td>Album</td>
<td>30</td>
<td>30 characters for the album name</td>
</tr>
<tr>
<td>Year</td>
<td>4</td>
<td>A four-digit year</td>
</tr>
</tbody>
</table>
The system uses an open source library **TagLib** to extract the relevant meta information from the ID3 Tags in an MP3 file source. The ID3v1 tag is stored in the last 128 bits of the file. The music genre is stored as a number ranging from 0 - 79 in the ID3v1 tag. To access this data the system reads the last 128 bytes to a string and parses it. The data in the string is than assigned to the relevant fields.

<table>
<thead>
<tr>
<th>Comment</th>
<th>28 or 30</th>
<th>The comment about the track</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-byte</td>
<td>1</td>
<td>If the track is stored, it contains a zero byte</td>
</tr>
<tr>
<td>Track</td>
<td>1</td>
<td>The track number on the album</td>
</tr>
<tr>
<td>Genre</td>
<td>1</td>
<td>The genre of the track</td>
</tr>
</tbody>
</table>

*Table 1 Metadata stored in an MP3 File*
Testing and Evaluation

The evaluation of the software indicates that the primary objective of the system has been achieved: using multiple techniques to perform emotion recognition and determining which has the highest success rates. This also suggests that the music classification section of the system has been achieved also. Hence the system uses the predetermined emotion to classify and sort out playlists for a user.

Testing the developed system was done with around 160 images from various databases. The images were tested using various feature extractors and using the two different classifiers. Each technique portrayed various results. The expressions considered were:

1. Neutral
2. Happy
3. Sad
4. Angry

6.1 During Development:

During the period of development, thorough unit testing occurred after new features had been implemented at the end of each sprint cycle. Testing was done on various subjects and the observations and comments were taken into account for future development.

6.2 Test Driven Development:

The entire system was designed with a test driven development approach. Final test cases of people portraying particular emotions was predetermined. A set of 160 images was used to test the system in the end. 40 images per each emotion was used to test the system thoroughly.

6.3 Results and Analysis:

The following section details on a few of the important findings of the system. Results from the basic use of SIFT and SURF were not very successful, but couple these 2 feature detectors with the Bag Of Words model, the results were much more accurate. PCA with BRIEF showed quite well results, but the highest accuracy rates were given by PCA with Harris Edge Detector, where the Euclidean distance was found between selective features from the face (figure 2) and the mean face produced by the PCA algorithm.
This section first compares the face recognition accuracy of a full frontal face to that of a sideways facing person. The Viola Jones technique (discussed in section 4.1.1) is extremely accurate in localizing a frontal face from a picture or a video. But problems arise when the user views sideways. This section compares the accuracy of local binary patterns and viola jones in face localization for a frontal face to that of a sideways facing subject.

A total of 80 subjects were tested, 40 were frontal facing and 40 were facing sideways either to the left or right.

<table>
<thead>
<tr>
<th></th>
<th>Viola Jones</th>
<th>Local Binary Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal Face</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>Sideways Face</td>
<td>17</td>
<td>14</td>
</tr>
</tbody>
</table>

*Table 2 Frontal and sideways comparison*

Viola Jones had a 100% success rate for frontal faces whilst local binary patterns were 97.5% accurate. For sideways faces Viola jones detected 42.5% of the faces whilst local binary patterns fared even worse with an accuracy of 35%.

<table>
<thead>
<tr>
<th></th>
<th>Frontal face (accuracy %)</th>
<th>Sideways Face (accuracy %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viola Jones</td>
<td>100</td>
<td>97.5</td>
</tr>
<tr>
<td>Local Binary Patterns</td>
<td>42.5</td>
<td>35</td>
</tr>
</tbody>
</table>

*Table 3 Frontal and sideways accuracy*

*Figure 14 Accuracy of Viola Jones against Local Binary Patterns*
Because of this, the system ensured that the user always faced the camera with a frontal face. This section produces a comparison of the following 4 results obtained using the system:

1. SIFT with Bag Of Words
2. SURF with Bag Of Words
3. PCA with BRIEF
4. PCA with Harris Edge Detector

### 6.4 Testing using Support Vector Machines

SVM’s are a leading approach to pattern recognition. SVM’s attempt to find the hyper plane that maximizes the margin between positive and negative observations for a specified class. The system uses CvSVM, OpenCv’s version of the SVM based on LibSVM. The choice of the kernel for a Support Vector Machine is of great importance. After using the 3 different kernels:

1. Linear Kernel
2. Polynomial Kernel
3. Radial Basis Function (RBF) Kernel

It was evident that the RBF kernel yielded the most accurate results. The RBF kernel is given by:

\[
K(x, x_i) = \exp\left(-\frac{||x-x_i||}{\sigma}\right)
\]

The data below is the result from the use of the SVM along with the different approaches considered. The data has been presented in a confusion matrix where the true positives are shown along with the false positives determined by the use of the classifier. Testing occurred on a set of 40 different individuals portraying the 4 different emotions.

#### SIFT with Bag of Words:

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Truth</th>
<th>Happy</th>
<th>Anger</th>
<th>Sad</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>35</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>34</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Sad</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Total Rate</td>
<td>87.5</td>
<td>85.0</td>
<td>75.0</td>
<td>87.5</td>
<td></td>
</tr>
<tr>
<td>Average Rate</td>
<td><strong>83.75</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 4 SIFT Confusion Matrix - Support Vector Machines*
### SURF with Bag of Words

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happy</td>
<td>Anger</td>
<td>Sad</td>
<td>Neutral</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>35</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>36</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Sad</td>
<td>3</td>
<td>2</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Neutral</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td><strong>Total Rate</strong></td>
<td>87.5</td>
<td>90.0</td>
<td>85.0</td>
<td>87.5</td>
</tr>
<tr>
<td><strong>Average Rate</strong></td>
<td></td>
<td></td>
<td></td>
<td>87.50</td>
</tr>
</tbody>
</table>

Table 5 SURF Confusion Matrix - Support Vector Machines

### PCA with BRIEF

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happy</td>
<td>Anger</td>
<td>Sad</td>
<td>Neutral</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>32</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Anger</td>
<td>2</td>
<td>28</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Sad</td>
<td>3</td>
<td>7</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>Neutral</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total Rate</strong></td>
<td>80.0</td>
<td>70.0</td>
<td>80.0</td>
<td>75.0</td>
</tr>
<tr>
<td><strong>Average Rate</strong></td>
<td></td>
<td></td>
<td></td>
<td>76.25</td>
</tr>
</tbody>
</table>

Table 6 PCA with BRIEF Confusion Matrix - Support Vector Machines
The classifier was highly successful compared to the others. It was second only to the Harris Edge Detector with PCA where face landmark points were manually located and fed to the classifier. Using the clustering technique of K means clustering in the Bag of Words approach, the classifier was highly accurate in distinguishing the emotion happy in majority of the subjects.

From the data above, the most successful technique used was the Principal Component Analysis with the Harris Edge Detector, with an average accuracy rate of 81.88%. Though this technique was exhaustive and time consuming as facial landmark points had to manually located and fed to the classifier.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happy</td>
</tr>
<tr>
<td>Happy</td>
<td>37</td>
</tr>
<tr>
<td>Anger</td>
<td>1</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>2</td>
</tr>
<tr>
<td>Total Rate</td>
<td>92.5</td>
</tr>
<tr>
<td>Average Rate</td>
<td></td>
</tr>
</tbody>
</table>

*Table 7 PCA with Harris Edge Detector Confusion Matrix - Support Vector Machines*
The technique that was most successful and accurate in classifying emotions was the Principal Component Analysis with the Harris Edge Detector. This technique, as stated in section 4.3.5, manually finds particular facial landmarks and using those feeds relevant inputs to a classifier over a range of varied subjects. The landmarks can be seen in figure 9. On average it gave an accuracy of 90%. In particular this gave an accuracy of 92.5% for detecting the happy emotion, the highest from any technique and classifier. The lowest accuracy was given by using PCA with BRIEF. Because BRIEF features are comprised of Binary data, applying PCA did not provide much gains in accuracy. As it can be seen, PCA works best with continuous data. It provided an average accuracy of 76.25% with the lowest value being 70% for detecting the angry emotion where it confused majority angry expressions with sad. SURF and SIFT both fared relatively well with the K means clustering approach of Bag Of Words, with the latter performing better providing an average accuracy of 87.5% whilst SIFT provided an average accuracy of 83.75%.
6.5 Testing using Naive Bayes Classifier:

SIFT with Bag of Words:

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happy</td>
</tr>
<tr>
<td>Happy</td>
<td>32</td>
</tr>
<tr>
<td>Anger</td>
<td>1</td>
</tr>
<tr>
<td>Sad</td>
<td>2</td>
</tr>
<tr>
<td>Neutral</td>
<td>4</td>
</tr>
<tr>
<td>Total Rate</td>
<td>82.5</td>
</tr>
<tr>
<td>Average Rate</td>
<td></td>
</tr>
</tbody>
</table>

*Table 8 SIFT Confusion Matrix - Naive Bayes*
### SURF with Bag of Words:

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Happy</th>
<th>Anger</th>
<th>Sad</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>36</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>35</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Sad</td>
<td>1</td>
<td>2</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>Neutral</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Total Rate</td>
<td>90.0</td>
<td>87.5</td>
<td>82.5</td>
<td>82.5</td>
</tr>
</tbody>
</table>

**Average Rate:** 85.63

*Table 9 SURF Confusion Matrix - Naive Bayes*

### PCA with BRIEF

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Happy</th>
<th>Anger</th>
<th>Sad</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>31</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Anger</td>
<td>2</td>
<td>29</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>3</td>
<td>2</td>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>Neutral</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>Total Rate</td>
<td>77.5</td>
<td>72.5</td>
<td>82.5</td>
<td>72.5</td>
</tr>
</tbody>
</table>

**Average Rate:** 76.25

*Table 10 PCA with BRIEF Confusion Matrix - Naive Bayes*

### PCA with Harris Edge Detector

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Happy</th>
<th>Anger</th>
<th>Sad</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>35</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Anger</td>
<td>2</td>
<td>36</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>2</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Neutral</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>Total Rate</td>
<td>87.5</td>
<td>90.0</td>
<td>85.0</td>
<td>87.5</td>
</tr>
</tbody>
</table>

**Average Rate:** 87.50

*Table 11 PCA with Harris Edge Confusion Matrix - Naive Bayes*
Figure 17 Comparing the accuracy rates for Naïve Bayes Classifier

The Naive Bayes Classifier performed similarly to the Support Vector Machine. PCA with Harris Edge gave the highest average accuracy with 87.5 %. SURF performed well with an average accuracy of 85.625%. These two gave the highest individual accuracy of 90% each, PCA with Harris Edge performing well on the angry expression whilst SURF performing well the happy expression. PCA with Brief performed the worst with an average accuracy of 76.25% whilst SIFT performed relatively well with an average accuracy of 81.88%

6.6 Evaluations on findings:

In general the Support Vector Machine performed better than the Naive Bayes Classifier. The SVM has an average accuracy rate of 84.375% whilst the Naïve Bayes gave an accuracy rate of 82.81% over all the training techniques. SVM gave the highest accuracy rate of 90% with the Harris Edge Detector using PCA. A trend from the data can be seen where the training techniques like PCA with Harris Edge and SURF give the highest recognition rates for both the classifiers whilst the lowest observed accuracy rate was given by BRIEF using PCA. Hence the technique which gave a high rate with SVM also gave a high rate with Naive Bayes and vice versa. But the Support Vector Machines outperformed the Naive Bayes classifier in each of the techniques, or at least matched it's accuracy as depicted in the use of PCA with BRIEF.
Figure 18: Accuracy of Support Vector Machines against Naive Bayes Classifier
Figure 19: Accuracy of Happy Support Vector Machines against Naive Bayes Classifier

Figure 20: Accuracy of Angry Support Vector Machines against Naive Bayes Classifier
Figure 21 Accuracy of Sad Support Vector Machines against Naive Bayes Classifier

Figure 22 Accuracy of Neutral Support Vector Machines against Naive Bayes Classifier
Further testing and research suggests that each and every emotion in the extreme could be confused to be the neutral emotion. Hence forth with each emotion and each technique, some of the emotions have always been misclassified as neutral. This occurs when a person has a slight smile and a barely noticeable frown etc. Hence only determining emotions using just facial cues might not be the most viable approach. Other features such as the hand movement or a person’s voice need to be taken into account and using all these features, a mind is able to accurately perceive the human emotions.
Conclusions

This chapter discusses the conclusions and limitations of the project. The project set out to investigate the impact of various techniques in classifying human emotion. It also set out to use the classified emotion to sort out music playlists for a person. This chapter also performs an evaluation on the approach taken. The lessons learnt during the course of this project are also discussed and finally the chapter concludes with the possible future directions of the project.

During the development of the project, existing techniques for feature extraction and emotion recognition were thoroughly researched, highlighting the benefits and problems with each associated. After thorough research 3 feature extraction techniques were decided upon and further techniques to improve classification accuracy were chosen. Further research dictated the use of the Support Vector Machines as the classifier to be used.

The evaluation of the project indicates that the preliminary objectives of the project have been met. The key objectives were:

1. Implement basic feature extraction algorithms (SIFT and SURF) to determine their impacts on an emotion recognition system.
2. Extract information from a music file to sort and classify a music playlist based on human emotion

Added functionality implemented for the project included using various techniques to improve the accuracy of the classifier. The two proposed models: PCA and Bag Of Words model, were implemented and their relative impacts were analyzed. Thorough testing showed that the use of these models vastly improved accuracy, the use of PCA in particular produced the highest accuracy rates in emotion classification.

This chapter further details the limitations of the project, the issues faced and the lessons learnt during the project development.

7.1 Limitations:

The biggest limitation of the project is in the general approach taken for emotion recognition using facial cues. The camera must take a full frontal image of the user to determine the emotion of the user accurately. Even though rotation invariance is taken into account for two of the feature extractors, accuracy rates of subjects facing the camera sideways was extremely low in comparison to frontal images.
Indeed SIFT depicted an average accuracy of 63% whilst SURF fared slightly better with an average accuracy rate of 68%.

Another massive limitation faced during the testing and development of the project was a relatively small sized training dataset. For ethical reasons express permission is required to use a database of faces depicting various emotions. Despite constant efforts only the Yale Face Database granted permission. Hence forth a small training data set was used which resulted in a relatively low accuracy rate for the system.

To develop a music information retrieval system, initially the library Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) was determined to be appropriate. The library turned out to be unstable and within a few weeks of development, it was unfeasible to use this with windows and OpenCV. Hence the project shifted from a machine learning approach to a meta data extraction for music classification. This was another hindrance and limitation faced during the project development.

7.2 Conclusions

In summary it is clear to see that the project set out and achieved its basic functionality. The feature extraction techniques could evidently be improved and much more advanced techniques exist to improve the accuracy of an emotion recognition system. Within this project there is scope for more work that could allow for the complexity of the modelling to be increased, which could evolve the result of the binding between the two main domains of current research: Music theory and Image analysis.
Bibliography


[6] Yale face database B


[8] Improved SIFT Matching for Pose Robust Facial Expression Recognition Hamit Soyel and Hasan Demire

[9] Emotion Recognition from an Ensemble of Features Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S. Huang, Xutao Lv and Tony X. Han

[10] Improved Emotion Detection by Regression Algorithm with SURF Feature and SVM Navdeep Kaur, Er. Varinderjit Kaur


[12] SURF: Speeded Up Robust Features Herbert Bay , Tinne Tuytelaars and Luc Van Gool


[14] Evaluation of SIFT and SURF using Bag of Words Model on a Very Large Dataset Kashif Ahmad, Nasir Ahmad
