Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Hedda Mathilde Sæther Langvik

Uniting Music and Painted Art Using Emotion Categories and Metadata

Master's thesis in Computer Science
Supervisor: Björn Gambäck
May 2022

NTNU
Kunnskap for en bedre verden

Hedda Mathilde Sæther Langvik

Uniting Music and Painted Art Using Emotion Categories and Metadata

Master's thesis in Computer Science
Supervisor: Björn Gambäck
May 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU

Kunnskap for en bedre verden

Abstract

Auditory art may provoke auditory stimuli, and visual art may provoke visual stimuli. Both of these types of stimuli can evoke emotions in the observer. Looking at paintings while listening to music can be even more influential than simply enjoying one art form. This Master's Thesis researches a way to unite auditory and visual art through emotions. The motivation for this Master's Thesis links to music and art's effect on people and ways to provoke certain feelings using these two art forms. The system described in this thesis may be helpful, for example, in selecting the correct paintings and music for a doctor's or therapist's waiting lounge and setting the patients in the correct emotional space. Music platforms such as Spotify may also use the system to display a suitable painting to the listener that should evoke the same emotion as the song.

A system is created which receives a song ID from TheAudioDB database as input and provides paintings from different image datasets as output. Russell's four quadrants (Q1 – happy, Q2 – angry, Q3 – sad, and Q4 – relaxed) which are based on valence and arousal, provided the foundation of emotion classification. The song is categorised into one of Russell's quadrants based on its metadata. A dataset containing information about 900 songs and their selected quadrants are used to train a model that can categorise never-before-seen songs based on their metadata.

Two image datasets are used. The first is WikiArt Emotions, which includes over 4000 paintings and pictures and metadata about the images' labelled emotions. The second dataset includes photographs of landscapes in different seasons and is created from scratch using public images from Flickr. Image-to-image translation with CycleGANs is used to transform the photographs into Monet-like paintings. The images from both datasets are categorised into Russell's quadrants. The emotion labels have been used to determine the images' quadrant in the first dataset. For the second dataset, a hypothesis is used as the foundation in quadrant categorisation. The hypothesis states that summer and spring landscapes fit well into Q1 and Q4, while autumn and winter landscapes are better suited in the Q2 and Q3 quadrants. The results from the thesis slightly support this hypothesis.

The system has been evaluated through a user survey. Five songs were selected to test the system, and a total of four images were selected as the output for each test song. The results show that the participants disagree with most of the system's song categorisations. Only one song received the same quadrant from the system and the survey participants. Some interviewed participants mentioned that it was difficult to pair modern pop songs with paintings from the last century. More tuning of system parameters and better use of datasets could improve this technology and create a fun and exciting way to pair music with art.

Sammendrag

Kunst i form av lyd kan fremprovosere audiell stimulus og visuell kunst kan provosere frem visuell stimulus. Begge disse typene av stimuli kan fremkalle følelser hos den som lytter eller observerer. Å se på malerier mens du lytter til musikk kan virke enda mer innflytelsesrikt enn å nyte kun én av kunstformene. Denne masteroppgaven prøver å finne en måte å forene auditiv og visuell kunst gjennom følelser. Motivasjonen for denne oppgaven handler om musikkens og kunstens effekt på mennesker og måter å fremprovosere visse følelser ved å bruke disse to kunstformene. Systemet som er beskrevet i denne oppgaven kan være nyttig for eksempel for å velge riktige malerier og musikk for en leges eller terapeuts venterom for å sette pasientene i det rette humøret. Musikkplattformer som Spotify kan også bruke dette systemet for å vise frem et passende maleri til lytteren som skal gi de samme følelsene som sangen.

Det er opprettet et system som mottar en sang-ID fra TheAudioDB-databasen som input og gir malerier fra ulike datasett som output. Russells fire kvadranter (Q1 – glad, Q2 – sint, Q3 – trist og Q4 – avslappet) som er basert på "valence" og "arousal" har lagt grunnlaget for klassifisering av følelser. Sangen er kategorisert i en av Russells kvadranter basert på metadataene. Et datasett som inneholder informasjon om 900 sanger og deres valgte kvadranter er brukt til å trene opp en modell som kan kategorisere aldri-før-sett sanger basert på deres metadata.

To datasett med bilder er brukt. Den første er WikiArt Emotions, som inkluderer over 4000 malerier og bilder og metadata om bildenes følelser. Det andre datasettet inkluderer fotografier av landskap i forskjellige årstider og er laget fra bunnen av ved hjelp av offentlige bilder fra Flickr. Bilde-til-bilde-oversettelse med CycleGAN har forvandlet disse fotografiene til Monet-lignende malerier. Bildene fra begge datasettene har blitt plassert i Russells fire kvadranter. Informasjonen om følelsene har blitt brukt til å bestemme bildenes kvadranter for det første datasettet. For det andre datasettet ble en hypotese brukt som grunnlag i kategorisering. Hypotesen sier at sommer- og vårlandskap passer godt inn i Q1 og Q4, mens høst- og vinterlandskap er bedre egnet i kvadrantene Q2 og Q3. Resultatene fra oppgavens system støtter forsiktig opp under denne hypotesen.

Systemet er evaluert gjennom en brukerundersøkelse. Fem sanger ble valgt ut for å teste systemet. Totalt fire bilder ble valgt som output for hver testsang. Resultatene viser at deltakerne er uenige med systemet i de fleste av kategoriseringene til sangene. Bare én sang fikk samme kvadrant fra systemet og deltakerne i undersøkelsen. Noen av de intervjuede deltakerne nevnte at det var vanskelig å pare moderne poplåter med malerier fra det forrige århundre. Mer justering av parametere og bedre bruk av datasett kan forbedre teknologien og skape en morsom og spennende måte å koble musikk med kunst.

Preface

This Master's Thesis was written during the spring semester of 2022. It is the final work of achieving the Master of Science degree from the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. The work has been supervised by Professor Björn Gambäck and conducted within the Data and Artificial Intelligence Group at the Department of Computer Science.

Professor Björn Gambäck deserves a special thank-you for his guidance and excellent feedback throughout the semester. Even though some of this research is outside his area of expertise, he has shown interest in my work and assisted me in any way he deemed possible. I have always enjoyed different art forms, and diving deeper into the area of computational creativity has been most inspiring.

Lastly, my time in Trondheim and at NTNU would not have been the same without the friendships I have made. A special thank-you to Abakus, the student association for Computer Science, is compulsory. My interest in music and art has been reinforced after five years in Abakusrevyen, and I will keep the memories with me as long as I live.

Hedda Mathilde Sæther Langvik

Trondheim, 13th May 2022

Contents

List of Figures

List of Tables

Acronyms

API Application Programming Interface. 23, 26, 27, 29, 35, 68, 77, 78, 85

CC Computational Creativity. 5, 7, 78

GAN Generative Adversarial Network. 5, 7–9, 15, 40, 69, 84

H-creativity Historical creativity. 7

IAPS International Affective Picture System. 27

MER Music Emotion Recognition. 5, 8, 11, 13, 14, 84, 85

MVP Minimum Viable Product. 4, 77, 79, 83

P-creativity Psychological creativity. 7, 78

SAD Seasonal Affective Disorder. 38

VA Valence and Arousal. 8, 13, 80, 83, 84

XGBoost eXtreme Gradient Boosting package. 10, 25, 30, 36, 65

1. Introduction

This Master's Thesis explores ways to combine existing art generation tools with music emotion recognition to create art that matches a song's emotions.

This chapter will first describe the background and motivation for this thesis. Next, the research goals and research questions are presented with some comments. Furthermore, the research method and main contributions of this project are given.

1.1. Background and Motivation

Music and paintings are two different art forms that are supposed to evoke an emotion in the observer. If a painting gives emotion and music gives emotion, what will happen if the observer listens to music while watching a painting? Will the emotional perception be more vital? Will it confuse the observer or enhance their experience?

This research aims to find a mapping between visual stimuli (paintings) and audio stimuli (music) to optimise their impact on a person's mood state. Hopefully, this will influence a person who looks at a painting whilst listening to calming music and give them a sense of relaxation and happiness. This feeling should be more potent or more precise than if the person was experiencing only one of the two forms of stimuli. The aim is also to see if there is any difference in the strength between paintings that the audience claim to give a particular emotion and a painting of natural environments in a specific season. Will a painting of a wet autumn day be perceived as more gloomy than a modern painting labelled with this mood?

A program translating a music piece to a painting can be used to investigate the statement and question in the last paragraph. This program could also be used as an artistic tool to aid painters or musicians in their creative work. A painter can use their favourite music as input and see what the program outputs to recreate the given song in a painting. Painter Mark Rothko was highly influenced by music and said that he wanted to raise painting to the level of the poignancy of music (Sarno, 2006). Rothko would paint in rooms filled with music to evoke the same energy and emotions that the music gave him. Hence, a tool that pairs music and art based on emotions could be of great use to painters like Rothko.

Figure 1.1.: Example of Spotify Audio Aura, retrieved from Spotify's newsletter.

On the other hand, a musician can use this tool to analyse other songs and then use the outputted paintings as inspiration for a new musical piece. In an interview, music artist Billie Eilish revealed that she sees all her songs in colours[1]. This condition is called synaesthesia. Her album "Happier Than Ever" has a beige theme all over, as seen on the cover, music videos, and advertisements. Therefore, it is clear that music and painted art have a connection and that a tool or program like the one created in this project can be valuable to enhance or elevate the level of creativity in both art forms.

Additionally, the program can be used by music platforms such as Spotify to present some fun elements to the listener. In 2021, Spotify introduced a new feature where users could see their audio aura for the previous year[2]. Essentially, it provides an image with different colours where each colour represents a mood or theme. For instance, the colour green reflects calm, analytical, and introspective moods. Figure 1.1 shows an example of this image. A program like the one described in this project could provide more concrete art to the listener and give one art piece for each song and not only once a year. It can also be useful in smart TVs or gadgets like a Chrome Cast or Apple TV to display a more interesting image while playing music on the telly, replacing the standard display.

[1] Retrieved from interview: https://youtu.be/uItbMBBHFmo.
[2] Retrieved from article: https://newsroom.spotify.com/2021-12-01/learn-more-about-the-audio-aura-in-your-spotify-2021-wrapped-with-aura-reader-mystic-michaela/.

Lastly, the program may be used by concert halls worldwide so that deaf and hard of hearing people can enhance their musical experience. Even though their hearing is impaired, they may still enjoy music and use the outputted painting and the songs' vibrations to understand the music better. The program can either analyse the music in advance and display it to the audience when it is played or demonstrate live generation.

1.2. Goals and Research Questions

This project aims to combine two art forms – painted art and music, that should unite in emotion. This can be done by creating a system that receives a song as input and delivers a painting as output. The system should be able to categorise the song's emotion(s) and then select or create a suitable painting that reflects the same emotion. Hence, Music Emotion Recognition (MER) and image generation must be utilised. The goal of this Master's Thesis is as follows:

Goal Unite visual and auditory art so that a music piece and an art piece share the same perceived emotion.

In order to verify that this goal is met, a user study will be completed, asking whether the visual and auditory art pieces match well the emotions portrayed to the participant.

Research question 1 What meteorological seasons couple best with which emotions?

A common perception is that a sunny summer is connected to happiness, and dark winter is connected with sadness (Watson, 2000). Is this, in fact, the case with a painting displaying a landscape in the summertime? Each season is coupled with a quadrant from Russell's model to test this (Russell, 1980), and questions in a user survey will disclose whether the participants agree with this common perception or not.

Research question 2 Will users prefer generated seasonal landscape paintings or original paintings labelled with emotions when listening to music?

This project aims to match a song with two different paintings. The first painting will be an original painting that already has emotion labels, and the second will be a landscape photo that will be translated into a painting. A question in the user survey can unveil which of the two a user prefers or which they feel matches best with the song, given the emotion they perceive from the song.

Research question 3 How can the system that pairs art and music be analysed and evaluated?

After the experiments, the results from *Research Questions 1 and 2* come in handy. The

system can be evaluated using these questions to find answers. In addition, the user survey could help research the music and art pairs in themselves. Lastly, the evaluation of the system should include all the steps taken in the experiments to ensure proper analysis.

All the research questions will be focused on throughout the project. The answers to these questions will be discussed in Chapter 8.

1.3. Research Method

Some research and relevant theories used in this project were initially gathered in the specialisation project from the fall. This fall project looked at the state-of-the-art techniques to generate art, particularly landscape paintings, and if there was a way to base the output on musical input. All the findings have been used as preparation for this Master's Thesis, including experiments and tests. However, more knowledge is needed before the experiments can start.

In order to answer the research questions, a theoretic and an experimental methodology will be applied. The research method for the theoretic part is mainly constrained to Snowballing sampling and citation networks. This technique consists of reading a report that includes some interesting findings regarding emotion recognition in art or music, image generation, or a combination of the two, and then checking out the references to dig deeper into the topics. The reverse is also possible: when an article is considered relevant, Google Scholar's search engine may be used to see which other articles have referenced the relevant article. The literature list in this report is put together using all of these methods.

In the experimental part, a system must be built to try different parameters or datasets. The system implementation process will start by building a working prototype or a minimum viable product (MVP). This MVP will take in a song and return an image. The next step is to analyse this and tweak parameters to reach a final product. Eventually, different experiments will be carried out using different datasets according to the experiment plan. The final step is to evaluate the results through user surveys and interviews.

1.4. Contributions

The main contributions to this Master's Thesis are listed below.

1. Implement a system using state-of-the-art MER data to categorise songs into

Russell's quadrants.

2. Implement a system that reuses state-of-the-art image-to-image translation with CycleGAN to create new landscape paintings based on photographs.

3. Implement a system that translates from song to image using Russell's circumplex model of affect as the foundation of emotion.

1.5. Thesis Structure

The rest of the thesis is structured as presented below.

1. In Chapter 2, the theory that is necessary to understand the relevant topics of this thesis is provided. It covers Generative Adversarial Network, Music Emotion Recognition and Computational Creativity, among other things.

2. In Chapter 3, some of the state-of-the-art techniques, previous experiments and other research related to music emotion recognition and painting generation are presented. The image-to-image translation technique used in the experiments is explained in detail here.

3. In Chapter 4, the datasets that are used in the experiments and the necessary transformation of the data are described. This chapter also presents some datasets or databases that were not used in this project due to license fees or lack of interesting data.

4. In Chapter 5, an overview of the system's architecture and the design approach used in this work are given. It reviews the mapping from song and painting to quadrants and presents the system's flow.

5. Chapter 6 details the experiments conducted, and their results are reviewed. This chapter provides the results from each step and shows some of the resulting song-image pairs.

6. In Chapter 7, the setup of the user survey is presented, and some of the results are discussed.

7. In Chapter 8, the answers and analysis of the user survey from Chapter 7 are used to evaluate the system. The findings, limitations and possible improvements are discussed. Answers to the research questions are also presented.

8. Finally, Chapter 9 concludes the project by presenting its main contributions, and some possible future work is discussed.

2. Background Theory

This chapter provides necessary background information about the technology, techniques and terminology used in the project and the report. The topics in this chapter do not have much in common, but all are important to understand before moving on to the following chapters. First, Section 2.1 defines computational creativity and different forms of creativity. Computational creativity is important to understand because it will be discussed in Chapter 8. Next, Section 2.2 presents some details on previous work on emotion classification that is good to know before reading more about this topic in Chapter 3. The background theory on this topic provides some information on how a song can be categorised and what Russell's quadrants are. In Section 2.3, Cycle-consistent adversarial networks are described, and their purpose is presented. Details about the image-to-image translation technique used in this project are also provided. This ensures some understanding of GANs before Chapter 5 presents how this has been used in this project. Lastly, Section 2.4 presents some basic machine learning information that is nice to know before reading Chapter 4. Note that some of the paragraphs in Section 2.2 and Section 2.3 are revised and renewed from the specialisation project from the fall.

2.1. Computational Creativity

Computational Creativity (CC) is a term used when computers generate a result that would be considered creative if it was produced by humans alone (Besold et al., 2015). Colton et al. (2012) define CC as the engineering and philosophy of computational systems. They say that, in some way, it exhibits manners that any random observer would deem to be creative. Colton et al. describe computational creativity as a subfield of Artificial Intelligence research working with computational systems to create new ideas.

According to Boden (1998), creativity is a feature of human intelligence, and a creative idea is something novel, surprising, and valuable. Creative ideas (novelties) which are novel only to the mind of the individual concerned are called P-creativity, where P is for psychological (or personal). If the idea is novel to the entire world and the whole of the previous history, it is called H-creativity, where H is for historical. Artificial intelligence should focus mainly on the former (Boden, 1998). There are three main types of creativity; combinational, exploratory, and transformational. The first type, combinational creativity, involves novel combinations of familiar ideas. Exploratory

creativity involves the exploration of structured conceptual spaces that result in new structures or ideas that are somewhat unexpected. Lastly, transformational creativity involves some transformation of the structure's dimension so that new structures can arise that were impossible before (Boden, 1998).

2.2. Emotion Classification and Music Emotion Recognition

Music Emotion Recognition (MER) is a widely used term describing the essence of understanding what emotion a musical piece evokes in the listener. Emotion in music can be studied in three different ways. The first is how the listeners perceive emotion. This means the emotion an individual identifies with when listening to music. The second is the felt emotion, which describes the emotional response a listener can feel inside when listening. The felt emotion may be different from the perceived one. For instance, a person may listen to a song and perceive it as a sad song but feel calm or glad due to some memory they connect to it. Last is the transmitted emotion, representing what the artist or composer wanted to convey to the listener.

There are different ways to extract emotion-related data from a musical piece, either through music features or ground truth data (Yang et al., 2018). Panda et al. (2018) used music features to extract emotion information from songs. They used 30-second clips of songs and used different algorithms to analyse features like tempo, rhythm and tonality. Their work is described in more detail in Section 3.2. Ground truth data is derived from emotion labels given by human beings to a piece of music and is therefore defined as the perceived emotions of humans. People are different, and hence, the resulting emotion labels may also vary for the same song. Combining these two to get both human and computer analysis of the same song is possible.

Classifying emotions is possible using arrays describing an emotional dimension model. The most commonly used dimension is a valence arousal plane, or a VA plane, such as the one from Russell's model in Figure 2.1. The valence scale ranges from pleasant to unpleasant, and the arousal scale ranges from calm (deactivation) to excited (activation). The two dimensions describe disconnected feelings that may not necessarily be felt towards a specific situation (Cowen and Keltner, 2017). When using dimensional emotion models, a musical piece will receive values on their level of valence and arousal, e.g., from -10 to 10 (Russell, 1980; Grekow, 2021; Aljanaki et al., 2017).

2.3. Cycle-consistent Adversarial Networks

Generative adversarial networks (GAN) are commonly used in image synthesis, style transfer, semantic image editing, and classification (Creswell et al., 2018). A forger-expert

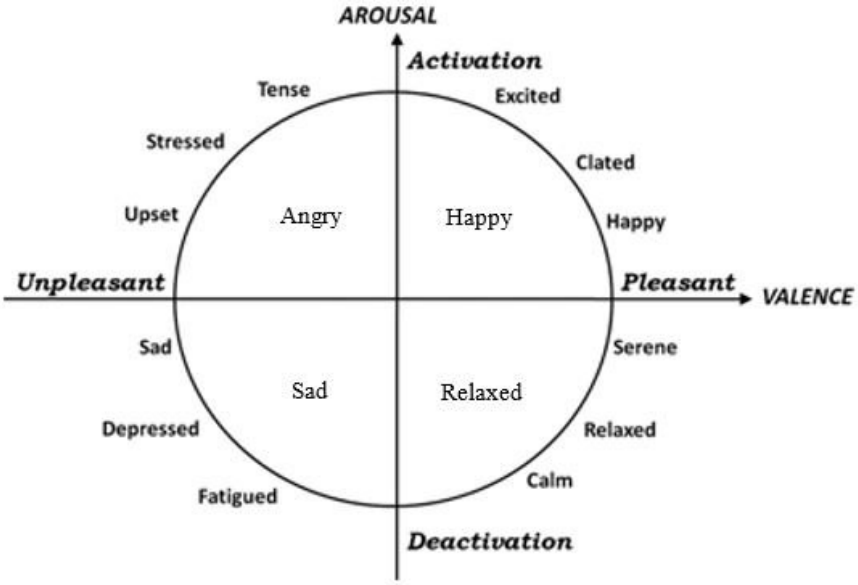


Figure 2.1.: Russell's circumplex model of affect. Retrieved from Bajada and Bonello (2021) with creative common license.

analogy is often used to depict the functionality, as shown in Figure 2.2. A forger, known as the *Generator*, shown as G in the figure, creates forgeries of art. Then the expert, known as the *Discriminator*, or D in the figure, receives some input and tries to distinguish genuine and authentic artwork from the generator's forgeries. The discriminator has access to authentic images and synthetic samples and answers to whether the input was real or fake. The generator uses the answers from the discriminator to learn to create better forgeries.

Generative Adversarial Network can be described as a two-player machine learning competition. The aim is to train a system to learn a loss function that recognises what is authentic and not. These loss functions may differ depending on what the generator is trying to achieve. Typically, the GAN uses a minimax loss function where the generator wants to minimise the loss, and the discriminator tries to maximise it.

CycleGAN introduces a cycle consistency loss that tries to preserve the original image after a cycle of translating and reverse translating it. A popular analogy is to describe the system as a language translator. When translating a sentence from English to French and then back to English again, the output should ideally be the same sentence as the input. The same goes for images. This removes the need for matching image pairs, making data preparation easier. Style transfer is possible due to this, e.g., transforming a photo picture into a Monet-like painting or adding snow and winter features to a picture taken in the summer (Zhu et al., 2017).

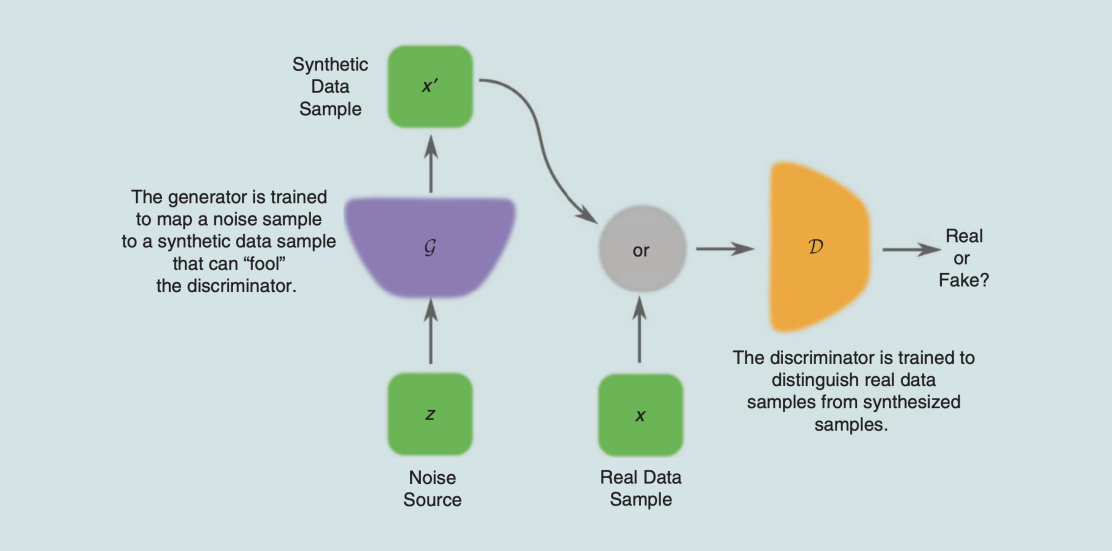


Figure 2.2.: Architecture of Generative Adversarial Networks. Reprinted, with permission, from Creswell et al. (2018). © 2021 IEEE.

2.4. Training and Testing with Datasets

As presented in Chapter 4, there are many suitable datasets with metadata about music and paintings with information about their emotions or moods. These datasets can be used for training to predict the same information in new songs or paintings. Different machine learning techniques may be used for this process, such as XGBoost. XGBoost is short for the "eXtreme Gradient Boosting package" (Chen et al., 2015). XGBoost is used for model improvement and is capable of handling missing values. It is also good with categorical encoding and sparse data. Data transformation is needed to achieve categorical encoding on datasets with text fields. This has been done for some of the datasets used in this project, and their transformation is described in Chapter 4.

2.5. F1 Score and Music Feature Extraction Tools

The F-measure or F1 score is a way to measure the performance of a model on a dataset. It uses the precision and recall of a test to calculate the value (Chicco and Jurman, 2020). The F1 score is invariant for class swapping and is independent of the number of samples correctly classified. Some criticism has been made about the measuring tool (Hand and Christen, 2018), but it is still widely used in most machine learning applications (Chicco and Jurman, 2020).

Three Music Emotion Recognition (MER) tools are mentioned in Chapter 3, MIR Toolbox, Marsyas, and PsySound. These tools can extract music features from audio files, which may be used in MER. The first tool, Mir Toolbox, is written in Matlab and provides an integrated set of functions to extract music features from audio files[1]. Marsyas, an acronym for *Music Analysis, Retrieval and Synthesis for Audio Signals*, is written in `C++` and has sound processing modules to extract music features[2]. Lastly, PsySound is a Matlab implemented environment used to analyse audio clips[3].

[1]More information about Mir Toolbox: `https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`.
[2]More information about Marsyas: `https://github.com/marsyas/marsyas`.
[3]More information about PsySound: `http://psysound.org/`.

3. Related Work

This chapter will describe state-of-the-art techniques for music emotion recognition, image generation and connections between music and images. Firstly, Section 3.1 and Section 3.2 detail the process of music emotion recognition and the use of valence and arousal in emotion. Painting generation using GANs is described in Section 3.3. In Section 3.4, other researchers' work on music and art are presented. The paragraphs in Section 3.4 are revised and renewed from the specialisation project from the fall.

3.1. Emotion in the Valence and Arousal Plane

When labelling music with emotion, two categorisation techniques are possible: dimensional and categorical. A categorical technique is stricter than a dimensional one because there only is a number of n categories to choose from. These may be common emotions such as awe, fear and joy (Colton et al., 2012). A dimensional technique provides more freedom because several dimensions can provide more details on each emotion category.

As presented in Section 2.2, the Valence and Arousal plane is a much-used dimensional emotion classification model. In 2020, Bliss-Moreau et al. did a study challenging this model. They were trying to see whether there exists a social dimension to emotion in addition to or instead of valence and arousal. They completed two studies, one looking at emotional words that were more social and one looking at whether priming social information would impact the structure. Figure 3.1 shows the results of the first study, including social-emotional words. Comparing this to Figure 2.1, the exact emotional words, i.e. *happy*, *calm* and *sad*, appear in the same quadrants. Hence, their results concluded that the priming of social information did not influence the dimensional model. The two-dimensional scale with valence and arousal best described the structure of emotion.

3.2. Music Emotion Recognition

Panda et al. (2018) researched new ways to advance the state-of-the-art Music Emotion Recognition (MER) using the music feature techniques presented in Chapter 2. They

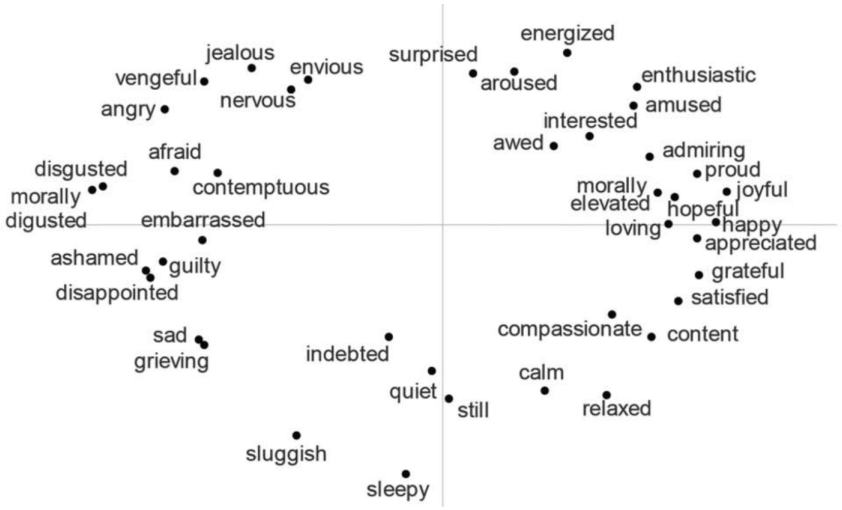


Figure 3.1.: Bliss-Moreau et al.'s research on valence and arousal in the foundation of emotion. Reprinted, with permission, from Bliss-Moreau et al. (2020).

introduced novel emotion-relevant audio features, which improved the F1 score compared to using only baseline features (Section 2.5). The authors wanted to introduce new audio features because many of the existing ones in MER were insufficient in their emotional relevance because they were created for other audio recognition applications (Panda et al., 2018). As indicated in Section 2.2, emotion may be studied as three different things: perceived, felt and transmitted, and Panda et al. focused on the first alternative in their study. The frameworks MIR Toolbox, Marsyas, and PsySound, presented in Section 2.5, were used to extract 1702 standard audio features, filtered based on their correlation according to the ReliefF feature selection algorithm (Panda et al., 2018). They introduced several novel features, for instance, *Note Smoothness Statistics* and *Register Distribution* within the category Melodic Features, and *Glissando Presence* and *Vibrato and Tremolo Features* in the category Expressively Features. Next, they tried different baseline and novel feature combinations and found the best combination. Table 3.1 shows the combinations they tried, where the ones in bold indicate the best pick for each category. A combination of 100 baseline and novel features gave the best results, which were used in the testing stages.

Panda et al. (2018) created a dataset to test and evaluate their work. They considered it necessary to create a new dataset as there was no public, widely accepted dataset that was also adequately validated. Hence, it was challenging to compare works. They created this dataset to evaluate their work and for others to use in further research and compare the results on the same dataset. This dataset is described in more detail in Chapter 4. Essentially, it contains metadata about 900 songs and what quadrant they are associated with based on their audio features described in the previous paragraph.

Table 3.1.: A table from Panda et al.'s study shows the results of the Classification by quadrants. Reprinted, with permission, from Panda et al. (2018). © 2018 IEEE.

Classifier	Feat. set	# Features	F1-Score
SVM	**baseline**	70	**67.5% ± 0.05**
SVM	baseline	100	67.4% ± 0.05
SVM	baseline	800	71.7% ± 0.05
SVM	baseline+novel	70	74.7% ± 0.05
SVM	**baseline+novel**	100	**76.4% ± 0.04**
SVM	baseline+novel	800	74.8% ± 0.04

3.3. Generating New Paintings with Adversarial Networks

Isola et al. (2017) investigated how adversarial networks could generate and manipulate images using different image-to-image translation techniques based on GANs. One method used conditional adversarial networks and trained them to learn mappings between input and output images. The technique maps pixel to pixel in the images and works well in image processing and translation (Isola et al., 2017). They need image pairs, or tuples, to do this. An image pair can consist of one image being the outline or sketch of a shoe and the other image being the filled-in ground truth photo of the shoe. The discriminator looks at each patch in image `x` and sees if the image from the generator matches the ground truth patch from image `y`. The process is shown in Figure 3.2. Isola et al. tried different patch sizes and image sizes to see what gave the best results and found that a 70×70 PatchGAN on a 256×256 pixel image was ideal. Their `pix2pix` framework[1] has become very popular, and other researchers and visual artists have modified and used it in other domains.

Some of the researchers from Isola et al. (2017) continued to work on image-to-image translation, now using unpaired data samples (Zhu et al., 2017). Figure 3.3 shows the difference between the two. For instance, digital photography can be transformed into the same image but look like a painting using the styles of famous artists, such as Monet and van Gogh. That means only the style of image Y is applied to the content of image X. The motivation for Zhu et al.'s research was that it might be challenging to find paired datasets, and creating one from scratch is very time-consuming. Hence, unpaired data is more cost-efficient. Zhu et al. introduced cycle consistency in Cycle-consistent adversarial networks (CycleGAN) using unpaired data to train and the 70×70 PatchGAN from Isola et al. (2017) as the discriminator. The essence of cycle-consistent adversarial networks is described in Section 2.3. Zhu et al.'s solution adds both an adversarial loss

[1] `https://github.com/phillipi/pix2pix`

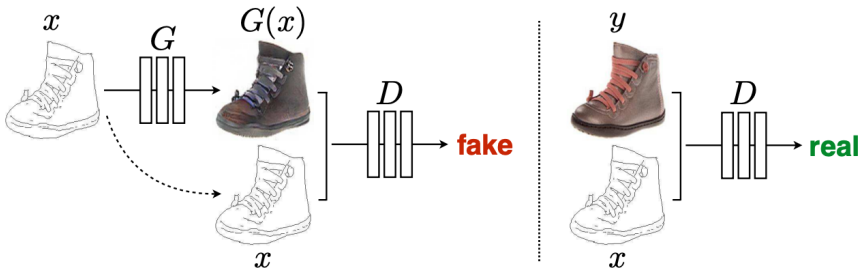


Figure 3.2.: Training a conditional GAN to map edges→photo. The discriminator, D, learns to classify between fake and real tuples. The generator, G, learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe the input edge map. Reprinted, with permission, from Isola et al. (2017). © 2017 IEEE.

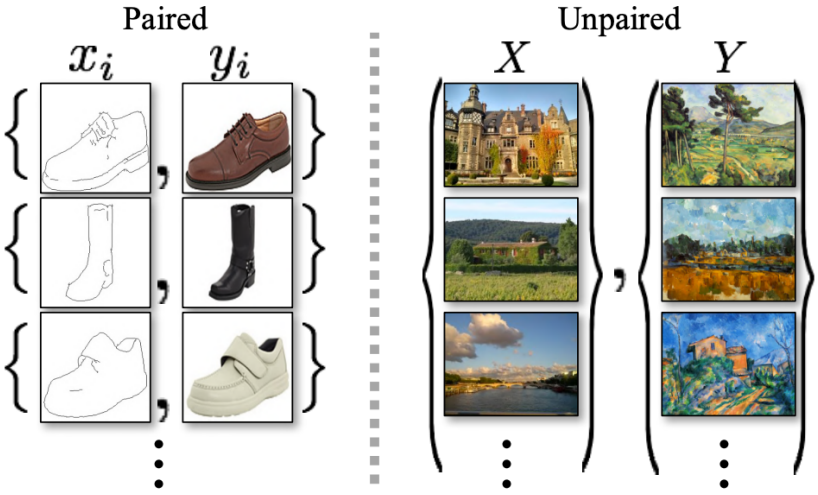


Figure 3.3.: Paired data tuples on the left where there exist a correspondence between x_i and y_i. Unpaired data on the right where there is no information on matches between the two sets. Reprinted, with permission, from Zhu et al. (2017). © 2017 IEEE.

and a cycle-consistent loss to image-to-image translation. It was also intended to be a general-purpose solution that can be modified to fit specific tasks or domains.

The system Zhu et al. created uses four sets of data to complete training and testing. Both training and testing data include two folders each, the training dimension has trainA and trainB, and the testing dimension has testA and testB. A and B, in this sense, correspond to X and Y in the unpaired example of Figure 3.3. First, the system trains on many images with the same subject, for instance, landscape images. If the

(a) A real photo from trainA.

(b) The real photo made to look like a painting.

(c) An actual Monet painting from trainB.

(d) The Monet painting made to look like a photograph.

Figure 3.4.: Recreated examples of the trained model from Zhu et al. (2017).

goal is to translate from landscape photos to paintings, trainA should contain photos of landscapes, and trainB could be Monet paintings. The system will translate images in both directions, making the photo look like a painting and vice versa. Figure 3.4 shows an example of the system when training on photos and paintings.

3.4. Music and Paintings

Zorić (2017) researched the features in music and art and created a system that made abstract computer-generated paintings based on musical input (Zorić and Gambäck,

2018). This system used the Spotify API to collect the musical features and map each feature to a corresponding painting feature. Examples of these mappings are *Tempo* → *Number Of Brush Strokes* and *Energy* → *Colour Degree*. Zorić created a new program where geometric shapes, brush strokes, and image effects were utilised to generate abstract art. When an image was created, evolutionary algorithms were used to test and improve it towards a user-preferred mapping table, freely selecting colours and using user preference without knowing the mapping table. Zorić found that using evolutionary algorithms meant that the system could not sufficiently evaluate the performance through a fitness function; human involvement was necessary. The system was also reliant on user involvement to create aesthetically pleasing and meaningful images. Zorić mentions some possible future work that includes (i) alternative approaches for the music-to-image transformation problem, (ii) using various fitness functions in the evolutionary algorithms, (iii) using other music and image systems, and (iv) interacting with the users in new ways. Numbers (i), (iii) and (iv) are highly relevant to this project.

Aleixo et al. (2021) also created abstract images based on musical inputs. They created a genetic algorithm that generated images using two inputs. The first was an image created randomly. The other was images generated where musical instruments were associated with specific shapes and figures, e.g., piano mapped to circles and guitar to rectangular spots. Their research looked into cross-domain associations to find appropriate mappings between music and image features. The approach mapped the music's melody and harmonies to the image's foreground and background, respectively. After creating images through many iterations, the results were evaluated through a digital survey where participants were asked different questions. The participants rated the images and musical pieces separately, deciding how much they preferred them and what mood they associated with them. Next, they were told that all images were created based on the music and were asked whether they agreed with the mappings. Aleixo et al.'s questioning approach is different from Zorić's because the participants are not told the connection between the music and the images from the start. In Zorić (2017), participants are asked to rate how well each image relates to their respective song and what changes would improve the results. Both studies received answers saying their abstract images fit well with the musical input on which it was based.

4. Datasets

There are many datasets and databases with paintings, photographs, and songs and with metadata about these. This chapter will present the databases and datasets used in this thesis and other useful ones that were not used. Some factors were important when considering whether or not to use a particular database or dataset. First and foremost, the contents and size of the dataset were important. What columns does it include? Does the dataset with songs have information about the song's genre, and if yes, is there more than one genre for each song? Does the database have more than one dataset with songs? Is it possible to join different datasets to get all the valuable information? Moreover, considering paintings, is there any information about the emotions they evoke in the observer? Is there any information about the painting's style or what time area they are from?

One dataset with paintings was chosen, and one dataset and one database were chosen to provide information about songs. The first four sections describe these, and Section 4.5 introduces some of the datasets that were not chosen and why.

4.1. WikiArt Emotions

WikiArt Emotions[1] is a dataset with information about over 4000 images that have annotations of emotions evoked by the observer (Mohammad and Kiritchenko, 2018). This dataset contains mostly paintings with western painting styles, such as post-renaissance art and contemporary art. Annotations were done by crowd-sourcing the paintings. The dataset includes these annotations for three groups; one group looked at only the image, one looked at only the title, and the last looked at both the image and title. The dataset does not contain folders with the images but rather URLs to their sources. Hence, it is possible to find and download every image to recreate the datasets. This project has utilised downloading the images based on their source URLs. Moreover, the pictures have been categorised in Russell's quadrants according to what emotions they were labelled with.

The dataset contains all the annotations of the images but has three different levels of aggregations: 30%, 40% and 50%. Suppose at least n% of the responses indicate that

[1] More information on http://saifmohammad.com/WebPages/wikiartemotions.html.

Table 4.1.: Columns from the WikiArt Emotions dataset from Mohammad and Kiritchenko (2018).

Column	Example 1	Example 2
ID	58c6237dedc2c9c7dc0de1ae	5772843bedc2cb3880fd334d
Style	Modern Art	Contemporary Art
Category	Impressionism	Minimalism
Artist	Charles Courtney Curran	Richard Serra
Title	In the Luxembourg Garden	One Ton Prop (House of Cards)
Year	1889	1969
Is painting	yes	no
Face/body	face	none
Ave. art rating	2.33	−1.7

a particular emotion applies. In that case, that label is chosen. If at least $n\%$ of the respondents indicate a neutral emotion and less than $n\%$ of the respondents indicate any of the other nineteen emotions, then the neutral label is chosen. The different percentages represent that n of 10 people gave an emotional response. Mohammad and Kiritchenko recommend using the 40% aggregation distribution in production, which was therefore selected for this project.

Table 4.1 shows two example rows and their column values from the dataset. Including this metadata, there are also sixty columns with one-hot encoded data on the annotated emotions. There are 19 emotions for each group mentioned earlier, along with a "neutral emotion" category. Hence, each group will have the value 1 in at least one of these 20 columns. Since only the annotations for the image only-group were used, 40 columns that contained annotations for the image and title and title only were removed. Next, all the rows were iterated through to translate from annotated emotion to quadrant.

The simple, original quadrant from Figure 2.1 and the research on Russell's quadrants done by Bliss-Moreau et al. (2020) shown in Figure 3.1 were used to categorise and translate the emotions. The mapping between the 19 emotions and a quadrant is shown in Table 4.2. All images that only had a label with neutral emotion were removed from the dataset because they do not belong to any quadrant. The mapping was done by trying to find a similar word from the WikiArt Emotions and the two figures. For instance, the word "happiness" from WikiArt Emotions was considered the same as "happy" from the two figures, and therefore, "happiness" was assigned to Q1. Table 4.2 shows three columns: the **Emotion** label retrieved from WikiArt Emotions, the **Quadrant** assigned to each emotion, and each emotion's mapping to words from **Russell's model** in Figure 2.1 and Figure 3.1.

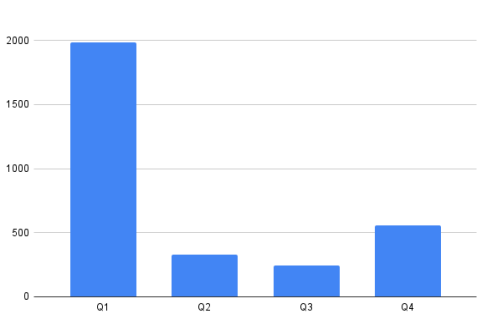
Many things may have gone wrong when mapping the emotions to quadrants. Some word pairs, such as anger – angry and sadness – sad, are so similar that they are easy to

Table 4.2.: Translating the WikiArt Emotion annotations to quadrants.

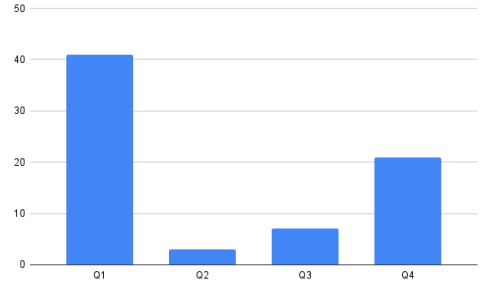
Emotion	Quadrant	Compared to word from Russell's model
Trust	Q1	Admiring
Happiness	Q1	Happy
Optimism	Q1	Hopeful
Anticipation	Q1	Interested
Surprise	Q1	Surprised
Pessimism	Q2	–
Disagreeableness	Q2	–
Fear	Q2	Afraid
Anger	Q2	Angry
Arrogance	Q2	Contemptuous
Disgust	Q2	Disgusted
Shame	Q3	Ashamed
Regret	Q3	Guilty
Sadness	Q3	Sad
Shyness	Q4	–
Agreeableness	Q4	–
Humility	Q4	–
Gratitude	Q4	Grateful
Love	Q4	Loving

compare. For others, a synonym was used to find their matching quadrant, e.g., optimism – hopeful. Furthermore, if no synonym of the feeling was found in the figures, the research from Panda et al. (2018) was used to find a suitable quadrant. Their dataset contains information about the moods of all the songs they analysed. Each song has information about the quadrant and moods associated with the given song. Hence, for the WikiArt Emotions labels "pessimism", "shyness", and "agreeableness", this comparison technique was used. Either the word itself or a synonym was found in Panda et al.'s mood column for several rows, and then that mood was compared to the quadrant of all the rows. The quadrant labelled for most of the rows containing that mood was chosen. The emotion "pessimism" was compared to the mood "negative" from Panda et al., "shyness" was compared to "reserved", and "agreeableness" to the same mood from the dataset. Only two emotions were left to categorise, "disagreeableness" and "humility". None of these words, or a synonym, were found in the earlier mentioned figures nor Panda et al.'s dataset. The former emotion, "disagreeableness", was set to Q2 merely as a guess since it is the opposite of agreeableness. The latter, "humility", was set to Q4 as it may be viewed as the opposite of "proud", which is in Q1. Possible pitfalls with this method are discussed in Chapter 8.

Figure 4.1 shows the distribution of the images from WikiArt Emotions into quadrants. Two different testing groups were created, one containing all the images from the dataset



(a) All images.



(b) Without modern images or images that have faces.

Figure 4.1.: Distribution of images from WikiArt Emotion into quadrants from Russell's model

(Figure 4.1a) and another excluding images labelled with the style "modern" and labelled to include a face (Figure 4.1b). Images are predominant in Q1, containing over 50% of the images in both categories. Q4 also dominates Q2 and Q3, which means that for songs categorised as Q2 or Q3, there are fewer images to choose from.

4.2. Flickr

Flickr[2] was used for the photograph database to gather pictures that would be translated into paintings. It is a photo management and sharing application containing millions of pictures and many of natural landscapes. A disadvantage with this platform is that there is no labelling regarding their emotions, which can be time-consuming and challenging to add manually. It is possible to create an automatic process that iterates through a set of images and labels each one. However, how a program should label each photo is difficult to determine. Some research found that weather may affect a person's mood (Denissen et al., 2008; Ennis and Mcconville, 2004). For instance, temperature and wind power may negatively affect a person's day-to-day mood.

Further, it is common to assume that sunlight will make someone happy and rain will make them sad. Songs and sayings include phrases such as "keep on the sunny side of life", "you are my sunshine", and "into every life, some rain must fall" because there is a cultural belief that mood and weather reflect each other (Watson, 2000). However, researchers continue to fail to prove that rain is related to negative moods and sun is related to positive moods (Watson, 2000; Denissen et al., 2008; Huibers et al., 2010). Notwithstanding, these researchers are under the impression that people still believe there is some truth to this, so maybe their beliefs would be enough evidence to use weather

[2] https://flickr.com/

as a criterion when creating datasets with emotion or mood labels. Hence, pictures of landscapes in the different seasons were downloaded into four different albums and were used as test images in the system by Zhu et al. (2017) to translate from photo to painting. The four different seasons were used and paired with the quadrants in two ways.

A script using an API endpoint was used to download up to 1000 pictures from each of the four seasons, summer, spring, winter and autumn, using the keywords *[season] landscape*, e.g., *spring landscape*. The pictures were resized to a 256×256 size to comply with Zhu et al.'s program's criteria. Images from the test sets of each season are shown in Figure 4.2. Even though "landscape" was specified as a keyword when fetching images from Flickr, not all images were of natural landscapes. As shown in Figure 4.3, some images did not match the keyword well, even though the script fetched images with the highest relevance to the keyword.

4.3. Panda et al.'s Dataset with Songs

As presented in Section 3.2, Panda et al. created an open dataset containing metadata about 900 songs, labelling them with quadrants from Russell's circumplex model of affect from Figure 2.1. Appendix B lists all the 209 emotion labels in the dataset (Panda et al., 2018). This dataset can be used as training data because it already has information about songs and their quadrant. All the original columns were the following:

- **Song**: The song name or ID used in AllMusic.
- **Artist**: Name of the artist(s).
- *Title*: Song title.
- **Quadrant**: The annotation, Russell's quadrant obtained.
- *PQuad*: Ratio of mood tags from "Quadrant" against all moods.
- *MoodsTotal*: Total number of moods associated with the song.
- *Moods*: Number of moods that matched the Warriner's list (Warriner et al., 2013).
- *MoodsFoundStr*: Moods that were found.
- **MoodsStr**: Original list of all moods associated with the song entry.
- *MoodsStrSplit*: Same as above but with moods split (for the few tags originally containing two words).
- *Genres*: Number of genres.

(a) Autumn

(b) Spring

(c) Summer

(d) Winter

Figure 4.2.: Four images retrieved from each of the four seasons.

- **GenresStr**: List of genres.
- *Sample*: 1/true, since all songs contain a sample.
- *SampleURL*: The sample URL.

(a) Image retrieved with keyword *summer landscape*

(b) Image retrieved with keyword *spring landscape*

Figure 4.3.: Not all images matched the keyword well.

Alterations

Alterations were made to use this dataset in the system. Some of the columns were considered irrelevant for this work and were removed. These are marked with italics in the list. The used columns are marked with bold: Song, Artist, Quadrant, MoodsStr and GenresStr. In order to use this data for training and testing, further alterations were necessary, as described in Section 2.4. Both MoodsStr and GenresStr are called lists above but were, in fact, strings. These were converted to array lists and then one-hot encoded, giving 209 moods or emotion labels and 21 genres. Next, the Quadrant column was transformed to values between 0-3 instead of Q1, Q2, Q3 and Q4 because XGBoost cannot work with strings. Finally, the artist was also label encoded, which means transformed to numbers instead of strings so that this information could be used. There are many artists in the world, so at first, this seemed like worthless data. However, since much data in TheAudioDB is sparse, keeping the artist information was valuable, as described in the next chapter. If two songs have the same artist, their chances of belonging to the same quadrant are more prominent than if they have nothing in common. After data cleaning, the dataset is ready for training and testing, as described in Chapter 5.

Table 4.3.: The columns on one track from TheAudioDB. The strDescription field had multiple columns for different country codes, e.g., strDescriptionEN.

idTrack	**strArtist**	**strMood**
strGenre	strDescription[country code]	strStyle
idAlbum	idArtist	idLyric
idIMVDB	intCD	strTrack3DCase
strTrackLyrics	strMusicVid	strMusicVidDirector
strMusicVidCompany	strMusicVidScreen1	strMusicVidScreen2
strMusicVidScreen3	intMusicVidViews	intMusicVidLikes
intMusicVidDislikes	intMusicVidFavorites	intMusicVidComments
intTrackNumber	strMusicBrainzID	strMusicBrainzAlbumID
strMusicBrainzArtistID	strLocked	strTrackThumb
strTheme	intLoved	intScore
intScoreVotes	intTotalListeners	intTotalPlays
intDuration	strArtistAlternate	strTrack
strAlbum		

4.4. TheAudioDB

The dataset used to retrieve metadata about test songs was TheAudioDB, a database of audio artwork and metadata with a JSON API[3]. Appendix A shows two examples of JSON responses from the database. Some of the fields in Example 1 have no values. The created system can handle these null values, but it is best if the response has values in the most critical data fields for this Master's Thesis. These fields are marked in bold in Table 4.3. Example 2 in the appendix shows a song that contains very little metadata, as almost all the fields are "null". With no information about the genre and mood of the song, it is challenging to categorise them in the correct quadrant in Russell's circumplex model of affect (Russell, 1980). The system will provide a quadrant, but it does not have enough information to give a reasonable estimate.

Alterations

Alterations were made to the data to fit the purpose of the experiments. Many columns were dropped as they did not provide relevant information about the songs. Table 4.3 below shows all the columns, and the ones that were not dropped are marked in bold. Over 90% of the columns were dropped, indicating that this was not the best database for this project. However, the information retained about the artist, mood and genre can be used alongside the dataset from Panda et al.

[3] https://theaudiodb.com/

Some of the moods from TheAudioDB do not exist in the Panda et al. dataset, for example, "In Love", "Philosophical", or "Troubled". Hence, these will have a lower chance of being categorised into the correct quadrant. The genres are also different from Panda et al.'s set. For instance, two genres in TheAudioDB are Pop and Rock, but these have been merged into one genre as Pop/Rock in Panda et al. Some issues with these inconsistencies are discussed in more detail in Chapter 8.

4.5. The Data That Was Not Chosen

Panda et al. (2018) used AllMusic API with data from Rovi Music[4] to retrieve metadata about songs and 30-second audio clips of the songs. Looking at the documentation, this seemed like an excellent database, providing information about a song genre and what emotion it was labelled with. Rovi Music claims to offer metadata on over 30 million tracks worldwide, including popular hits and cult favourites to minor works and classical masterpieces. The website linked in the footnote is end-of-life, but licensing the metadata is still possible in exchange for a fee. Since there are other free and open music databases, Rovi Music was not chosen for this Master's Thesis.

The Spotify API[5] has endpoints to retrieve information and metadata about songs, artists, and playlists. The information about a song includes duration, acousticness, instrumentalness, and valence. This metadata can be used to create new mappings from song $\rightarrow$ emotion. However, using this endpoint and creating a new mapping is very time-consuming. Considering that the focus of this project is to look at ways to generate art based on the music's emotion, such a technique might be a trial-and-error method that would take too much time. It may also lead to poor results because the mappings may be subjective to their creator. Hence, it was considered better to use datasets that already had a suitable mapping function to extract the emotion or, as in TheAudioDB: already had information about mood.

The International Affective Picture System (IAPS)[6] is a database of photos like WikiArt Emotions. These photos have been labelled with an emotion consistently evoked in the viewers. They have used the valance and arousal scale, such as in Russell's model, in addition to a dominance/control scale to determine the emotions in a picture. The latter scale ranges from "in control" to "dominated", and the viewer can determine how they experience the image in terms of feeling either in control or under control. A weakness with IAPS is that it only contains 700 images. It has also only been tested on so-called WEIRD-people: white, educated, industrialised, rich and democratic people. Studies show that colour has different meanings in different cultures. For instance, white is

[4] http://developer.rovicorp.com/docs
[5] More information on Spotify's API: https://developer.spotify.com/documentation/web-api/.
[6] More information on IAPS: https://imotions.com/blog/iaps-international-affective-picture-system/.

associated with being *clean* and *pure* in western cultures but symbolises *death* in some places in Asia (Saito, 1996). Therefore, the dataset might not give a universal emotion classification, and it has fewer images than WikiArt Emotions, so this was chosen instead of the IAPS dataset.

5. Architecture

This chapter presents some of the technical work and set-up in this project. The created algorithm has three steps, but this chapter also presents a minor set-up step. Selection of test songs was needed as a prerequisite to step one. The procedure for selecting test songs is explained in Section 5.1. Step one was processing and transforming the data that would be used. The data transformation and cleaning are described in more detail in Chapter 4. Next, Section 5.2 describes step two in the algorithm, the training and testing of songs to categorise the chosen testing songs into the correct quadrants. Lastly, Section 5.3 and Section 5.4 present two methods for step three in the algorithm, which is transforming an image into a painting to create a set of fitting output paintings for a given song. The results of all these steps are presented in the next chapter, Chapter 6.

5.1. Selecting the Test Songs

TheAudioDB was selected as the database to retrieve test songs. The variety between the songs was substantial, and in order to provide songs with as few null values as possible, the selection of songs was made manually. Some famous artists were chosen and searched for in TheAudioDB's search engine. When selecting a song, an information page displays most of the available information about that song[1]. This is the same information that is fetched through the API. A song was chosen if it had metadata on the essential fields described in Chapter 4. This search method went on until eleven songs were chosen. These songs are listed in Table 6.1. Selecting songs this way was time-consuming because the database is sparse with metadata. As shown in Figure 5.1[2], gathering detailed metadata about the albums and songs has not been a priority.

5.2. Tuning, Training and Testing Songs

Alterations to the metadata about the songs were necessary before training and testing. The alterations are described in more detail in Section 4.3 and Section 4.4 and are shown

[1]Information page of the song Thriller by Michael Jackson: https://www.theaudiodb.com/track/32822998.

[2]Screenshot in figure retrieved from https://www.theaudiodb.com/stats.php

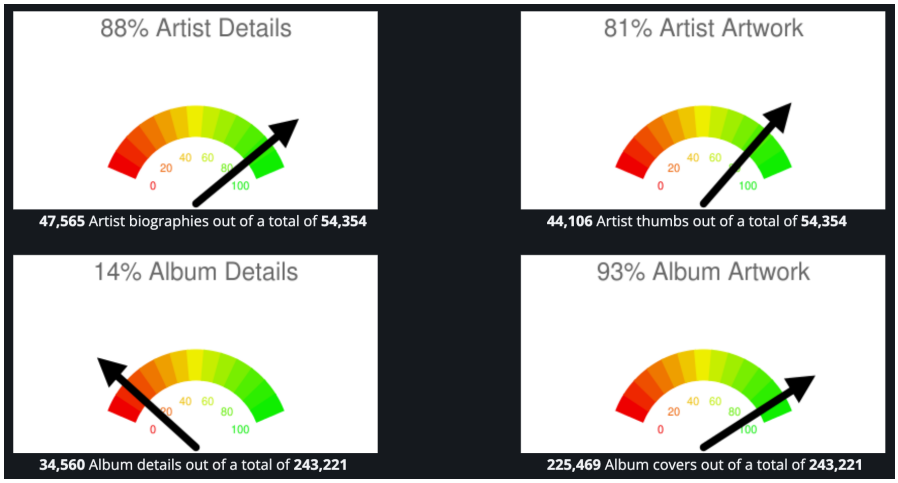


Figure 5.1.: Statistics on metadata from TheAudioDB website.

as step 1 in Figure 5.2. After the data transformation, the processed data was sent into the training algorithm, shown in step 2 in the figure. The metadata in the dataset from Panda et al. (2018) was split into trainX, trainY, testX and testY, where the Y sets contain all the labelled quadrants in the dataset. XGBoost's XCBClassifier was used to find the optimal parameters using manual parameter tuning[3]. The XGBClassifier is a scikit-learn class for classification. GridSearchCV[4], a model selection step, was used to test 5-15 different values for each parameter to find the one that gave the highest score. A code snippet is shown below. This code was run for each parameter, and the best parameters were selected and used for further tuning. All of the values that were tested were sent in with `param_test`, e.g., `param_test = {'max_depth':range(1,12,3)}`. The `tuning_estimators` parameter was the XGBClassifier, which in the beginning had no arguments, i.e., `tuning_estimators = XGBClassifier()`.

```
gsearch = GridSearchCV(
    estimator = tuning_estimators,
    param_grid = param_test,
    n_jobs=4,
    cv=5)
gsearch.fit(trainX, trainY)
gsearch.cv_results_, gsearch.best_params_, gsearch.best_score_
```

The parameters and their tested values are shown in Table 5.1 on page 32. The best ones are marked in bold and are: `{max_depth=4, min_child_weight=0, gamma=0.05,`

[3]Documentation on XGBClassifier: https://xgboost.readthedocs.io/en/stable/python/python_api.html#xgboost.XGBClassifier

[4]Documentation on GridSearchCV: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

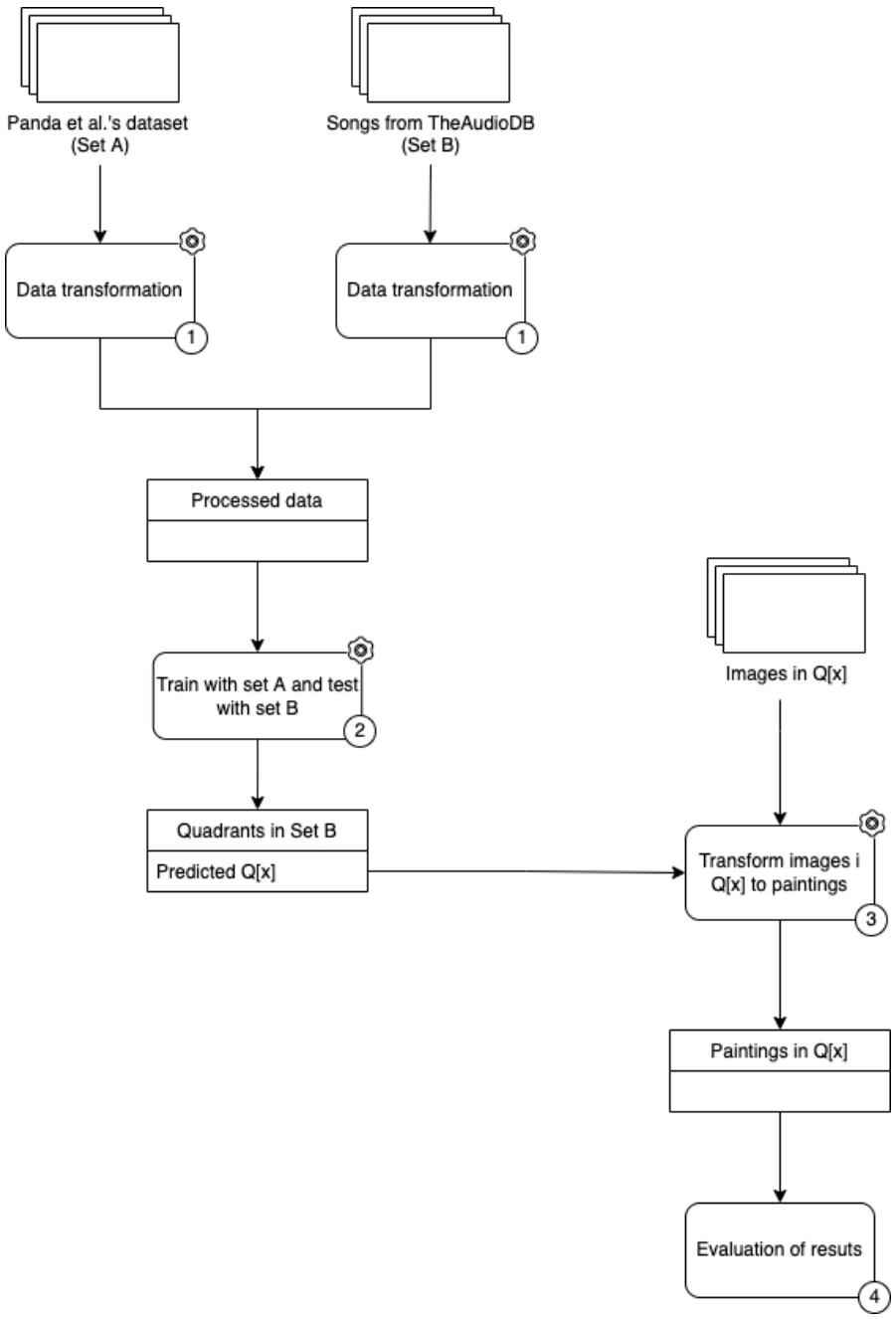


Figure 5.2.: The different steps of the system.

`colsample_bytree=0.4, subsample=0.6}`. A 5-fold cross-validation was used to achieve these values. The parameters `reg_alpha`, `reg_lambda` and `eta`, ended up with their default values, the same as the best value. The model was fit using these parameter values, and trainX and trainY. The model's accuracy was checked using testX and testY,

Table 5.1.: All the values that were tested for each parameter

Parameter	Values tested
max_depth	1, 2, 3, **4**, 5, 6, 7, 10
min_child_weight	**0**, 1, 2, 4, 6
gamma	0.0, 0.001, 0.01, **0.05**, 0.1, 0.2, 0.25, 0.3, 0.35 0.4
colsample_bytree	0.001, 0.01, 0.1, **0.4**, 0.45, 0.5, 0.55, 0.6, 0.8, 1.0
subsample	0.001, 0.01, 0.1, 0.4, 0.45, 0.5, 0.55, **0.6**, 0.8, 1.0
reg_alpha	**0**, 0.001, 0.01, 0.1, 1, 10, 100
reg_lambda	0, 0.001, 0.01, 0.1, 0.5, 0.7, **1**, 2, 5, 10, 100
eta	0, 0.001, 0.01, 0.1, **0.3**, 0.5, 0.8, 1

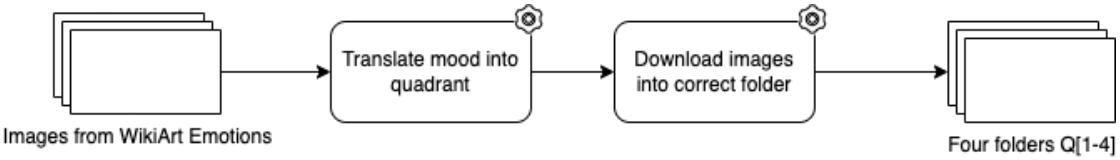


Figure 5.3.: Downloading the images from WikiArt into correct folders.

and the results gave 95.19% accuracy with the given parameters. This model was used to test new songs from TheAudioDB, also in step 2 in the figure. The algorithm's output in step 2 is the predicted quadrant of the song from TheAudioDB.

5.3. Using Image-to-Image Translation on Photographs

Two methods have been used in step 3 from Figure 5.2. The first is presented here, and the second is in Section 5.4. The first method used Zhu et al.'s image-to-image translation from the season photographs presented in Section 4.2. The season photographs were divided into four folders: Q1, Q2, Q3, and Q4. The folder matching the quadrant of the song was sent into Zhu et al.'s algorithm, and the output was all the images transformed into paintings with the same style as Monet. They were in the style of Monet's painting because his paintings were used to train Zhu et al.'s cycle-consistent model. The translation was only performed in one direction when testing with the season photographs: from picture to painting. However, it is possible to transform in both directions simultaneously, as shown in Figure 3.4 from Section 3.3. No alterations have been made to the image-to-image translation algorithm by Zhu et al. (2017). Only the dataset and the options for training and testing were provided. The options used to test and train with the season photographs are presented and discussed in Section 6.2.

5.4. Using the WikiArt Emotions Dataset

Using the WikiArt Emotions images was an alternative to step 3 from Figure 5.2. The first step of this method was to download all the images and categorise them into correct quadrants. The categorisation is presented in more detail in Section 4.1, and the flow of this algorithm is shown in Figure 5.3. After downloading all the images into the correct folder, different methods can be used to select one painting as output for the input song. These methods are presented and discussed in Section 6.2.

6. Experiments and Results

This chapter describes the experimental plan, set-up and results. In Section 6.1, four steps to the experiment are presented. Next, Section 6.2 presents the necessary set-up to complete the four steps from Section 6.1. Lastly, Section 6.3 presents the results when going from song to image.

6.1. Experimental Plan

The experiments have been completed stepwise, all shown in the list below. The first step was to categorise songs into Russell's quadrants. The technique will work for any song that has the same metadata. The set-up for this step is explained in more detail in Subsection 6.2.1.

1. Categorise songs into Russell's quadrants
2. Create and prepare datasets with pictures
3. Transform from song to image
4. Evaluate the system

The second step was to create datasets with pictures labelled with a quadrant. The system will only use images in the same quadrant as the input song when going from song to image. Two different image datasets have been used. The first was from WikiArt Emotions, described in Chapter 4. These images are labelled with emotions, so Subsection 6.2.2 describes the process of categorising them into quadrants. The second dataset was created from scratch using Flickr's API to fetch images based on keywords. The categorisation of these images is also explained in Subsection 6.2.2.

The next step was to transform from song to image. This step uses the results from the former two, and the set-up is presented in Subsection 6.2.3. The results from this step are the central part of this project and are explained in detail in Section 6.3. Moreover, this step was also the most time-consuming. The results are also the central part of the survey used to evaluate the system, presented in Chapter 7. In Subsection 6.2.4, the

set-up of the two evaluation methods is presented, and the final evaluation of the system is discussed in more detail in Section 8.1.

6.2. Experimental Set-up

This section describes the experimental set-up, all the parameters used to achieve the results and all the work done to prepare for the experiments.

6.2.1. Step 1: Choosing and Categorising Songs

TheAudioDB was used to gather metadata about the test songs used in step 3 when going from song to image. These songs were categorised into quadrants. As explained in Chapter 4, many songs had incomplete information where many fields had the value `Null`. It was important to choose songs with information about their genre and mood to ensure valid results. The process of selecting songs is described in Section 5.1. Since many of the songs in TheAudioDB do not have sufficient information to get an accurate classification, the selection was made manually, searching for popular songs with the search engine until eleven songs were collected. It is essential to mention that the program will work with any given song ID from TheAudioDB. The results, however, will not necessarily be satisfying if the metadata is incomplete.

In order to use the dataset from Panda et al. (2018) as training data, data cleaning and transformation have been completed. The data cleaning and transformation are described in more detail in Section 4.3. The data columns needed to be the same for Panda et al.'s dataset and TheAudioDB, and they needed to be transformed from tabular data to a numeric matrix. For Panda et al.'s dataset, this meant going from 14 to 235 columns due to the necessary one-hot encoding of the moods in `MoodsStr` and the genres in `GenresStr`. The transformation is depicted in Figure 6.1. Figure 6.1a shows the dataset before transformation, whereas Figure 6.1c shows the values closer. Figure 6.1b shows the dataset after transformation, where Figure 6.1d is zoomed in. As figures *c* and *d* show, the values have gone from a mixture of strings and numbers to become strictly numeric.

After structuring Panda et al.'s dataset to suit the system, the metadata about the testing song from TheAudioDB was retrieved. The structure of the JSON response is shown in Appendix A. This data also needed some cleaning and transformation to be in the same format as the training set. The process of this transformation is presented in Section 4.4. After cleaning, the system trained with XGBoost using Panda et al.'s data to learn what genres and moods lead to which quadrants and then used this training model on the testing song. The training and testing process is described in more detail

	Song	Artist	Title	Quadrant	PQuad	MoodsTotal	Moods	MoodsFoundStr	MoodsStr	MoodsStrSplit	Genres	GenresStr	Sample	SampleURL
0	MT0000004637	Charlie Poole	Bulldog Down in Sunny Tennessee	Q3	0.666667	3	3	circular, greasy, messy	Circular, Greasy, Messy	Circular, Greasy, Messy	2	Country, International	1	http://rovimusic.rovicorp.com/playback.mp3?c=1...
1	MT0000011357	Dismember	Reborn in Blasphemy	Q2	0.666667	3	3	jittery, negative, nervous	Negative, Nervous/Jittery	Negative, Nervous, Jittery	3	Electronic, International, Pop/Rock	1	http://rovimusic.rovicorp.com/playback.mp3?c=0...
2	MT0000011975	Curse of the Golden Vampire	Ultrasonic Meltdown	Q2	0.666667	6	5	fierce, harsh, hostile, menacing, outrageous	Fierce, Harsh, Hostile, Menacing, Outrageous, ...	Fierce, Harsh, Hostile, Menacing, Outrageous, ...	1	Electronic	1	http://rovimusic.rovicorp.com/playback.mp3?c=_...
3	MT0000040632	Gipsy Kings	Flamencos en el Aire	Q1	0.750000	4	3	fiery, sexy, spicy	Cathartic, Fiery, Sexy, Spicy	Cathartic, Fiery, Sexy, Spicy	2	International, Jazz	1	http://rovimusic.rovicorp.com/playback.mp3?c=G...
4	MT0000044741	Little Walter	Last Night	Q3	0.750000	4	4	greasy, gritty, gutsy, lazy	Greasy, Gritty, Gutsy, Lazy	Greasy, Gritty, Gutsy, Lazy	1	Blues	1	http://rovimusic.rovicorp.com/playback.mp3?c=k...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
895	MT0035332835	Disincarnate	In Sufferance	Q2	0.571429	7	4	angry, harsh, hostile, reckless	Angry, Harsh, Hedonistic, Hostile, Malevolent,...	Angry, Harsh, Hedonistic, Hostile, Malevolent,...	1	Pop/Rock	1	http://rovimusic.rovicorp.com/playback.mp3?c=b...
896	MT0035334027	Lieutenant Stitchie	Ego Tripping	Q2	0.666667	3	3	harsh, reckless, snide	Harsh, Reckless, Snide	Harsh, Reckless, Snide	1	Reggae	1	http://rovimusic.rovicorp.com/playback.mp3?c=e...
897	MT0036111736	Subhumans	Peroxide	Q2	0.666667	3	3	bitter, harsh, outraged	Bitter, Harsh, Outraged	Bitter, Harsh, Outraged	1	Pop/Rock	1	http://rovimusic.rovicorp.com/playback.mp3?c=R...
898	MT0036368550	Taio Cruz	Dynamite	Q1	0.555556	9	9	bright, carefree, energetic, euphoric, excitin...	Bright, Carefree, Energetic, Euphoric, Excitin...	Bright, Carefree, Energetic, Euphoric, Excitin...	1	Pop/Rock	1	http://rovimusic.rovicorp.com/playback.mp3?c=_...
899	MT0040033011	Product	Read	Q3	0.600000	5	3	bitter, bleak, snide	Bitter, Bleak, Snide, Somber, Thuggish	Bitter, Bleak, Snide, Somber, Thuggish	1	Rap	1	http://rovimusic.rovicorp.com/playback.mp3?c=E...

900 rows × 14 columns

(a) Panda et al.'s original dataset columns

	Song	Artist	Quadrant	MoodsTotal	Genres	Avant-Garde	Blues	Children's	Classical	Comedy/Spoken	...	Visceral	Volatile	Warm	Weary	Whimsical	Wintry	Wistful	Witty	Wry	Yearning
0	MT0000004637	103	2	3	2	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	MT0000011357	175	1	3	3	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	MT0000011975	134	1	6	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	MT0000040632	246	0	4	2	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	MT0000044741	358	2	4	1	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
895	MT0035332835	174	1	7	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
896	MT0035334027	350	1	3	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
897	MT0036111736	553	1	3	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
898	MT0036368550	558	0	9	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
899	MT0040033011	462	2	5	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

900 rows × 235 columns

(b) Panda et al.'s dataset after transformation

MoodsStrSplit	Genres	GenresStr
Circular, Greasy, Messy	2	Country, International
Negative, Nervous, Jittery	3	Electronic, International, Pop/Rock
Fierce, Harsh, Hostile, Menacing, Outrageous, ...	1	Electronic
Cathartic, Fiery, Sexy, Spicy	2	International, Jazz
Greasy, Gritty, Gutsy, Lazy	1	Blues

(c) Zoomed in on Panda et al.'s original dataset columns

Blues	Children's	Classical
0	0	0
0	0	0
0	0	0
0	0	0
1	0	0

(d) Zoomed in on Panda et al.'s dataset after transformation

Figure 6.1.: Transformation of Panda et al.'s dataset.

in Section 5.2.

Table 6.1 shows the results from categorising the songs based on Panda et al.'s information. The column "Mood" reveals the mood that TheAudioDB has assigned each song. As seen in the table, no song was associated with quadrant Q3. As a result of this, many different songs were investigated using TheAudioDB's search engine to find one that the system categorised as Q3, but no such song was found. This result is discussed in Section 8.2. Five of the songs from Table 6.1 were selected to test the program further: two from Q1, two from Q4 and the single one from Q2. The chosen songs are marked in

Table 6.1.: Results after categorising songs. The lines marked in green were selected for further testing. The white rows are shown in the table to prove that no songs were categorised in Q3 – most songs were placed in Q1 and Q4.

Song ID	Song Title	Artist	Mood	Quadrant
32822998	Thriller	Michael Jackson	Quirky	Q4
32802707	Bohemian Rhapsody	Queen	Rousing	Q1
32767148	The A Team	Ed Sheeran	Philosophical	Q4
35027980	Scorpion	Drake	Passionate	Q4
32978788	Dangerously in Love	Beyoncé	Energetic	Q1
35787816	Easy on Me	Adele	In Love	Q4
33033015	Hurt	Johnny Cash	Provacative	Q1
32734677	The Way I Am	Eminem	Angry	Q2
32724218	Fix You	Coldplay	Relaxed	Q4
33155473	Imagine	John Lennon	Dreamy	Q4
32801239	Can't Slow Down	Lionel Richie	Cheerful	Q1
32769006	Rehab	Amy Winehouse	Troubled	Q4

light green on the table.

6.2.2. Step 2: Categorising and Filtering Pictures

The second step of the experiment was to create testing sets of different images. Two of the testing sets used images from Flickr, and another two used images from WikiArt Emotions. This subsection describes how the four test sets were created and how the photographs from Flickr were transformed into paintings using Zhu et al.'s algorithm.

Creating the picture datasets

The two testing sets from Flickr both contained the same images, but they were categorised differently. The process of downloading the images is described in Section 4.2. After the images were downloaded, they were separated into different folders representing their matching quadrant. Test sets A and B are shown in Table 6.2. Essentially the autumn and winter seasons swap between Q2 and Q3, whilst summer and spring swap between Q1 and Q4.

The reasoning for this distribution is the research by Watson (2000) and Huibers et al. (2010). They investigate the correlation between mood, weather and seasons. Even though neither study proved a strong correlation between, e.g., bad weather and sadness, this was the foundation for the test set categorisation. Seasonal Affective Disorder (SAD) affects a large group of the world's population, especially those that live far away from the

Table 6.2.: Distribution of season images in test sets A and B.

Season	Test set A	Test set B
Autumn	Q2	Q3
Winter	Q3	Q2
Summer	Q1	Q4
Spring	Q4	Q1

Table 6.3.: The final four test sets A, B, C and D. Q stands for quadrant.

Q	A - Flickr	B - Flickr	C - WikiArt Emotions	D - WikiArt Emotions
Q1	Summer	Spring	All images in Q1	Landscape paintings in Q1
Q2	Autumn	Winter	All images in Q2	Landscape paintings in Q2
Q3	Winter	Autumn	All images in Q3	Landscape paintings in Q3
Q4	Spring	Summer	All images in Q4	Landscape paintings in Q4

equator (Melrose, 2015). SAD is often called *the winter blues*, which led to the decision to categorise winter and autumn with Q3 and Q2 and consequently labelling summer and spring with Q1 and Q4.

After creating test sets A and B, the next step was to go through the pictures from Mohammad and Kiritchenko (2018) in the WikiArt Emotions dataset. As described in Chapter 4, all the images are labelled with one or more emotions in three different categories: image only, image and title, and title only. The labels from the image only category were used to place the images into different quadrants. The distribution of images in each quadrant is shown in Section 4.1.

Two testing folders were created, one containing all the images from WikiArt Emotions that have an emotion label (test set C) and one containing only landscape motives (test set D). In order to filter out the paintings with landscape motives, all images in the dataset labelled with the style *modern* were excluded. All images marked to contain one or more faces were also removed. Finally, the images were iterated over manually to remove any of the remaining paintings that were not of landscapes.

Table 6.3 shows the final four test sets used when transforming from song to image.

Transforming the Flickr photos into paintings

The image-image translation from Zhu et al. (2017) was used to transform the seasonal landscape photographs to look like Monet paintings. The seasonal folders were used as testing sets for a one-way image translation with Zhu et al.'s algorithm. This process is described in more detail in Section 5.3. The algorithm was trained using one of Zhu

et al.'s datasets, and then the seasonal folders were used for testing. The results are a new folder with all the images transformed to imitate Monet's paintings.

Training and testing using Zhu et al.'s system were completed using NTNU's High-Performance Computing Group's Idun cluster system. *Idun* is a project that aims at providing a high-availability and professionally administrated compute platform for NTNU[1]. It allowed for the use of GPUs and was practical in the training phase of this experiment step. All the code run in this step was submitted to Idun in the form of scripts.

Training using Zhu et al.'s CycleGAN model was completed with the command below. `train.py` is the original Python code that runs the training algorithm. The other values are options sent in through the command line. The options and the reasoning for their selected values are explained below the code. It is also possible to use pretrained models directly. Zhu et al.'s monet2photo dataset has pretrained models that can be downloaded, enabling translation from Monet's painting to picture. Therefore, running the training algorithms from scratch is unnecessary if there is a time or resource constraint, and translation is only necessary for this direction.

```
python3 train.py --dataroot ./datasets/monet2photo/ --name
    photo2art --model cycle_gan --display_id 0 --n_epochs 50 --
    n_epochs_decay 50
```

- `-dataroot`: Path to images. The dataset `monet2photo` from Zhu et al. was used, which was stored in the `datasets` folder.
- `-name`: The name of the experiments – decides where to store samples and models. This experiment was called photo2art, but it may be anything.
- `-model`: Which model to use. This experiment used the CycleGAN model because this was recommended by the system's creators if testing image-to-image translation.
- `-display_id`: Window ID of the web display. Set to 0 in this experiment as there was no need to view the training results in real-time. The testing algorithm was used for over 12 hours when running on GPUs. Therefore looking at the results for 12 hours did not make much sense.
- `-n_epochs`: Number of epochs with the initial learning rate. Default is 100, but it was set to 50 due to time constraints. Using 100 epochs would result in twice the time and use more resources. However, the results would most likely have been better, which is discussed further in Chapter 8.
- `-n_epochs_decay`: Number of epochs to linearly decay learning rate to zero. Default is 100, but it was set to 50. The reasoning is the same as above.

[1]Read more about Idun here: `https://www.hpc.ntnu.no/idun/`

When the training was completed, it was time to test the model on the images collected from Flickr. The code below shows the line in the script that completed the testing of the datasets. `test.py` is the original Python code that runs the testing algorithm. The list under the code explains the options from the code line. The test script was run four times, once for each season. If all the season photographs were stored in the same folder, the test script would only need one submission.

```
python3 test.py --dataroot ./datasets/landscapes/${season} --
    name photo2art --model test --no_dropout --results_dir
    results
```

- `-dataroot`: Path to images. The landscape datasets were saved in the subfolder `./datasets/landscapes`. The final value of this option, `${season}` indicates that the images from the four seasons were stored in different folders, so this code had to be run four times, once for each season.
- `-name`: The name of the experiments – decides where to store samples and models. The name of the testing was the same as the training: photo2art, because it is still the same experiment.
- `-model`: Which model to use. Zhu et al. created a separate model for testing: `test` that would test one side only. As the goal of this experiment was to translate the season photographs into paintings, only one way was necessary.
- `-no_dropout`: The generator cannot have any dropout.
- `-results_dir`: The directory where the results will be saved. It can be anything.

Two examples of the translation are shown in Figure 6.2. Figure 6.2a and Figure 6.2c show the original photographs of a spring and an autumn landscape, and Figure 6.2b and Figure 6.2d show the result after they have been transformed into paintings.

6.2.3. Step 3: Transform from Song to Image

Step 3 used the results from steps 1 and 2 to transform a song into an image. The first step in this transformation was to know what quadrant the song would fit in. A song ID from TheAudioDB was passed as input, and the categorisation algorithm from Subsection 6.2.1 provided the output, which was the given quadrant for that song. Next, the quadrant was passed as input to the image picker. The system chose a random image from the folder of the correct quadrant. This algorithm was run four times, choosing one image from all four test sets in Subsection 6.2.2.

Figure 6.3 shows the different steps in this process. The song ID “32822998”, the ID for

(a) A real photograph of a spring landscape.

(b) Spring photo in painting style.

(c) A real photograph of an autumn landscape.

(d) Autumn photo in painting style.

Figure 6.2.: Two images retrieved from Flickr that have been translated to look like a Monet paintings.

Thriller by Michael Jackson, is sent as input. The system uses the song's information and Panda et al.'s dataset to place the song into a quadrant. This categorisation is described in more detail in Section 5.2. Next, the quadrant "Q4" is used as input in the next step, which sets the correct folder to select images from. The final step is to select a random image from the folder and return the image name as output. The results of this step are presented in Section 6.3.

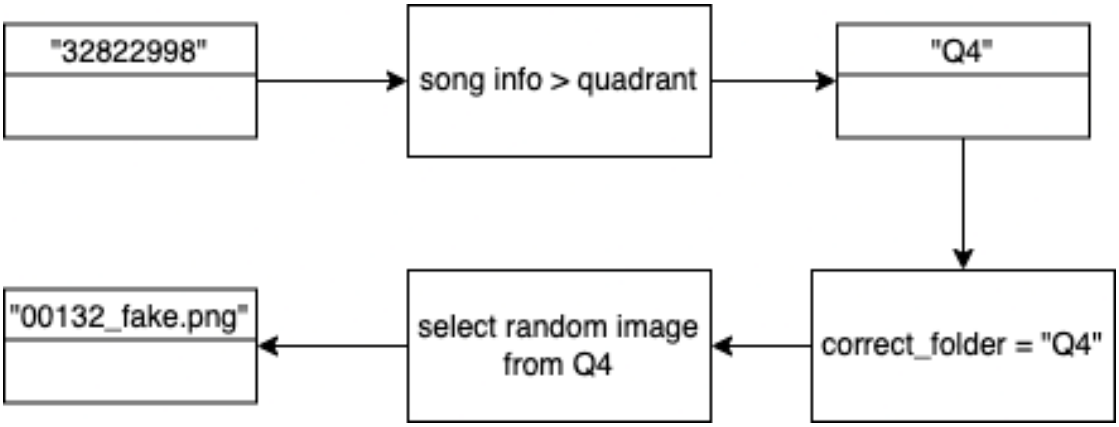


Figure 6.3.: A diagram showing how the system works.

6.2.4. Step 4: Evaluate the System

The evaluation was conducted in two different ways. The first was through a digital survey that participants answered using their digital platforms. A minimum of 100 participants was the wanted amount for the evaluation to get as general an opinion as possible. The participants were presented with all the song-image-pairs, and they had to answer questions related to the pairs. All the questions and their answers are presented and discussed in Chapter 7.

The second evaluation method was to interview five people. The interviews would consist of the participants answering the same digital survey but at the same time thinking out loud on every question. The interviewer wrote down all their thoughts. The goal of the interviews was to get a broader understanding of why participants answered the way they did. The participants were asked to alliterate every thought on every question.

In both evaluation methods, the participants listened to a minimum of 50 seconds of a song. Then they were asked to place it into one of Russell's quadrants. Next, they were presented with the selected result images and were asked to select the image they thought best matched that song. Finally, two follow-up questions related to the song-image pairs were asked.

6.3. Results of Song to Image

This section presents the final results for all the chosen songs from Section 5.1. The experiment's goal was for the system to select four images for each song that should evoke the same emotion in the observer. The emotional response should also be more potent when listening to the music whilst looking at the images. The first two images on

each song, images A and B, were photographs retrieved from Flickr using the keywords *autumn landscape*, *spring landscape*, *summer landscape*, and *winter landscape*. These were transformed into paintings using the image-to-image translation from Zhu et al. (2017) and categorised into quadrants as described in Subsection 6.2.2.

The other two images connected to each song, C and D, were selected from the WikiArt Emotions dataset from Mohammad and Kiritchenko (2018). These were already labelled with emotions used when categorising them into Russell's quadrants. The use of these emotion labels in categorisation is described in Section 4.1. In each subsection of every song below, the labelled emotion for each image is presented. Some images only have one emotion label, while others have up to three emotion labels.

The four images connected to each song were selected at random from the correct quadrant folder in the test sets. Therefore, it is not possible to recreate the resulting song-image-pairs precisely as it is improbable that the system will randomly select the same photos from all the sets. However, the system may still be used on the same or new test songs and provide images that the system has paired with the song.

The system has successfully categorised each song into a quadrant and chosen four images that should match the perceived emotion in the song. Whether or not the images were a good match to each of the songs is discussed in Section 7.3. Nevertheless, looking at the results, speculation is that the survey participants will not agree with the system, so possible improvements are discussed in Section 8.2.

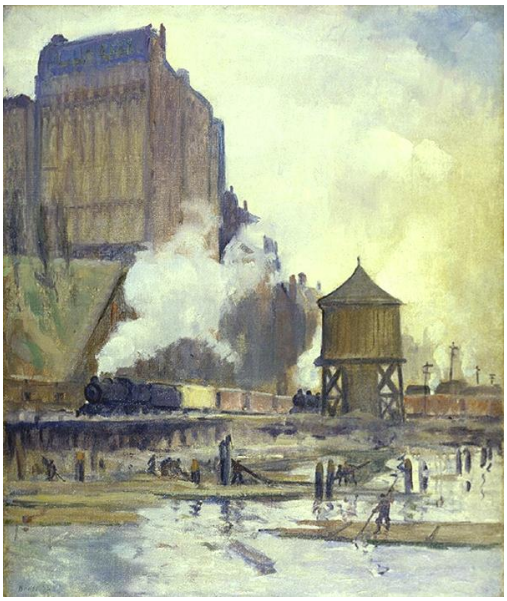
6.3.1. Thriller by Michael Jackson

Thriller, written by Rod Temperton for Michael Jackson, has the song ID 32822998 in TheAudioDB[2]. It is a mix of post-disco and funk, but TheAudioDB has labelled it with the mood "Quirky" and the genre "Pop". The song has horror film sound effects such as footsteps, thunder, and wind[3]. The system placed this song into quadrant Q4 – relaxed – which does not match the horror film atmosphere. Nevertheless, images in the four test sets from Q4 were selected. All the resulting images are shown in Figure 6.4. Image A is a spring landscape photograph translated into a painting, and image B is a summer landscape photograph transformed into a painting. Images C and D are retrieved from WikiArt Emotions. Image C was initially labelled with the emotion "Humility" in Mohammad and Kiritchenko's dataset, and image D was labelled with three emotions: "Happiness", "Humility", and "Optimism".

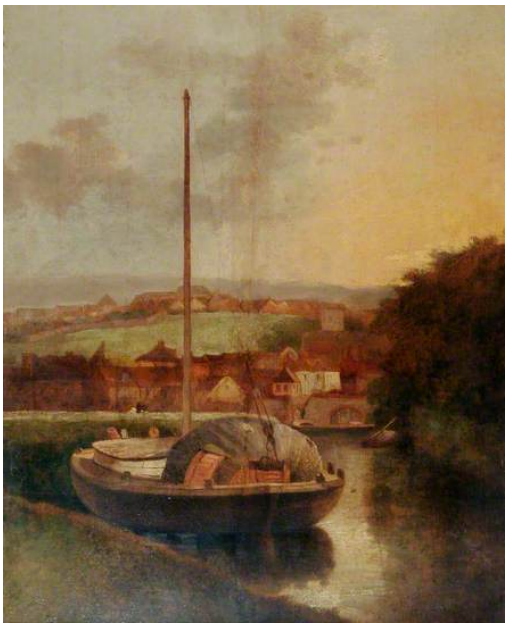
[2]Information page on Thriller: `https://www.theaudiodb.com/track/32822998`
[3]Wikipedia page for Thriller: `https://en.wikipedia.org/wiki/Thriller_(song)`

(a) Landscape image A

(b) Landscape image B



(c) WikiArt Emotions image C



(d) WikiArt Emotions image D

Figure 6.4.: The four images that were matched with Thriller.

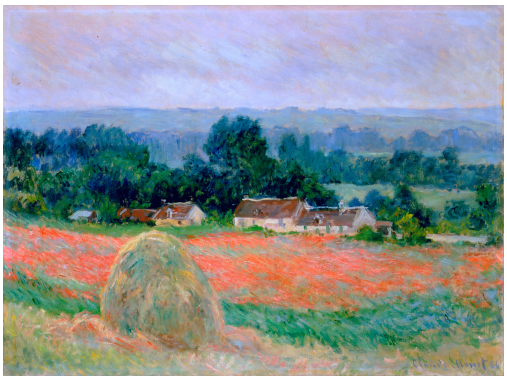
6.3.2. Dangerously in Love by Beyoncé

Dangerously in Love, written and produced by Beyoncé Knowles and Errol McCalla Jr., has the song ID 32978788 in TheAudioDB[4]. The song is a ballad in the genres R&B and

[4]Information page on Dangerously in Love: https://www.theaudiodb.com/track/32978788

(a) Landscape image A

(b) Landscape image B



(c) WikiArt Emotions image C

(d) WikiArt Emotions image D

Figure 6.5.: The four images that were matched with Dangerously in Love.

Soul[5] but is labelled with the mood "Energetic" and the genre "Funk" in TheAudioDB. The system placed the song into quadrant Q1 – happy, and all the image results are presented in Figure 6.5. Images A and B are the landscape photographs of summer and spring, respectively, transformed into paintings. Images C and D are from WikiArt Emotions, where both the former and the latter are labelled with "Happiness" in the dataset.

6.3.3. Hurt by Johnny Cash

Hurt was originally a rock song written by Trent Reznor for the band Nine Inch Nails, but a cover by Johnny Cash became famous in the early 2000s as an alternative acoustic

[5]Wikipedia page for Dangerously in Love: https://en.wikipedia.org/wiki/Dangerously_in_Love_2

(a) Landscape image A

(b) Landscape image B

(c) WikiArt Emotions image C

(d) WikiArt Emotions image D

Figure 6.6.: The four images that were matched with Hurt.

rock/country song[6]. The song has the ID 33033015 and is labelled with the mood "Provocative" and the genre "Country" in TheAudioDB[7]. The mood "Provocative" might relate to the lyrics in the song, covering topics such as self-harm and drug abuse. Q1 – happy – was the quadrant that the system assigned to this song, which does not fit the lyrical content. The image results can be seen in Figure 6.6. Images A and B are from

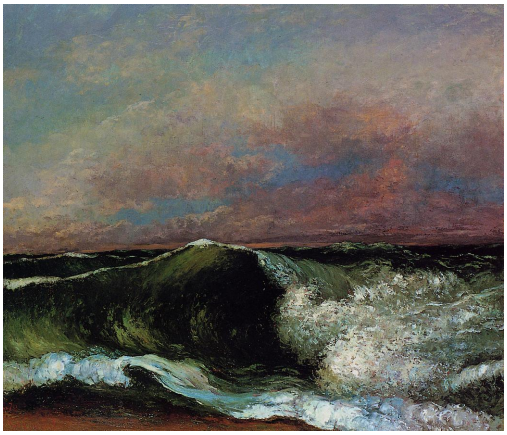
[6]Wikipedia page for Hurt: https://en.wikipedia.org/wiki/Hurt_(Nine_Inch_Nails_song)#Johnny_Cash_version

[7]Information page on Hurt: https://www.theaudiodb.com/track/33033015

(a) Landscape image A

(b) Landscape image B

(c) WikiArt Emotions image C



(d) WikiArt Emotions image D

Figure 6.7.: The four images that were matched with The Way I Am.

the same test sets as in the last song: A comes from the summer landscape set, and B comes from the spring landscape set. "Surprise" is the emotion label that Mohammad and Kiritchenko created for image C, and D has been labelled with "Happiness".

6.3.4. The Way I Am by Eminem

The Way I Am, a hip-hop song by the rapper Eminem, has the song ID 32734677 in TheAudioDB[8]. The song's theme is the rapper's frustration with being pressured by fans

[8]Information page on The Way I Am: https://www.theaudiodb.com/track/32734677

and critics to be or act a certain way[9]. The song is labelled with the mood "Angry" and the genre "Hip-Hop" in the database. This song was placed in Q2 – angry – by the system and was matched with the images in Figure 6.7. Image A is an image from the set of autumn landscapes, and image B comes from the winter landscape set. Images C and D are from the WikiArt Emotions set where image C represents "Anger" and image D had been labelled with both "Fear" and "Happiness".

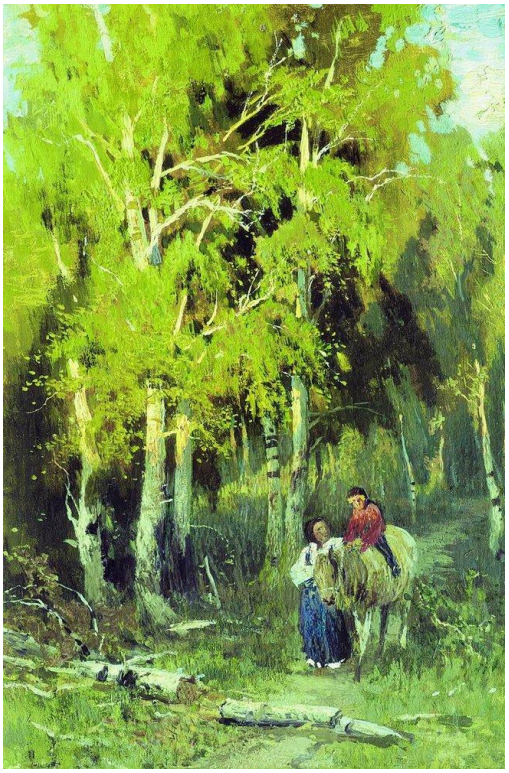
6.3.5. Rehab by Amy Winehouse

Rehab, written and recorded by Amy Winehouse, has the song ID 32769006 and is labelled with the mood "Troubled" and the genre "Soul" in TheAudioDB[10]. The lyrics address Winehouse's refusal to enter a rehabilitation clinic[11], making "Troubled" a suitable mood for the song. This was the last test song, and the system categorised it into the Q4 quadrant – relaxed. Figure 6.8 (page 50) shows the images that were paired with this song. Image A was retrieved from the spring landscape set, while image B was selected from the summer landscape set. The images in C and D come from the WikiArt Emotions dataset, where the former image has been labelled with "Happiness", "Humility", and "Love", and the latter, image D, has been labelled with only "Humility".

[9] Wikipedia page for The Way I Am: https://en.wikipedia.org/wiki/The_Way_I_Am_(Eminem_song)
[10] Information page on Rehab: https://www.theaudiodb.com/track/32769006
[11] Wikipedia page for Rehab: https://en.wikipedia.org/wiki/Rehab_(Amy_Winehouse_song)

(a) Landscape image A

(b) Landscape image B



(c) WikiArt Emotions image C

(d) WikiArt Emotions image D

Figure 6.8.: The four images that were matched with Rehab.

7. User Survey

A subjective evaluation was completed after achieving the resulting song-image-pairs presented in Section 6.3. A total of 102 people responded to the questionnaire, in which five of these were interviewed while they were answering. This chapter presents the set-up decisions, questions, the participant representation and the results. The questions and answers can also be found in Appendix C and Appendix D. In Section 7.1, the reasoning for including the different questions is presented. Some of the decisions were based on research, and others on the creator's preferences and theories. This section also exhibits how the survey was distributed to receive enough answers. The participant representation is presented in Section 7.2. This section also provides some discussion around the representation. Finally, Section 7.3 elaborates on the survey results and presents some interesting findings. The results are discussed in more detail in Section 8.1.

7.1. Set-up and Questions

The survey was created using Google Forms[1], and 25 questions were presented to the respondent. The first five were asking for some information about the person: their gender, age, cultural background, and their knowledge of music and art. These questions were vital to getting an idea of who the participants were. Information about their cultural background was interesting due to the previously mentioned research on how colours may have different meanings in different cultures (Saito, 1996).

A good distribution of gender, age, cultural background and musical knowledge was desirable. The optimal representation would be an even dispersal of people in all categories. The alternatives in all the categories are listed below, along with some hypotheses on what the answers may indicate.

- Gender: made sure to create an inclusive survey by adding the alternatives *non-binary* and *other*
  - Female
  - Male

[1]More information at their website: https://www.google.com/forms/about/.

  - Non-binary
  - Other
  - Prefer not to say: participants should be able to choose not to answer
- Age: four different age groups that indicate what person they might be
  - Below 18: Probably teenagers in high school
  - 18 - 25: Probably students of universities
  - 26 - 39: Probably a recent graduate in a job
  - Over 40: Probably an adult in a job
  - Prefer not to say: participants should be able to choose not to answers
- Cultural background: only included five options, which are the seven continents, excluding Antarctica and merging North- and South-America
  - African
  - American
  - Asian
  - European
  - Oceanian
  - Prefer not to say: participants should be able to choose not to answers
- Knowledge about music and art: two questions on a scale of 1 to 7. The layout of these questions is shown in Figure 7.1.
  - 1 (music): I barely know what music is and never listen to it
  - 7 (music): I have a degree in/am currently studying music
  - 1 (art): I never go to museums and do not enjoy art
  - 7 (art): I have a degree in/am currently studying art

The following section asked five sets of four questions, one set for each of the five songs from Section 6.3. Creating the questions for the songs was challenging. It was important not to ask too many questions, as this might have reduced the number of participants

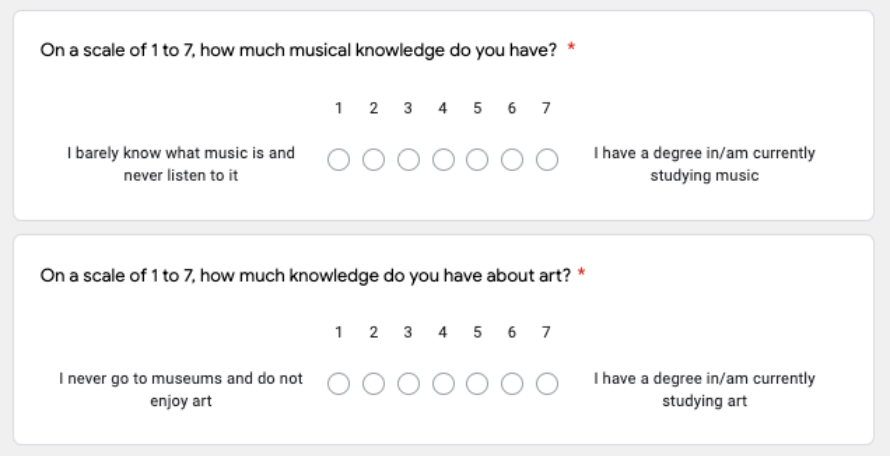


Figure 7.1.: Layout of the questions asking about the participant's knowledge on art and music.

finishing the survey. Furthermore, it was essential to ask precise questions, so there would be no room for interpretations, leading to a confusing span of answers. None of the questions had a text box answering alternative to ensure clear answers. They were either radio buttons, checkboxes or a scale from 1 to 7. The range 1 to 7 was consciously chosen instead of 1 to 5 or 1 to 10. Using a 1 to 5-scale might result in most respondents picking the median value of 3 for most questions where they were unsure. The value 2 might seem too low, and 4 might seem too high. Therefore, using the 1 to 7-scale, the median value 4 might not be used as much because the values 3 and 5 might not seem too generous as 2 and 4 in the former scale option. Lastly, using a 1 to 10-scale might overwhelm the respondents with too many options. Therefore, a scale ranging from 1 to 7 was chosen for all scaling questions.

The participants would need to listen to the songs before answering the questions. A minimum of 50 seconds was deemed necessary. Panda et al. (2018) used music clips of 30 seconds to analyse the songs that they categorised into quadrants. However, some of the songs chosen for testing in this project use more than 30 seconds to reach the chorus (e.g., Thriller by Michael Jackson), while others began with the chorus (e.g., Rehab by Amy Winehouse). To ensure that the respondents got a decent impression of each song, they were asked to listen to at least 50 seconds. Next, the first question asked which quadrant they would place the song in, given the model in Figure 3.1. The second question presented the four resulting images for the song and asked which image they thought matched the mood in the song best. It was up to the participants to determine what "best" meant in this question. If the question had a different formulation, there would be less room for individual interpretation. However, the third question asked the participants to rate how well the song-image-pair they chose was using a scale of 1 to 7. Finally, the participants were asked to select from eleven checkboxes if they had any preferences or ideas on what sort of image would match the song. All the questions about

all the songs are depicted in Appendix C.

Since the survey has a population size above ten million people (any person in any country), the margin of error was set to ±10. According to different web articles, a sample size of 100 was necessary to ensure statistically valid results[2]. To receive a minimum of 100 responses, the survey was shared on the creator's Facebook page, sent to friends and family, and posted on two subreddits on Reddit[3]. Sharing it with friends, family and fellow Master's students could be a way to ensure a large number of answers. This was theorised by the author of this report based on personal experience. However, since the author lives in Europe and most of their friends and family are also European, the participant representation might not include people of different cultural backgrounds. Therefore, the decision to post it on Reddit, an international community network, could help collect answers from people with different cultural backgrounds.

7.2. Participant Representation

The wanted number of responses (100) was achieved with 102 people that answered the survey. This section presents the answers to the first part of the survey, which collected some data on the participants. The first question asked about gender information. As seen in Figure 7.2, it was a good spread of genders at almost fifty-fifty between males and females. No answers were given to "Prefer not to say" or "Non-binary". The age groups did not have quite as good dispersal since over 57% were in the category "18-25". Most of these are probably students, as the survey was distributed to the author's friends, also within this age span. The spread in the cultural background is also poor, as over 90% of the respondents are European.

Next, the participants were asked how much knowledge they had of music and art. The scale went from barely knowing what music/art is to having or currently achieving a degree in music/art. Some of the interview subjects mentioned that they thought it was difficult to place themselves on this scale because the span was so vast. Checkboxes or a floating scale could have been better alternatives to pinpoint their knowledge. If checkboxes were used, values such as "I listen to music every day", "I play a musical instrument", or "I paint as a hobby" might appeal to more people.

Figure 7.3 shows the distribution of answers on the participants' knowledge. Only two of them have a musical degree, while none have a degree in art. None of the participants answered the value 1 in any of the categories, but over 60% categorised themselves as a level of 2 and 3 on their knowledge about art.

[2]Web articles: https://www.cloudresearch.com/resources/guides/statistical-significance/determine-sample-size/ and https://www.surveymonkey.com/curiosity/how-many-people-do-i-need-to-take-my-survey/.
[3]More information on Reddit can be found on their homepage: https://www.redditinc.com/

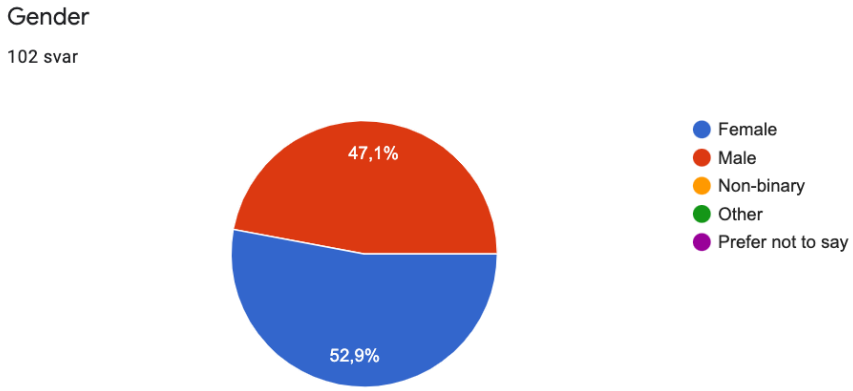
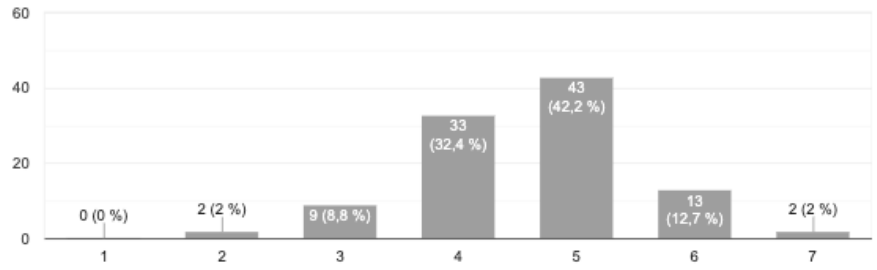
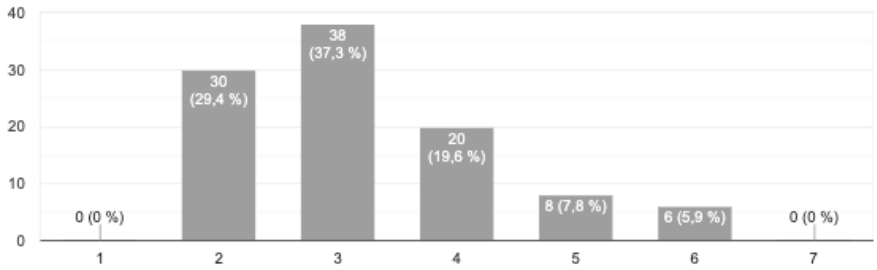


Figure 7.2.: Distribution of genders in the survey



(a) Musical knowledge



(b) Art knowledge

Figure 7.3.: Responses on the participants knowledge in art and music, using a scale of 1 to 7.

7.3. Survey Results

This section reviews the results for all the songs in the survey. The results from the interviews are also presented in this section. Each subsection is divided into two parts – the first looks at how the quadrant categorisation was. The second part focuses on the selection of images and how well the participants felt that their selected images matched

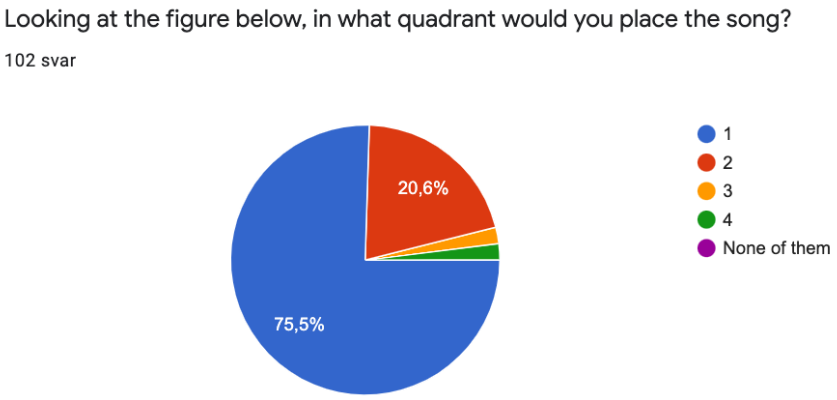


Figure 7.4.: Survey results: quadrant for Thriller. The numbers represent the quadrant, $1 = Q1$ etc.

the song's emotion. The thoughts of the five interview subjects are presented in both parts to get a deeper understanding of the results. The system is evaluated in Section 8.1 based on these results. Appendix D includes all the answers from the entire survey.

7.3.1. Thriller by Michael Jackson

The results show that the participants disagree with the systems quadrant categorisation for the song Thriller. Furthermore, almost 70% of the respondents scored the matching between the best-suited image and the song below average, with the values 1, 2 and 3. Only 2.9%, which is three people, rated their song-image-pair with the value 6. Hence, the system failed to categorise the quadrant according to the participants, and the resulting images did not match the song's emotion. One of the interview subjects also mentioned that it was difficult to form a new opinion of the song because they already have strong memories connected to the song.

Quadrant Categorisation

The system categorised Thriller into the Q4 quadrant. However, 75% of the respondents categorised the song into the Q1 quadrant, as shown in Figure 7.4. In the interviews, the participants mentioned words from the Q1 quadrant in the model, such as "Energised", "Enthusiastic", and "Joyful". They all mentioned the beat and that they felt happy. One said she wanted to get up and dance while listening to the song.

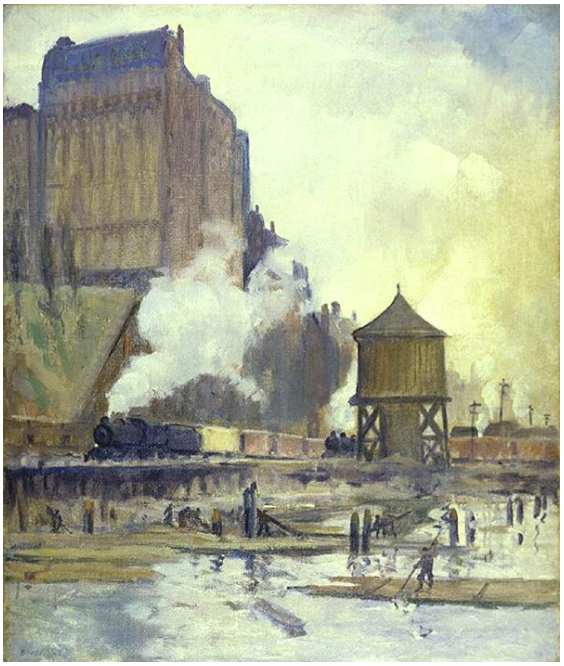


Figure 7.5.: Survey results: image for Thriller.

Image Pairs

Over 55% selected image C for this song, shown in Figure 7.5. Almost all of the interview subjects agreed with this decision. However, one participant thought image C was too dark. They preferred image B because it evoked positive emotions, which they felt matched the song.

The final question asked what sort of painting they would prefer or expect to see with the song Thriller, which provided three exciting results. The first is that 82% would expect an image consisting of dark colours, even though they categorised the song as happy and energising. These emotions are commonly associated with light and bright colours (Hemphill, 1996; Boyatzis and Varghese, 1994). The second interesting finding was that almost 52% would prefer an image with people. In some of the interviews, the participants said they expected people due to the song's danceability. Finally, almost 43% answered that they would prefer an autumn landscape. Some of the interview subjects also mentioned this.

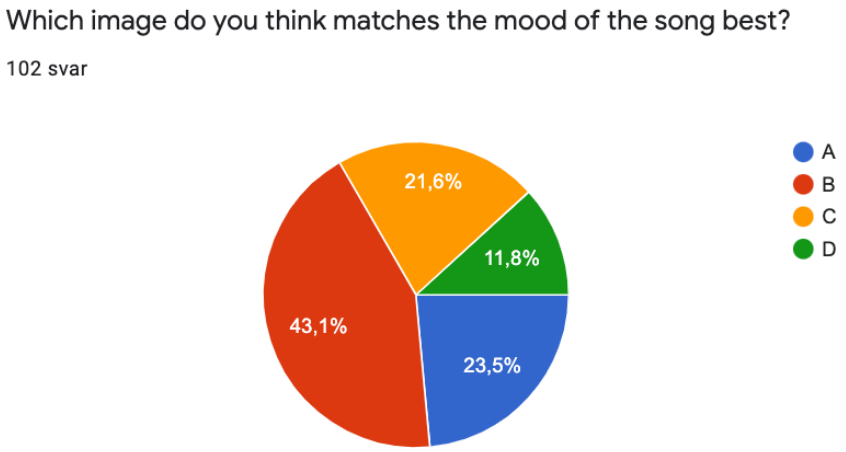


Figure 7.6.: Survey results: answers to the image for Dangerously in Love.

7.3.2. Dangerously in Love by Beyoncé

For the song Dangerously in Love, the participants did not agree as much as with the song Thriller. There was more dispersion in answers on both quadrant categorisation and image selection. The system categorised the song differently than most of the participants. However, most of the participants thought the image they selected was a good match with the song's mood.

Quadrant Categorisation

Two quadrants received the majority of the votes for this song. Q4 was most preferred and was selected by 49%, and the second most selected quadrant was Q3 at 33%. The system categorised this song into Q1, which only received around 10% of the votes from the survey participants. In the interviews, it was mentioned that the song evoked a sad feeling. The song was described as calm, and the lyrics were a mixture of compassionate and sad, placing it between Q3 and Q4, which also seem to be the general opinion of the respondents.

Image Pairs

About 43% chose image B as the best match to the song's emotion. Around 20%, however, chose images A and C, as shown in Figure 7.6. The most selected image, B, is shown in Figure 7.7. 47% of the respondents felt the image they chose was a suitable match to the song, scoring it with the values 5 and 6. The interview subjects had diverging opinions

Figure 7.7.: Survey results: image for Dangerously in Love.

as well. All four images were selected by someone as the best match. Still, almost all of them placed the song in quadrants Q3 or Q4.

Finally, the last question revealed that most participants would prefer an image with people, using light, warm colours and displaying a summer or spring landscape.

7.3.3. Hurt by Johnny Cash

The system categorised this song into Q1, on which the participants disagreed. None of the respondents felt the song belonged in this quadrant. Moreover, they disagreed on which images were the best for this song.

Quadrant Categorisation

Circa 90% of the respondents placed the song into the Q3 quadrant, as shown in Figure 7.8. The remaining responses were equally distributed between Q2 and Q4. The system categorised the song into Q1, which received none of the responses. All the interview subjects agreed that Q3 seemed like the best fit. Many of them mentioned how sad the lyrics were, and therefore they decided Q3 was most fitting.

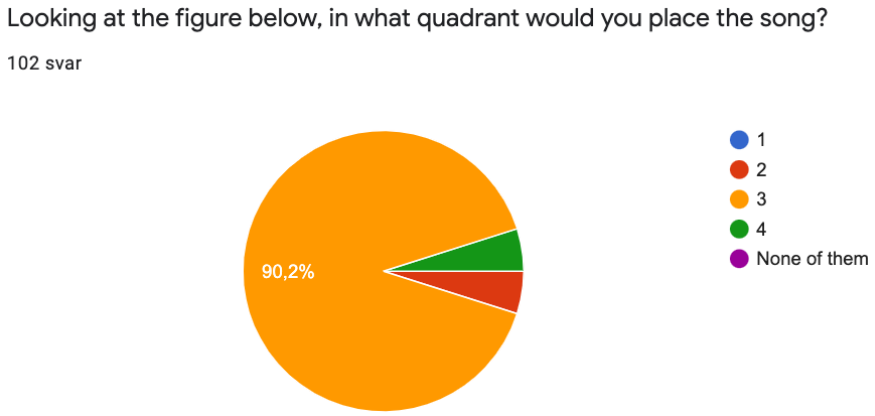


Figure 7.8.: Survey results: quadrant for Hurt.

Figure 7.9.: Survey results: image for Hurt.

Image Pairs

There was no absolute favourite among the four images, but images A and D received the majority of the votes. The former got 30%, and the latter got the most votes with 34%. Image D is shown in Figure 7.9. The interview subjects were also torn between C and D. Some felt that C was too chaotic, and others felt this chaos matched the feeling in the song. Most of them also thought A and B had too light colours and gave a happy feeling that did not match the emotions in the song. Most participants thought the image they chose matched well, with over 60% rating the song-image-pair above or equal to average, with the values 4, 5, 6 or 7.

When selecting features on the final question, over 74% answered that they would expect

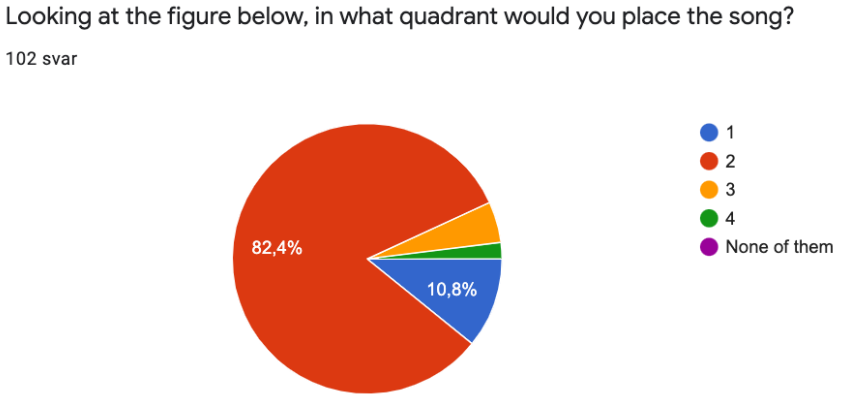


Figure 7.10.: Survey results: quadrant for The Way I Am.

to see an image with dark colours. This was also mentioned in the interviews. Almost 57% also said they would prefer an image without people. One of the interview objects also mentioned that they would prefer a winter landscape due to the winter depression and that they felt this was a fitting season for a sad song. However, only 32% of the respondents agreed with this assessment, and autumn landscapes were more preferred and received 41% of the votes.

7.3.4. The Way I Am by Eminem

The system categorised this song into the same quadrant as the participants selected in the survey. Two images got the majority of the votes in the second question, and most of the participants thought the song and the image they paired with it matched the emotions they evoked.

Quadrant Categorisation

The system categorised The Way I Am into the Q2 quadrant, and over 82% of the respondents agreed with the system, as shown in Figure 7.10. This was the only song that the system and the participants categorised uniformly. All of the interview subjects said the song was rather angry. The words vengeful and disgust were also mentioned. As shown in the figure, over 10% said they felt Q1 was the most fitting quadrant. In one of the interviews, a participant mentioned that they felt energised listening to the song. The participant listened to songs like this one when working out and felt excited. The word energised is placed in Q1, and even though this interview subject categorised the song into Q2, Q1 was considered a decent fit as well.

Figure 7.11.: Survey results: image for The Way I Am.

Image Pairs

Two images were preferred for this song, images C and D. The former received almost 58% of the votes, and the latter received 38%. Images A and B had less than 4% in total. Image C is shown in Figure 7.11. In the interviews, there were dispersed opinions as well. Half of the participants chose D, and the other half chose C, so they agreed with the other survey participants. Image C was said to show the same anger as in the song due to the people's facial expressions. In particular, the woman in front was mentioned for her unpleasant facial expression. Image D depicts waves in the sea, and some of the participants felt a "rough" sea suited the anger in the song. Others connected the sea and waves with calm emotions and did not see it as a suitable match.

The matching rate of the song-image-pairs was reasonable because over 70% said they thought the match was equal to or better than average. Over 38% answered that they would expect to see their chosen image. Almost 76% thought dark colours in the image would be expected. Half of the participants would also prefer an image with people with cool-toned colours. There was also a slight preference for winter and autumn landscapes. However, these received less than 25% of the votes each.

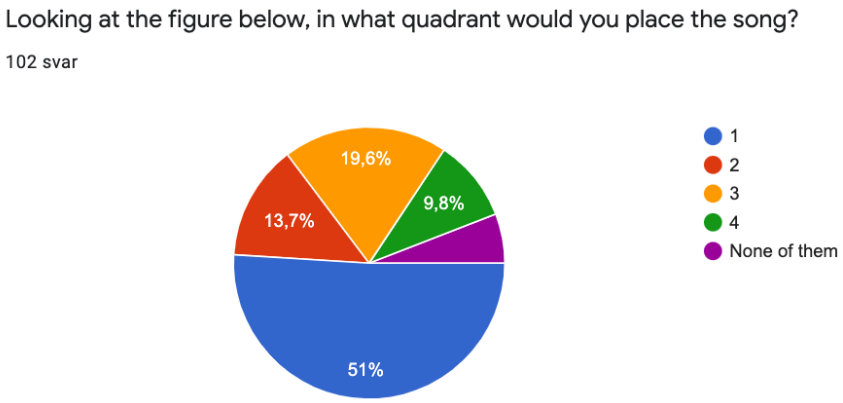


Figure 7.12.: Survey results: quadrant for Rehab.

7.3.5. Rehab by Amy Winehouse

This song was the only one that some participants did not categorise into a quadrant. It was also a wide dispersal of votes on the images, and very few of the participants felt that the song-image-pair they chose was a good match.

Quadrant Categorisation

For the song Rehab, 51% of the participants in the survey placed the song in Q1. The system placed the song into Q4, which only received 9% of the votes. Almost 6% answered that they felt none of the quadrants was a good match for the song. Figure 7.12 shows the results of the quadrant categorisation in the survey. In the interviews, words such as energised, joyful, proud and happy were selected from the model. Several interview subjects mentioned that the song was complex to categorise because the beat and the lyrics gave different emotions. The beat was perceived as happy and energising, but the lyrics are about rehab and drug use. However, the participants mentioned that the lyrics are performed "sassy" and proud. Therefore, Q1 was still the best match.

Image Pairs

Only two answers separate images B and D with 33% and 35% of the votes, respectively. Both of the images are shown in Figure 7.13. Most of the participants in the interview disagreed that D was the best match. They selected image A as the image that matched the song's mood best. They thought image D had too dark colours and did not match the joyful and energising beat in the song. One participant thought it was not easy to select a song-image-pair because they did not think landscape paintings matched the

(a) Rehab image B

(b) Rehab image D

Figure 7.13.: Survey results: images for Rehab.

feeling of being energised.

Most of the survey participants thought the song-image-pair they chose had a match equal to or below average at the value 4. Only 20% thought the match they made was above average at values 5, 6 and 7. No image feature received more than 55% of the votes in the final question. Half of the participants would prefer an image with people, using dark and warm-toned colours. However, in the interviews, many participants said they preferred a mixture of dark and light colours to reflect the contrast between the beat and the lyrics. Some of them also mentioned that they would prefer images with people because of the song's danceability.

8. Evaluation and Discussion

This chapter presents an evaluation of the system followed by a discussion that elaborates further on the topics from the evaluation. In Section 8.1, four subsections are presented to evaluate the different steps when creating the system, the system's results and the survey. The survey results from Section 7.3 will be used in the evaluation. The section will also evaluate the survey itself and suggest improvements that could better evaluate the system. Next, Section 8.2 will discuss the system's potential and limitations and present some possible improvements. This section will also discuss possible answers to the research questions from Section 1.2.

8.1. Evaluation

This section will evaluate the steps used when creating the system, the system's results, and the survey used to evaluate these results. How the different datasets were used, tested, and trained with has also been evaluated.

8.1.1. Categorising Songs

The dataset from Panda et al. (2018) was used as training data to categorise new songs from TheAudioDB into quadrants. The training and testing algorithms used XGBoost to fit a model. Some hyperparameters were tuned and used in training. These were the following: `{max_depth=4, min_child_weight=0, gamma=0.05, colsample_bytree=0.4, subsample=0.6}`. The rest of the parameters were set to default. The values of these parameters affect into what quadrant each song is categorised. For instance, using these values, Thriller was categorised in Q4. However, using only the default parameters, the song is categorised into Q1. Notwithstanding, using the default parameters achieves a lower accuracy on the predictions of the testing set. Using the default parameters results in prediction accuracy of $\sim$ 93%, whilst using the tuned parameters, the prediction accuracy is equal to $\sim$ 95%. A potential problem might be that the model is overfitted using the tuned parameters. Overfitting means that the model is too dependent on the training data, and it will have more errors when trying to fit new unseen data.

Song	Quadrant with DP	Quadrant with TP
Thriller	Q1	Q4
Dangerously In Love	Q1	Q1
Hurt	Q4	Q1
The Way I Am	Q2	Q2
Rehab	Q4	Q4

Table 8.1.: Quadrant categorisation using default parameters (DP) versus using the tuned parameters (TP).

As shown in Table 8.1, the songs marked with orange received a different quadrant using the tuned versus using the default parameters. The survey results in Section 7.3 show that using the default parameters, the system categorised Thriller into the same quadrant as 75% of the survey participants: Q1. Consequently, choosing the correct hyperparameters when fitting the model is very important and may lead to more consistent results with human perception.

In addition to hyperparameters, the metadata from TheAudioDB affects the accuracy of the quadrant categorisation. As presented in Section 4.3, the genres and moods from TheAudioDB and Panda et al.'s dataset do not overlap. Some of the genres in one of the datasets do not exist in the other. One genre in the first dataset is called Pop/Rock, which is divided into two genres in the other dataset; Pop and Rock. However, one of the columns dropped from Panda et al.'s dataset, MoodsStrSplit, could solve this problem since this column splits the values containing two genres. Moreover, some of the moods from TheAudioDB are not found in the other dataset, and hence it is not easy to compare this information with the metadata from Panda et al..

Furthermore, artist information is label encoded. Hence, if an artist from TheAudioDB has the same name as one from Panda et al.'s dataset, they will receive the same encoded label. However, if there is a misspelling in either name, they will not be matched. If TheAudioDB also includes information about featured artists that Panda et al. excluded, they will not receive the same label. Finally, there are less than 900 artists in Panda et al.'s dataset, while TheAudioDB has songs from over 50 000 artists.

The categorisation will be difficult if a song from TheAudioDB misses information about genre and mood and the artist is not present in Panda et al.'s dataset. The system has no applicable information about the song, and the categorisation will probably happen at random. Unfortunately, this is most likely the case for around 80% of the songs from TheAudioDB since the metadata about mood and genre are Null[1]. If this part of the system is to be improved, using another database with more information about its songs is beneficial. More ideas on improvements are listed in Section 8.2.

[1]Statistics from `https://www.theaudiodb.com/stats.php`.

8.1.2. Image Datasets

Two datasets were used to retrieve images from the quadrants, the WikiArt Emotions' annotated images and seasonal photographs from the Flickr database. The first had images annotated with emotion labels, and the second was pictures taken by Flickr users. The latter was used in Zhu et al.'s image-to-image translation to transform the images into paintings. This subsection evaluates the use of these two datasets in the system.

WikiArt Emotions

The WikiArt Emotions dataset contains images that have been labelled with emotions through crowdsourcing. The emotion labels in the dataset were used in this project to categorise the images into quadrants. The words from the Bliss-Moreau et al. (2020) model in Figure 3.1 were used to find the emotions in the WikiArt dataset to categorise each emotion label into a quadrant. This process is described in more detail in Section 4.1. If a word was not found in the model, either a synonym was used, or the metadata in Panda et al.'s dataset was used as inspiration. For some of the emotion labels, their quadrant was merely a guess. These methods present some potential shortcomings, as discussed below.

The first method seems to be the most accurate; finding the exact word or a stemmed version of this word in the model to find the most suitable quadrant. For instance, "happy" and "happiness" both involve the same emotion as the latter means "the state of being happy", according to the Cambridge Dictionary[2]. The second method is not as accurate. Most of the synonyms were paired from the creator's mental dictionary. For instance, "admiring" was chosen as a synonym for "trust". However, they are not found to be accurate synonyms in the Cambridge Thesaurus. Panda et al.'s database as a synonym dictionary is not an optimal solution either. However, the worst method was the final one: guessing which quadrant an emotion belonged to based on the existing words in each quadrant. These guesses were solely subjective from the creator's mind and may not be the general population's opinion.

A prestudy could have been performed to categorise these emotion labels better. Participants could be presented with all the emotional labels and the models from Russell (1980) and Bliss-Moreau et al. (2020) and then categorise all the labels into quadrants. Even though this would result in yet another subjective categorisation, it would present the average of many participants rather than simply the opinion of one individual.

Correct quadrant categorisation of these images is essential to ensure a sound system. When a song has been categorised into a quadrant, an image from WikiArt Emotions that belongs in that quadrant will serve as output. However, if the initial categorisation

[2]Definition of happiness: https://dictionary.cambridge.org/dictionary/english/happiness.

(a) Keyword: happy winter landscape

(b) Keyword: sad winter landscape

Figure 8.1.: Winter landscape images from Flickr.

of WikiArt images is flawed, the output image may not evoke the same feeling in the observer as the song did. Consequently, the user may think the system itself provided poor results when, in fact, the system's preparation was poorly done.

Seasonal images

Images were fetched from the Flickr database using their API with a keyword such as *[season] landscape* and a sorting method. The parameters used were the following: `text=keyword, tag_mode="all", tags=keyword, extras="url_c", content_type=1, per_page=100, sort="relevance"`. Around 1000 images were fetched for each season: autumn, spring, summer, and winter. The sorting method selected the images that Flickr estimated as most relevant to the keyword. Even though the keyword specified landscape images, some images did not fulfil this demand, as shown in Figure 4.3. Hence, the keyword could have been improved to ensure that all the images were, in fact, of season landscapes. This improvement could involve a longer keyword such as "Landscape images of winter nature" or variations of the same words, e.g., "winterlandscape landscapswinter winter landscape".

Words from Russell's model could also be used as keywords to get landscapes images that are already categorised with different emotions. For instance, there is a big difference between a happy winter landscape picture where the sun is shining and a sad or gloomy winter landscape with fog. An example of two different winter landscape photos is shown in Figure 8.1. The image in Figure 8.1a may have been placed in Q1 due to the word "happy", and Figure 8.1b may have been placed in Q3 due to the word "sad" in Figure 2.1. The picture datasets might have been more suitable if these keywords had been used to download images into different quadrants.

After the images were downloaded, they were transformed into Monet-like paintings using image-to-image translation. As described in Subsection 6.2.2, Monet paintings and

(a) Real Monet painting

(b) Real autumn picture

(c) Picture to painting with 50 epochs

(d) Picture to painting with 100 epochs

Figure 8.2.: Difference between 50 and 100 epochs in image-to-image translation

landscape photographs collected by Zhu et al. (2017) were used to train their CycleGAN. However, the system was trained using only 50 epochs due to time constraints, even though the default number is set to 100. Using fewer epochs creates worse results, which means that the translated paintings might not look like actual paintings, but rather blurry photographs. An example of a poor image-to-image translation with 50 epochs is shown in Figure 8.2. The figure shows four images: an actual Monet painting, a real photograph and the photograph translated using two models, one trained with 50 and the other with 100 epochs. Training the system with 100 epochs was conducted after the experiments to see if there would be better results. Figure 8.2c and Figure 8.2d display the autumn photograph after being translated into paintings. Comparing them to the

actual Monet painting, the model trained using 100 epochs made a more convincing impersonation due to the textures and colours in the image. Hence, training the model using at least 100 epochs might improve the quality of the resulting images.

When the photo-to-painting translation is good, the output painting looks more convincing or authentic and is perhaps easier on the eyes. The system's purpose is, first and foremost, to unite visual and auditory art so that a music piece and an art piece share the same perceived emotion. Nonetheless, a pleasing painting to look at may ensure that this unity appears more potent or compelling to the user.

8.1.3. Survey Set-up and Possible Improvements

The survey was created in order to evaluate the system. Four images were presented for each song, and the participants were asked to evaluate the song-image-pairs that the system created. These images were retrieved from the quadrant that the system categorised the song into. Since all the images were from one quadrant, it is understandable that the survey participants rated the song-image-pair poorly if they felt the song belonged to a different quadrant than the system. For instance, Thriller was categorised in Q4 by the system, so the four images were all chosen from Q4. The participants categorised the song into Q1, and over 87% of them rated the song-image-pairs equal to or below average, with the values 1, 2, 3 and 4 out of 7. Considering all the images were fetched from the Q4 quadrant, this makes sense.

As mentioned, four images were chosen for each song. One of the interview subjects mentioned that presenting an increased amount of images would be more beneficial. The reasoning was that the participant took the time to analyse each one with only four images, which only made the selection process more demanding. The subject suggested using more images, perhaps ten images for each song. The theory was that the participant would glance through all the images quickly and select the one they felt stood out as the best match.

More than five songs could also be included in the test and the survey. Including more songs might make it easier to discover biased outliers or specific patterns. Nevertheless, more songs would make the survey longer, potentially reducing the number of participants finishing the survey.

Even though one of the questions reveals the participants' general knowledge of music, none of the questions asks whether or not the participants are familiar with the test songs. One of the interview subjects mentioned that it was difficult to form a new opinion of the songs they already knew since they had strong memories connected to the songs. This might be the case with multiple participants because the songs are relatively famous. Hence, a question for each song asking if the participant is familiar with the song or not would make it easier to rule out or detect preconceptions or bias.

Two of the resulting images for each song came from the WikiArt Emotions dataset. The other two were seasonal images placed in different quadrants based on the hypothesis presented in Subsection 6.2.2. Since there was no guarantee that the participants would categorise the song into the same quadrant as the system, it might have been better if there were images for each song that would represent each quadrant in Russell's model. For instance, each song could present eight images, one from each quadrant from WikiArt Emotions and one from each season. This strategy might answer the following question: Do the participant select an image from the same quadrant that they categorised the song? It might also give answers to research question 2 from Chapter 1. With the existing survey, it is problematic to conclude whether or not the images' quadrant categorisation was accurate. Including more images in the survey might be a way to evaluate how well the quadrant classification of the images was. If the classification was well executed, the images that belong to the same quadrant as the participants selected should also receive the most votes.

Some of the interview subjects and others who participated in the survey remarked how difficult it was to pair pop songs with landscape images. As the survey results show, images including people for some of the songs would be preferred or expected. Hence, it could have been interesting to have at least one image for each song from WikiArt Emotions that were labelled to include a body or a face. Some images that have been labelled with the face or body tag are shown in Figure 8.3. In theory, for the songs where the participants would prefer images with people, these alternatives should receive the most votes. Notwithstanding, most of the images from WikiArt Emotions have western painting styles from areas like romanticism, realism or neoclassicism. Unfortunately, there are few images of people dancing, which was mentioned as a preference from some participants. Hence, the images that include a face or body may still be challenging to compare and match with modern pop songs.

The final question on each song in the survey asked what painting the participants would expect or prefer to see with the specific song. This question included alternatives of [season] landscape for all four seasons. Another way to present these alternatives would be to include a separate question that showed one or more images from each season and asked the participant to select the season image they preferred for the song.

8.1.4. System Results

This subsection evaluates the system's results on each test song based on the responses from the user survey and the interviews. Four images for each song were selected, whereas two of them stemmed from the WikiArt Emotions dataset and the other two were seasonal images. The survey results are presented in Section 7.3, and screenshots of all the questions and answers are shown in Appendix C and Appendix D.

(a) Image that was labelled with face

(b) Image that was labelled with face

(c) Image that was labelled with body

Figure 8.3.: Three images from WikiArt Emotions including face or body

Thriller by Michael Jackson

The survey results show that the participants categorised the song into Q1, while the system placed the song into the Q4 quadrant. The mood of the song was labelled as "Quirky" in TheAudioDB. Scanning through Panda et al.'s dataset, 9 out of 16 songs labelled with "Quirky" were categorised into Q4. 4 out of 16 were labelled with Q1, and the remaining three songs were categorised into Q3. Hence, the system utilised the training data well since most of the training songs with the same mood were categorised in Q4. Consequently, 900 songs may be too few for training because there are very little data that decides which quadrant to select. The song was labelled with the genre "Pop", which does not exist in Panda et al.'s dataset because they have merged pop and rock into "Pop/rock". Hence, this information was not used. Had this genre been split in this metadata so that it would be compared to TheAudioDB's genres pop and rock, the results may have improved.

The images chosen for this song were one landscape image from WikiArt, one random image from WikiArt, one image of a spring landscape, and the last was a summer landscape image. The image that the most participants selected was the random image from WikiArt Emotions (image C). This image was labelled with the emotion "Humility". The second most preferred image was the other image from WikiArt Emotions that displayed nature or a landscape (image D) and was labelled with three emotions: "Happiness", "Humility", and "Optimism". These images were placed in Q4 due to the "Humility" label. As described in Section 4.1, this word was not categorised in Q4 based on research or facts but as a guess or speculation. Hence there is no guarantee that Q4 is the correct quadrant. Nonetheless, the participants were asked to rate how well the image they chose matches the song on a scale of 1 to 7. The average score for those who chose image C was 2.74, and the average score from the participants who chose image D was 2.97. This may indicate that the images labelled with "Humility" should not be categorised into Q1, and therefore, Q4 is still a decent guess.

Less than 10% of the participants thought an image of a winter, summer or spring landscape was expected with this song. Over 51% would expect to see an image including people, and almost 43% of the respondents thought an autumn landscape could be suitable. Even though quadrant Q1 includes emotions such as "joyful", "enthusiastic", and "energised", an autumn landscape was more preferred, which does not support the hypothesis from Subsection 6.2.2. Speculation is that many people have seen the original music video for this song[3]. This video consists of people dancing in the street at night and based on the surroundings, it seems like it might be late autumn or early winter. Hence, the participants might expect to see an image similar to the music video.

Since the participants would prefer to see an autumn landscape, it would have been

[3]The original music video to Thriller on YouTube: https://www.youtube.com/watch?v=sOnqjkJTMaA&t=277s

interesting to see if they had chosen this over image C. Around half of the participants selected image C, and around half expected to see an image of an autumn landscape. Therefore, it would have been exciting to see how much this distribution changed had an autumn landscape been an alternative.

Dangerously in Love by Beyoncé

The system categorised Dangerously in Love in the Q1 quadrant. However, less than 10% of the participants agreed with this categorisation. 49% of the participants placed this song in the Q4 quadrant, and 33% preferred the Q3 quadrant for this song. TheAudioDB labelled Beyoncé's song with the mood "Energetic", similar to the word "Energised" in the Q1 quadrant in Bliss-Moreau et al.'s model in Figure 3.1. "Energetic" is also used in Panda et al.'s dataset, and 35 out of 37 songs labelled as energetic are also categorised in Q1. The song was labelled with the genre "Funk", which is not found in the testing set. For this song, it seems like it was the mood labelling in TheAudioDB that ensured poor results.

The most selected image for this song was image B - the spring landscape photograph translated to look like a painting. In some of the interviews, the flowers were mentioned to be decisive since the song is about love. The second most popular image was image A - the painting created using a photograph of a summer landscape. Hence, the two seasonal images were most preferred next to the song that most participants categorised in Q4. Out of all the participants that chose Q4, 56% preferred image B. These results support the hypothesis that spring landscapes are well suited to the Q4 quadrant.

The final question for this song also supports the hypothesis from Subsection 6.2.2. Over 45% of the participants would expect to see an image of a summer landscape, and spring landscapes were expected by over 35%. It is below half of the participant population, and therefore no grand conclusion can be drawn. However, it is noteworthy that this is more than the number who would prefer either winter or autumn landscapes.

Hurt by Johnny Cash

This song appears to be the biggest flop of the system. The quadrant categorisation of the song placed it into Q1. However, over 90% of the participants categorised this song into the Q3 quadrant. In the interviews, the word "Sad" from Q4 was mentioned by all the subjects, which is quite the opposite of the emotions in Q1 in Figure 3.1. TheAudioDB provided information about this song, giving it the genre "Country" and the mood "Provocative". Hurt's mood is not present in Panda et al.'s dataset, but its genre is shared with 85 songs. 21 out of these 85 are also labelled with Q1, but as many as 33 were labelled with Q3. This suggests that the categorisation algorithm that uses

Panda et al.'s dataset as training data is underfitting the model during training. This underfitting might happen because there were little data to train with.

Two images received around 30% of the votes, image D with 34% and image A with 30%. The former is from the WikiArt Emotions dataset and was labelled with "Happiness". It is interesting that the image with the most votes had this label since the survey participants categorised the song as "Sad". Some of the interview subjects that selected this image explained that the dark colours in the image concluded their choice. However, another interview subject explained that the image reminded them of summer vacation because of the city skyline and the sunset. Speculation is that since the image was pretty small on the screen, the participants might overlook the happy motives and that their emotional response was provoked strictly by colour and weather. Furthermore, the participants who chose image D rated the song-image-pair at an average of 4.3, which is slightly above average on a scale of 1 to 7.

The final question for this song asked the participants to select the image features they would expect or prefer to see along with this song. Almost 75% answered that they would expect to see an image with dark colours and without people. For the song Thriller, it was mentioned that an image with people might enhance the feeling of happiness if the people were, for instance, dancing in the image. Hence, the participants might have chosen the opposite for the song Hurt because they categorised it as sad and not happy. Furthermore, the survey results show that the participants prefer a winter or autumn landscape with this song, which supports the previously mentioned hypothesis.

The Way I Am by Eminem

The system and the participants both categorised this song into Q2. Over 82% of the respondents placed the song in this quadrant. TheAudioDB had labelled this song with the mood "Angry" and the genre "Hip-Hop". The genre is not present in the training dataset, but the mood is found in 74 songs. 71 out of the 74 songs labelled with "Angry" were also categorised in Q2. Songs that include one of the four emotions from the quadrants in Figure 2.1: happy (Q1), angry (Q2), sad (Q3) or relaxed (Q4), might be more straightforward for the system to categorise. Looking at Panda et al.'s dataset, one can see that songs labelled with one of these four emotions are categorised into their belonging quadrant in 90% of the cases. However, it is optimal that the system categorises the songs correctly even though they are not labelled with either of these emotions.

Image C and image D received the most votes for this song, where the former got the most at almost 58%. This image originated from WikiArt Emotions' dataset and was labelled "Anger". Hence, it is understandable why most participants thought this was the best match for the song. The average match rate for image C was 4.7, where 36 out of 59 people rated the match above the value of 4 on a scale of 1 to 7.

Few participants, i.e. only four people, preferred the season images A and B, displaying autumn and winter, respectively. This is reflected in the final survey question for the song, where only around 22% would expect or prefer an image with either autumn or winter landscapes. However, as explained in Subsection 8.1.3, it can be a big difference between winter landscapes based on the weather. Since 75% of the participants answered that they would expect an image with dark colours, it would be interesting to see if they had chosen the winter landscape if it depicted a gloomier day with bad weather instead of a sunny mountain. Nevertheless, these results do not support the hypothesis that winter or autumn landscapes match well with songs from the Q3 quadrant. Moreover, over 50% responded that they would expect an image with people, so a landscape painting may not be suitable.

Rehab by Amy Winehouse

The system wrongfully categorised the song into Q4 since 51% of the participants placed the song Rehab into the Q1 quadrant. The song was labelled with the mood "Troubled" and the genre "Soul" by TheAudioDB. Neither of these was found in Panda et al.'s dataset, so there was little information to use in order to categorise the song into a quadrant. When testing with different songs from TheAudioDB that did not contain any information about genre or mood, the system almost exclusively categorised it into Q4. Since there are 225 songs from each quadrant in Panda et al.'s dataset, it would be more comprehensible if the system selected each quadrant 25% of the time if no data foundation was present. However, this was the only song that some participants failed to categorise as almost 6% selected the option "None of them" instead of a quadrant. It might have been difficult for the respondents to categorise the song since the beat and the lyrical topic are different.

There was an even dispersal of votes regarding the best matching image. The most selected image, D, only received 35% of the participants' votes. The second most preferred received 33% of the votes, image B. Image D was labelled with "Humility" in the WikiArt Emotions dataset, one of the most challenging emotions to categorise. It was merely a guess that this emotion belonged in the Q4 quadrant. However, among the people who selected image D as the best match, only two participants categorised the song into Q4. Most of the participants selected either Q1 or Q3 along with image D.

Finally, half of the participants responded that they would expect or prefer images with people, a mixture of light and dark colours, or warm-toned colours. Landscapes depicting summer, autumn or spring all received over 25% of the votes, and images with summer landscapes received the most votes at almost 40%. These results support the hypothesis that summer or spring landscapes are suitable for Q1 songs. However, almost 55% prefer an image with people, which might be favourable over a landscape without people.

Conclusion

The results show that it was disputed whether the WikiArt Emotions images or seasonal landscape images were the most accurate dataset to use in this system. The former dataset had predefined emotion labels, and the latter had a hypothesised quadrant categorisation to split the images into four quadrants. Even though the seasonal images were not preferred for all the songs, it seems as though the hypothesis was correct. Summer and winter landscapes were more preferred for songs in Q1 and Q4, while autumn and winter landscapes were more preferred for songs in Q2 and Q3. This was true for all test songs except Thriller.

For the WikiArt Emotions dataset, the emotion labels were used to categorise each image into a quadrant. The technique used for this is described in Section 4.1. The results show that the categorisation was not well executed for all emotion labels. Especially the label "Humility" provided conflicting results. For the songs *Thriller* and *Rehab*, an image with this label was selected as the most popular. The images were retrieved from Q4 because "Humility" was categorised in this quadrant. However, both of the songs were categorised in Q1 by the participants. Hence, images labelled "Humility" might be better suited in Q1 than Q4.

Based on the data from TheAudioDB, the system's song categorisation technique seemed to work for around half of the songs. Thriller, Dangerously in Love and The Way I Am all had information present in Panda et al.'s dataset and used this training data correctly to categorise the songs. However, the song Hurt had metadata familiar with many songs in the training data, but this information was not appropriately utilised, and the song was categorised into another quadrant. Since Rehab shared no information with the training data, it was blindly categorised into Q4. These results may indicate that TheAudioDB is not an excellent database to retrieve metadata about songs since it does not include the same information or metadata as Panda et al.'s dataset. This dataset has retrieved song information from the AllMusic API mentioned in Section 4.5. It could be beneficial to use this database for metadata about never-before-seen songs to improve the system.

8.2. Discussion

Due to time constraints, only a Minimum Viable Product of a creative system was created. The system can receive a song ID as input and provide some suitable images as output. However, as concluded in the previous section, the system presents some issues where the output image does not perfectly match the input song. The level of how well the images matched the song presents the most room for improvement. This section discusses the system's potential and limitations and possible alterations or techniques that might improve the results.

8.2.1. Potential

The project's goal is to unite auditory and visual art in the forms of music and paintings. This has been accomplished using different machine learning techniques. Moreover, this system shows potential in the field of computational creativity. It pairs songs and images based on their emotion. However, the system only uses ground truth data about songs to assume what feeling they may evoke in the listener. In Section 2.1, Computational Creativity is presented in more detail, and the system creator has experienced P-creativity throughout the project. Novel and exciting ideas have been tossed around and incorporated into this project. Comparing and pairing the four meteorological seasons to Russell's four quadrants have never been done before, even though researchers have investigated the relationships between mood and weather (Watson, 2000; Ennis and Mcconville, 2004; Denissen et al., 2008; Huibers et al., 2010; Melrose, 2015). As presented in Section 2.1, there are different types of computational creativity. This project might fit nicely under the first type: combinational creativity, which involves novel combinations of familiar ideas.

The system has the potential to become a valuable tool for music distributors, musicians, painters, and for people to create or select artwork that match a specific song or album. TheAudioDB and the AllMusic API include metadata about songs that can be used in the system. Hence, musicians can use this tool when creating new music based on their old songs. TheAudioDB was used in this project, but the AllMusic API seems like a good database for further work. Painters may also use the system as an inspiration tool. If they want to construct a new art piece, they may use the system with songs they are inspired by to see what the system considers a matching painting to that song. Furthermore, it is a fun system that can be used by people who want to play around with pairing songs and images.

For the system to have value for music distributors such as Spotify, some improvements are necessary to ensure more suitable song-image-pairs. When the system provides good matches, it can be used in Spotify's "discover weekly"-playlist or their "Spotify wrapped" to present interesting paintings to their customers unique to their favourite songs.

8.2.2. Limitations

Using TheAudioDB as the database for songs has limited the system. As described in Section 4.4, not all the moods or genres from this set were present in the training set from Panda et al. (2018). The latter used song clips and metadata, i.e. both music features and ground truth data from AllMusic API, to categorise each song into quadrants. This system only uses ground truth data from another database when categorising never-before-seen songs. Since there are inconsistencies in the training and testing data, the system is not always able to correctly categorise which quadrant a song belongs in.

Another reason for poor quadrant categorisation is the lack of metadata from TheAudioDB. Most songs do not have any information about genre or mood, which was most important for the system to ensure good categorisation. As discussed in Subsection 8.1.4, lack of information resulted in random categorisation that often placed songs into Q4, even though the songs were from different artists and eras.

None of the songs was categorised in Q3 by the system when selecting the test songs. This has limited the chances of adequately evaluating the system and the hypothesis from Subsection 6.2.2. Since no song was categorised as Q3, none of the images from Q3 was used in the evaluation. Hence, it was not possible to test whether or not the labels from WikiArt Emotions were categorised correctly for Q3, as presented in Section 4.1.

When running the system for this Minimum Viable Product, it was evident that the first set of results was poor. As presented in Section 6.3, some of the quadrant categorisations seemed odd. However, the creator wanted to stay subjective and not change anything based on personal preference. Hence, the first set of results was presented in the user survey and evaluated. Not surprisingly, most of the survey participants had the same perception as the system's creator. Thus, the goal to stay objective limited the possibility of ensuring good results with one iteration.

The results show how important it is to have enough metadata and choose an excellent database to provide this. It has been suggested multiple times throughout this chapter that TheAudioDB may not have been the optimal choice. It has limited the accuracy of quadrant categorisation of the songs due to the previously mentioned lack of information or inconsistencies with the training set. Using the same database as Panda et al., the AllMusic API, might have improved the results since the testing and training data most likely would be more consistent.

Not using the correct keyword to fetch images from Flickr is also a limiting factor. As presented in subsubsection 8.1.2, a winter landscape can look somewhat different based on weather or location. Hence, if different keywords had been used to create a better dataset with seasonal landscape images, the survey results might have shown a more significant preference for the landscape images.

Furthermore, only displaying four images from the same quadrant for each song may have limited how useful the survey results were. Had there been more images from different quadrants, more participants might have agreed on what image they thought matched the songs best. Including an image from each quadrant would also test if participants chose the image from the same quadrant as the one they selected for the song. Hence, the placement of images in quadrants would also be evaluated better.

Using Russell's quadrant as the foundation for emotion classification may also provide some limitations. As seen in the interviews detailed in Section 7.3, some participants experienced a mixture of emotions on different songs. For instance, it was mentioned

that Rehab was difficult to categorise into one singular quadrant since the beat and the lyrical content were evoking different emotions. Hence, Russell's model might be too restricting when only using the quadrant as four pieces. Had the participants been asked to rate the song's Valence and Arousal on a scale of -10 to 10, it might be easier for them to focus on both the lyrics and the beat.

8.2.3. Improvements

The training, testing and parameter tuning described in Section 5.2 and Subsection 8.1.1 can be improved to enhance the system. Since the training and testing data were in different formats, more testing could ensure good parameter values that avoid over- or underfitting. For instance, songs from the training set could be fetched from TheAudioDB and tested using different parameters until the system successfully categorised them into the same quadrant as in Panda et al.'s dataset.

Using some of the songs from Panda et al.'s dataset in the survey would be interesting to test the system's quadrant categorisation before evaluating results through a survey. This would test whether the system could categorise the songs into the same quadrant based on metadata from TheAudioDB. Using songs without lyrics would also be interesting to test because the listener would not experience the confusion as they did with Rehab. Consequently, more people might categorise the song equally as there is no difference between those who emphasise the melody versus the lyrics.

Finding better keywords when downloading images from Flickr might also improve the system. Since the previously mentioned hypothesis states that winter and autumn landscapes would fit well with the quadrants Q2 and Q3, words from these quadrants could be included in the keyword. For instance, instead of searching for "winter landscape", the keyword could be "sad angry winter landscape". The same applies to summer and spring landscapes where the keyword could reflect quadrants Q1 and Q4, for instance, "happy calm summer landscapes". Better datasets would hopefully improve the match rate in the song-image-pairs.

Categorising the emotion labels in the WikiArt Emotions set into quadrants shows room for improvement. Some of the labels were placed into a quadrant due to a guess of the project's creator, which is highly subjective. A prestudy could ensure that the categorisation was done more objectively. Crowdsourcing a user study to people from different age groups, nationalities, genders, and with different levels of artistic knowledge could provide a good foundation for categorisation. This technique would still give a subjective result, but it would be more objective than a guess from one individual.

Increasing the number of epochs parameter, `n_epochs`, to at least 100, the default for Zhu et al.'s system, might improve the image-to-image translation. Even though this does not directly affect the matching or creation of song-image-pairs, the transformed

paintings might be more pleasing to the eye. This might improve the overall impression of the system's user and, therefore, might indirectly improve the match rate and preference of the season images.

8.2.4. Research Questions

Throughout this project, the research questions were used as guidance when creating the system and when creating the user survey. This subsection will discuss the questions and how the system can be used to answer these.

Research question 1: What meteorological seasons couple best with which emotions?

A hypothesis based on research was created that winter and autumn would couple best with being sad or angry and that summer and spring were best coupled with happy or relaxed. Participants in a user survey were asked which seasonal landscape images they would prefer to see while listening to a song. The results show that the hypothesis was true for all except one song.

Research question 2: Will users prefer generated seasonal landscape paintings or original paintings labelled with emotions when listening to music?

The participants were asked to select one out of four images in the user survey. Two of them were generated seasonal landscape paintings, and the other two were original paintings labelled with emotions. The results from the survey show an even dispersal of answers, but the labelled images from WikiArt Emotions were chosen almost 65% of the time. Hence, it can be concluded that the users preferred these over the seasonal landscapes. However, a poorly constructed user survey might also contribute to these results. They might have been different if the survey had included images from all four quadrants for each song.

Research question 3: How can the system that pairs art and music be analysed and evaluated?

The system that was created was analysed and evaluated through a user survey. However, the participants of this survey were mostly Europeans in the age group 18-25. Therefore, the system has not been sufficiently evaluated by people from different cultural

backgrounds or with an even dispersal of age groups. Nonetheless, a user survey is an excellent way to evaluate such a system because the goal is to create a system that makes choices as humans would. If a person would categorise a song into Q1, then the system should too, and it is difficult to analyse whether the system did a good job or not without human involvement. Furthermore, interviewing candidates is also a suitable evaluation technique as it provides room for discussions and explanations.

9. Conclusion and Future Work

This chapter will conclude the work done in this thesis. First, the research goal from the introduction is used to draw some conclusions. Next, the main contributions of this project are presented. Finally, some proposals for future work are introduced.

9.1. Conclusion

The goal of this project was, first and foremost, to unite visual and auditory art so that they evoke the same perceived emotion. This has been accomplished by creating a Minimum Viable Product (MVP) of a system that uses Russell's quadrants to establish emotion categorisation. However, the system did not provide satisfying results as an MVP. The system has many limitations and should be improved through multiple iterations to correspond with human perception.

Utilising only the four quadrants in Russell's model and not the levels of the two dimensions of Valence and Arousal limits the range of emotions. Regardless, it turned out to be a solid and easy-to-use foundation. The user survey showed that half or more participants agreed on a quadrant for each song. However, the system categorised four out of five songs into a different quadrant than the majority of the survey participants. Hence, even though the quadrants are a good foundation, the system utilised them poorly.

The system's categorisation algorithm needs improvement to secure a better accuracy of human emotional perception. The system struggled to categorise the songs into quadrants matching the survey results using only metadata, especially since this metadata was sparse. Tuning the parameters better with more thorough testing through iterations could improve the training of songs to provide a more suitable quadrant for never-before-seen songs. Furthermore, TheAudioDB turned out to be a poor choice of a database for test songs since it has few songs with high data density. Using this database led to diverting results from the survey participants and the system.

The mapping between emotions in the WikiArt Emotions Dataset and Russell's model was imprecise and subjective. Guessing what quadrant some of the emotions could map to was vague and more work should be put into this process. Performing a prestudy could provide a more objective and meticulous mapping which might improve the system's

results.

Moreover, there was no obvious winner among the survey participants between the WikiArt Emotions dataset and the seasonal painting dataset made from scratch. The results might be different and more definite if the latter were formed using more specific keywords related to each emotion or quadrant. Keywords such as "sad angry winter landscape" might provide images that would support the hypothesis from Subsection 6.2.2 better than those used, e.g., "winter landscape".

Furthermore, evaluating the system with a user survey was a helpful way to receive many answers in a short period. Over 100 answers were collected in one week. However, showing images from only one quadrant did not give room for a thorough evaluation of the mappings and keywords in the previous paragraphs. Only one song included images from Q2 and Q3, while the other four songs depicted images from Q1 and Q4. Therefore, the evaluation of quadrant categorisation and image selection for Q2 and Q3 was insufficient.

Despite the system's limitations, the results show a slight preference for winter and autumn landscapes in songs categorised as Q2 and Q3. Furthermore, there is a slight preference for summer and spring landscapes with songs from Q1 and Q4. Including images from all seasons on each song could further support these findings.

9.2. Contributions

In this Master's Thesis, a system is implemented utilising state-of-the-art MER data from Panda et al. (2018) to categorise songs into Russell's quadrants. When Panda et al. created their dataset, they wanted it to be a standard tool for further work in the MER field. This project proves that it is possible and provides a solid foundation to learn from when music features are not used to categorise data.

Furthermore, the created system reuses state-of-the-art image-to-image translation with CycleGAN to create new landscape paintings based on photographs. The landscape paintings were used to prove a hypothesis that winter and autumn landscapes are associated with the quadrants Q2 and Q3, whilst summer and spring landscapes are associated with Q1 and Q4. The images were translated to paintings to see if survey participants would prefer paintings that had already been labelled with an emotion. Moreover, the system's creator thought the technology was exciting and wanted to use it, even though it was not used in a novel way but solely using a new dataset.

Finally, the system translates from song to image using Valence and Arousal as the foundation of emotion. Even though exact values of these two dimensions are not utilised, Russell's quadrants are the foundation of emotion categorisation in the project. This

proved to be a good foundation because it is easy to understand. The survey participants categorised the songs into quadrants, even though many of them have probably never used this model.

9.3. Future Work

As presented in this and the previous chapter, the song categorisation shows the most room for improvement. Using TheAudioDB did not provide accurate results based on the user survey. Hence, fetching metadata from other databases, such as the AllMusic API, might improve this step.

This system only utilises metadata to analyse and categorise songs into one of Russell's quadrants. Including music features to analyse audio clips of the songs might improve the Music Emotion Recognition (MER) step and, hence, the quadrant categorisation. Panda et al.'s novel audio features could be reused to ensure a more precise state-of-the-art MER. Moreover, combining metadata/ground truth data and music features might ensure a solid foundation to categorise the songs into a quadrant.

Creating a better mapping between the emotion labels and quadrants in the WikiArt Emotions dataset could significantly improve how well the images match the song. Crowdsourcing a prestudy could ensure participation from all nationalities, gender and different levels of artistic knowledge. A survey could be posted on different platforms where participants receive a fee for answering and, in that way, ensure participant diversity.

The image-to-image translation and style transfer from Zhu et al. (2017) can be explored further. The research on mood and weather preference presented in this thesis may be used to create a new hypothesis, e.g., rain and snow are best coupled with Q2 and Q3. Hence, any song in Q2 or Q3 could be coupled with an image transformed to have snow or rain using Zhu et al.'s algorithm.

Incorporating text analysis of the song's title might also improve the system. If a song title includes the words sad, gloomy, unhappy, or heartbreaking, it is more likely suited for Q3 – sad than Q1 – happy. The models from Russell (1980) or Bliss-Moreau et al. (2020) could be used as a foundation to select what words belong in which quadrant and find other research that uses Russell's model with quadrants in text analysis.

Some of the survey participants mentioned difficulties in selecting one quadrant if the lyrics and beat of a song provoked different emotions, as suggested with the song Rehab from Subsection 8.1.4. Since Panda et al. (2018) only focused on the melodic aspects of each song and not the lyrics, it could have been clarified to the survey participants that they also should focus on the melody alone. This might have provided different results

to the user survey that may be used to improve the system.

The system's results are only written to a `.out` file stating the file name of the selected images for a song. Creating a frontend web app could make this system more accessible for people who want to play with the technology or make testing more enjoyable and efficient. Instead of running the system on a song, receiving a `.out` file and searching through the database to look for that image, the web app could display the images right away to the user.

Furthermore, if a frontend web app is created, utilising participants' opinions to improve the system is also possible. If the participants were asked to rate each song-image-pair, this could be used to create a better mapping. For instance, if the system had categorised a song into Q1 and hence displayed Q1-images to the user, the feedback can be used to see if Q1 was a good match or not. If all the participants testing the same song rated the song-image-pair with high values, the system has successfully selected a correct quadrant for that song. The images in that quadrant might also have been categorised well.

Lastly, a full web app with a user-friendly frontend and a database in the backend could make a fun tool. The results for each song could be stored in the backend to remove the need to run the system on the same song multiple times. The system could use its own data to categorise new songs by continuously learning from the existing data.

Bibliography

Luís Aleixo, H Sofia Pinto, and Nuno Correia. From music to image a computational creativity approach. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pages 379–395. Springer, 2021.

Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392, 2017.

Josef Bajada and Francesco Borg Bonello. Real-time EEG-based Emotion Recognition using Discrete Wavelet Transforms on Full and Reduced Channel Signals. *CoRR*, 2021.

Tarek R Besold, Marco Schorlemmer, and Alan Smaill. *Computational Creativity Research: Towards Creative Machines*, volume 7 of *Atlantis Thinking Machines*. Atlantis Press, 2015.

Eliza Bliss-Moreau, Lisa A Williams, and Anthony C Santistevan. The immutability of valence and arousal in the foundation of emotion. *Emotion*, 20(6):993, 2020.

Margaret A. Boden. Creativity and artificial intelligence. *Artificial Intelligence*, 103: 347–356, August 1998.

Chris J. Boyatzis and Reenu Varghese. Children's emotional associations with colors. *The Journal of Genetic Psychology*, 155:77–85, 1994.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, and Kailong Chen. XGBoost: eXtreme Gradient Boosting. *R package version 0.4-2*, 1 (4):1–4, 2015.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

Simon Colton, Geraint A Wiggins, et al. Computational creativity: The final frontier? In *20th European Conference on Artificial Intelligence*, volume 20, pages 21–26. Montpellier, 2012.

Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114 (38):E7900–E7909, 2017.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35:53–65, 2018.

Jaap JA Denissen, Ligaya Butalid, Lars Penke, and Marcel AG Van Aken. The effects of weather on daily mood: a multilevel approach. *Emotion*, 8(5):662, 2008.

Edel Ennis and Chris Mcconville. Personality traits associated with seasonal disturbances in mood and behavior. *Current Psychology*, 22(4):326–338, 2004.

Jacek Grekow. Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, 57(3):531–546, 2021.

David Hand and Peter Christen. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539–547, 2018.

Michael Hemphill. A note on adults' color–emotion associations. *The Journal of Genetic Psychology*, 157:275–280, 1996.

Marcus JH Huibers, L Esther de Graaf, Frenk PML Peeters, and Arnoud Arntz. Does the weather make us sad? Meteorological determinants of mood and depression in the general population. *Psychiatry research*, 180(2-3):143–146, 2010.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, Berkeley, 2017. IEEE.

Sherri Melrose. Seasonal affective disorder: an overview of assessment and treatment approaches. *Depression research and treatment*, 2015.

Saif M. Mohammad and Svetlana Kiritchenko. WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 1225 – 1238, Miyazaki, Japan, May 7-12, 2018 2018.

Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, 11(4):614–626, 2018.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Miho Saito. Comparative studies on color preference in Japan and other Asian regions, with special emphasis on the preference for white. *Color Research & Application*, 21 (1):35–49, 1996.

Ann Sarno. Mark Rothko: A Cross-Modal Approach. *Elements*, 2(1), 2006.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

David Watson. *Mood and temperament*. Guilford Press, New York, 2000.

Xinyu Yang, Yizhuo Dong, and Juan Li. Review of data features-based music emotion recognition methods. *Multimedia systems*, 24(4):365–389, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, Berkeley, 2017. IEEE.

Viktor Zorić. Computer generated art based on musical input. Master's thesis, Norwegian University of Science and Technology, Department of Computer Science, 2017.

Viktor Zorić and Björn Gambäck. The Image Artist: Computer Generated Art Based on Musical Input. In *In Proceedings of the 9th International Conference on Computational Creativity*, pages 296–303, Salamanca, Spain, 2018. Association for Computational Creativity.

A. TheAudiDB response

Below are two examples of the JSON output from TheAudioDB API. Example 1 contains a lot of metadata and provides valuable information. Example 2 has very little data, which is hard to work with. The fields “strGenre”, “strMood”, and “strStyle” are essential fields concerning this project.

A.1. Example 1


```
{"track": [
    {
        "idTrack": "32802707",
        "idAlbum": "2116679",
        "idArtist": "111238",
        "idLyric": "401344",
        "idIMVDB": "0",
        "strTrack": "Bohemian Rhapsody",
        "strAlbum": "A Night at the Opera",
        "strArtist": "Queen",
        "strArtistAlternate": null,
        "intCD": null,
        "intDuration": "355106",
        "strGenre": "Rock",
        "strMood": "Rousing",
        "strStyle": "Rock/Pop",
        "strTheme": "...",
        "strDescriptionEN": "\"Bohemian Rhapsody\" is a song by
            the British rock band Queen. It was written by
           Freddie Mercury for the band's 1975 album A Night at
            the Opera. The song has no chorus, instead
           consisting of several sections: a ballad segment
           ending with a guitar solo, an operatic passage, and
           a hard rock section. At the time, it was the most
           expensive single ever made and it remains one of the
            most elaborate recordings in popular music history
```

```
        ."
      "strDescriptionDE": null,
      "strDescriptionFR": null,
      "strDescriptionCN": null,
      "strDescriptionIT": null,
      "strDescriptionJP": null,
      "strDescriptionRU": null,
      "strDescriptionES": null,
      "strDescriptionPT": null,
      "strDescriptionSE": null,
      "strDescriptionNL": null,
      "strDescriptionHU": null,
      "strDescriptionNO": null,
      "strDescriptionIL": null,
      "strDescriptionPL": null,
      "strTrackThumb": "https://www.theaudiodb.com/images/
        media/track/thumb/trvqxt1582885477.jpg",
      "strTrack3DCase": null,
      "strTrackLyrics": "",
      "strMusicVid": "https://www.youtube.com/watch?v=
        fJ9rUzIMcZQ",
      "strMusicVidDirector": "Bruce Gowers",
      "strMusicVidCompany": "Hollywood records",
      "strMusicVidScreen1": "https://www.theaudiodb.com/
        images/media/track/mvidscreen/vxquvq1582885556.jpg",
      "strMusicVidScreen2": "https://www.theaudiodb.com/
        images/media/track/mvidscreen/tsytup1582885566.jpg",
      "strMusicVidScreen3": "https://www.theaudiodb.com/
        images/media/track/mvidscreen/vpyyuw1582885574.jpg",
      "intMusicVidViews": "754654610",
      "intMusicVidLikes": "4295822",
      "intMusicVidDislikes": "151879",
      "intMusicVidFavorites": "0",
      "intMusicVidComments": "345587",
      "intTrackNumber": "11",
      "intLoved": "0",
      "intScore": "9.5",
      "intScoreVotes": "8",
      "intTotalListeners": "1399221",
      "intTotalPlays": "9723231",
      "strMusicBrainzID": "ebf79ba5-085e-48d2-9eb8-2
        d992fbf0f6d",
```

```
        "strMusicBrainzAlbumID": "6b47c9a0-b9e1-3df9-a5e8-50
            a6ce0dbdbd",
        "strMusicBrainzArtistID": "0383dadf-2a4e-4d10-a46a-
            e9e041da8eb3",
        "strLocked": "Unlocked"
        }
    ]
}
```


A.2. Example 2


```
{
  "track": [
    {
      "idTrack": "35570951",
      "idAlbum": "2353024",
      "idArtist": "113672",
      "idLyric": null,
      "idIMVDB": null,
      "strTrack": "Plastic Hearts",
      "strAlbum": "Plastic Hearts",
      "strArtist": "Miley Cyrus",
      "strArtistAlternate": null,
      "intCD": null,
      "intDuration": "205724",
      "strGenre": null,
      "strMood": null,
      "strStyle": null,
      "strTheme": null,
      "strDescriptionEN": null,
      "strDescriptionDE": null,
      "strDescriptionFR": null,
      "strDescriptionCN": null,
      "strDescriptionIT": null,
      "strDescriptionJP": null,
      "strDescriptionRU": null,
      "strDescriptionES": null,
      "strDescriptionPT": null,
      "strDescriptionSE": null,
      "strDescriptionNL": null,
      "strDescriptionHU": null,
```

```
      "strDescriptionNO": null,
      "strDescriptionIL": null,
      "strDescriptionPL": null,
      "strTrackThumb": null,
      "strTrack3DCase": null,
      "strTrackLyrics": null,
      "strMusicVid": null,
      "strMusicVidDirector": null,
      "strMusicVidCompany": null,
      "strMusicVidScreen1": null,
      "strMusicVidScreen2": null,
      "strMusicVidScreen3": null,
      "intMusicVidViews": null,
      "intMusicVidLikes": null,
      "intMusicVidDislikes": null,
      "intMusicVidFavorites": null,
      "intMusicVidComments": null,
      "intTrackNumber": "2",
      "intLoved": "0",
      "intScore": null,
      "intScoreVotes": null,
      "intTotalListeners": null,
      "intTotalPlays": null,
      "strMusicBrainzID": "40d992ca-efe8-4610-99bd-071d7248a9bf
          ",
      "strMusicBrainzAlbumID": "0663bcbd-e202-4aeb-b003-6
          d49f0a4d152",
      "strMusicBrainzArtistID": "7e9bd05a-117f-4cce-87bc-
          e011527a8b18",
      "strLocked": "Unlocked"
    }
  ]
}
```

B. All Moods and Genres in Panda et al.'s dataset

Table B.1.: Emotions in Panda et al.'s dataset

Acerbic	Aggressive	Agreeable
Amiable/Good-Natured	Angry	Angst-Ridden
Anguished/Distraught	Atmospheric	Austere
Autumnal	Belligerent	Bitter
Bittersweet	Bleak	Boisterous
Brash	Brassy	Bravado
Bright	Brittle	Brooding
Calm/Peaceful	Campy	Carefree
Cathartic	Celebratory	Cerebral
Cheerful	Circular	Clinical
Cold	Complex	Confident
Confrontational	Crunchy	Cynical/Sarcastic
Delicate	Detached	Difficult
Dramatic	Dreamy	Druggy
Earnest	Earthy	Eccentric
Eerie	Effervescent	Elaborate
Elegant	Energetic	Enigmatic
Epic	Erotic	Ethereal
Euphoric	Exciting	Explosive
Exuberant	Fierce	Fiery
Flowing	Fractured	Freewheeling
Fun	Gentle	Giddy
Gleeful	Gloomy	Greasy
Gritty	Gutsy	Happy
Harsh	Hedonistic	Hostile
Humorous	Hungry	Hypnotic
Indulgent	Innocent	Insular
Intense	Intimate	Ironic
Irreverent	Jovial	Joyous
Knotty	Laid-Back/Mellow	Lazy
Light	Literate	Lively

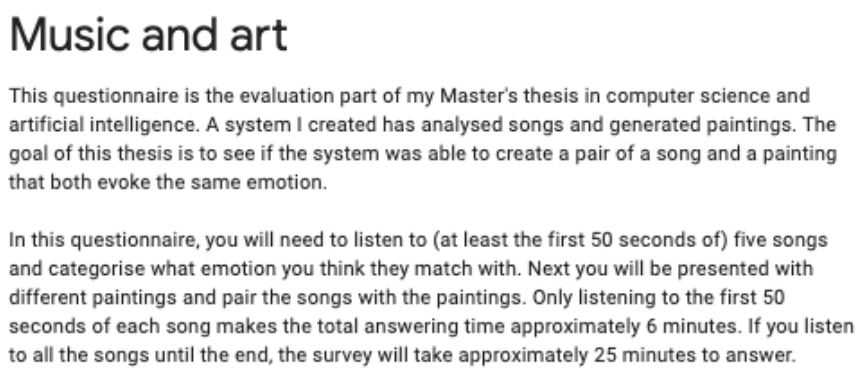
Lonely	Lush	Majestic
Malevolent	Manic	Marching
Meandering	Meditative	Melancholy
Menacing	Messy	Mighty
Mystical	Naive	Negative
Nervous/Jittery	Nihilistic	Nocturnal
Nostalgic	Ominous	Optimistic
Organic	Outraged	Outrageous
Paranoid	Passionate	Pastoral
Plaintive	Playful	Poignant
Positive	Powerful	Precious
Provocative	Pulsing	Pure
Quirky	Rambunctious	Ramshackle
Raucous	Rebellious	Reckless
Refined	Reflective	Regretful
Relaxed	Reserved	Restrained
Reverent	Rollicking	Romantic
Rousing	Rowdy	Rustic
Sad	Sardonic	Searching
Self-Conscious	Sensual	Sentimental
Serious	Sexual	Sexy
Silly	Sleazy	Slick
Smooth	Snide	Soft/Quiet
Somber	Soothing	Sophisticated
Spacey	Sparkling	Sparse
Spicy	Spiritual	Spooky
Sprawling	Springlike	Stately
Street-Smart	Striding	Strong
Stylish	Suffocating	Sugary
Summery	Swaggering	Sweet
Swinging	Technical	Tender
Tense/Anxious	Theatrical	Thoughtful
Threatening	Thrilling	Thuggish
Tragic	Trashy	Trippy
Uncompromising	Unsettling	Uplifting
Urgent	Visceral	Volatile
Warm	Weary	Whimsical
Wintry	Wistful	Witty
Wry	Yearning	

Table B.2.: Genres in Panda et al.'s dataset

Avant-Garde	Blues	Children's
Classical	Comedy/Spoken	Country
Easy Listening	Electronic	Folk
Holiday	International	Jazz
Latin	New Age	Pop/Rock
R&B	Rap	Reggae
Religious	Stage & Screen	Vocal

C. All Survey Questions

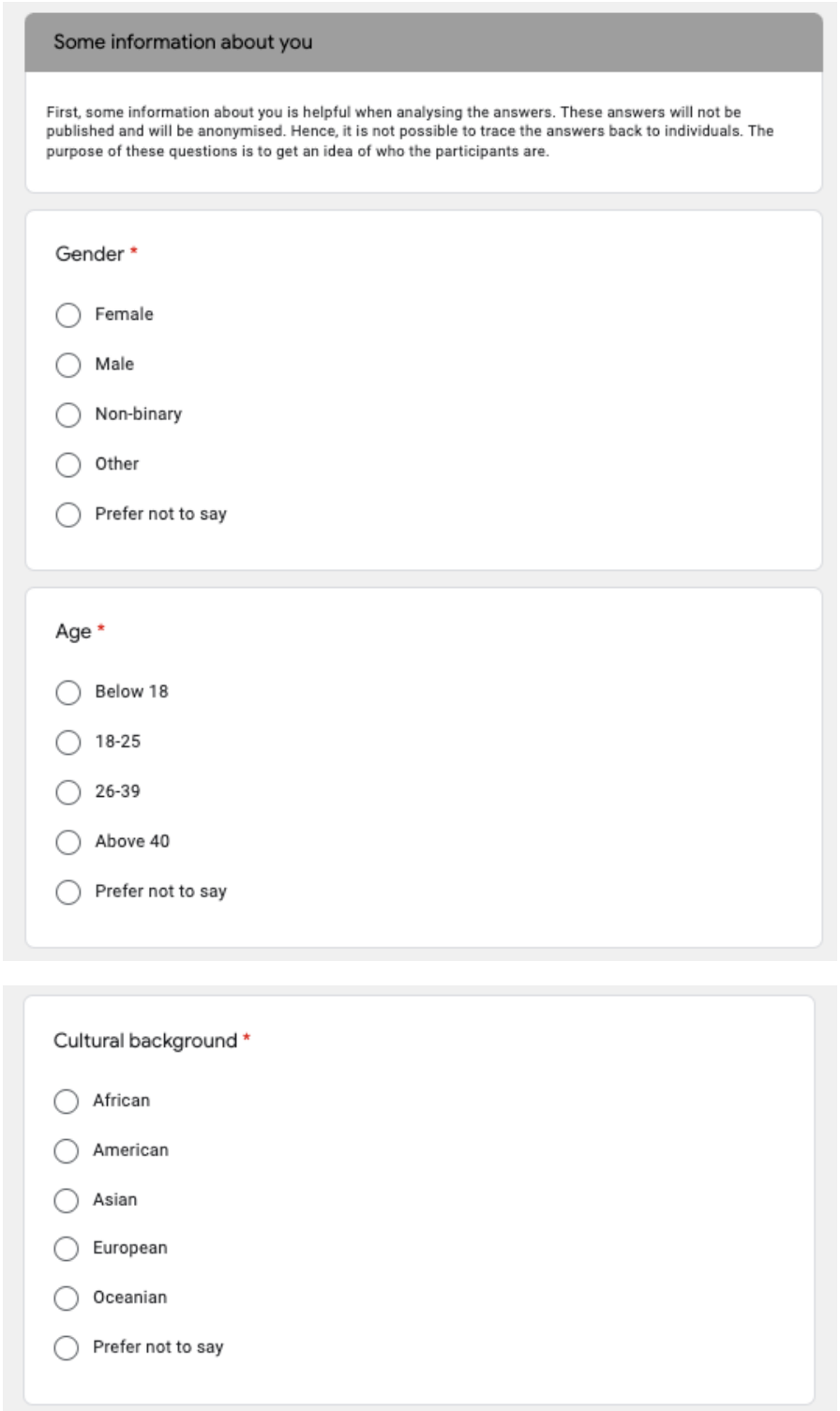
All the questions and information boxes are shown below in the same order as the survey.



Music and art

This questionnaire is the evaluation part of my Master's thesis in computer science and artificial intelligence. A system I created has analysed songs and generated paintings. The goal of this thesis is to see if the system was able to create a pair of a song and a painting that both evoke the same emotion.

In this questionnaire, you will need to listen to (at least the first 50 seconds of) five songs and categorise what emotion you think they match with. Next you will be presented with different paintings and pair the songs with the paintings. Only listening to the first 50 seconds of each song makes the total answering time approximately 6 minutes. If you listen to all the songs until the end, the survey will take approximately 25 minutes to answer.



Some information about you

First, some information about you is helpful when analysing the answers. These answers will not be published and will be anonymised. Hence, it is not possible to trace the answers back to individuals. The purpose of these questions is to get an idea of who the participants are.

Gender *

- Female
- Male
- Non-binary
- Other
- Prefer not to say

Age *

- Below 18
- 18-25
- 26-39
- Above 40
- Prefer not to say

Cultural background *

- African
- American
- Asian
- European
- Oceanian
- Prefer not to say

On a scale of 1 to 7, how much musical knowledge do you have? *

	1	2	3	4	5	6	7	
I barely know what music is and never listen to it	◯	◯	◯	◯	◯	◯	◯	I have a degree in/am currently studying music

On a scale of 1 to 7, how much knowledge do you have about art? *

	1	2	3	4	5	6	7	
I never go to museums and do not enjoy art	◯	◯	◯	◯	◯	◯	◯	I have a degree in/am currently studying art

Thriller - Michael Jackson

Listen to at least the first 50 seconds of the song Thriller by Michael Jackson (single version).
https://open.spotify.com/track/20efeySIfZoiSaISGLBbNs?si=c32bfa75c2114478

Looking at the figure below, in what quadrant would you place the song? *

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ None of them

Which image do you think matches the mood of the song best? *

A

B

C

D

On a scale of 1 to 7, how well do you think the painting you chose above matches the song? *

	1	2	3	4	5	6	7	
Bad match, but better than the other paintings	◯	◯	◯	◯	◯	◯	◯	Excellent match, they both evoked the same emotion.

What sort of painting would you expect or prefer to see with this song? Select all you see fit

- [ ] The one I chose
- [ ] With people
- [ ] Without people
- [ ] With dark colours
- [ ] With light colours
- [ ] With warm toned colours
- [ ] With cool toned colours
- [ ] With winter landscapes
- [ ] With summer landscapes
- [ ] With autumn landscapes
- [ ] With spring landscapes

Dangerously in Love - Beyoncé

Listen to at least the first 50 seconds of the song Dangerously in Love by Beyoncé.
https://open.spotify.com/track/75n8NNxozHYy7THEzzpWtX?si=b7bf3664ad2a4400

Looking at the figure below, in what quadrant would you place the song? *

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ None of them

Which image do you think matches the mood of the song best? *

A

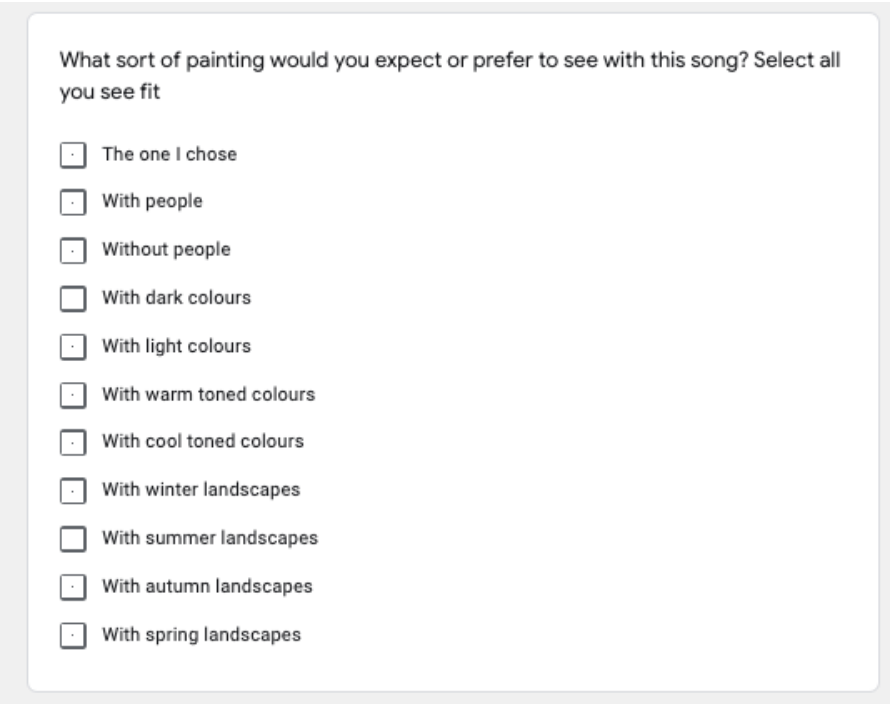
B

C

D

On a scale of 1 to 7, how well do you think the painting you chose above matches the song? *

	1	2	3	4	5	6	7	
Bad match, but better than the other paintings	◯	◯	◯	◯	◯	◯	◯	Excellent match, they both evoked the same emotion.


What sort of painting would you expect or prefer to see with this song? Select all you see fit
The one I chose
With people
Without people
With dark colours
With light colours
With warm toned colours
With cool toned colours
With winter landscapes
With summer landscapes
With autumn landscapes
With spring landscapes

Hurt - Johnny Cash

Listen to at least the first 50 seconds of the song Hurt by Johnny Cash.
https://open.spotify.com/track/28cnXtME493VX9NOw9clUh?si=c500aba20a6d4c68

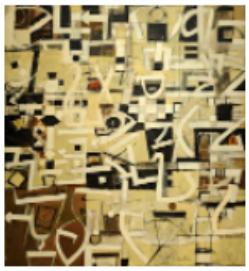
Looking at the figure below, in what quadrant would you place the song? *

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ None of them

Which image do you think matches the mood of the song best? *

A

B



C

D

On a scale of 1 to 7, how well do you think the painting you chose above matches the song? *

	1	2	3	4	5	6	7	
Bad match, but better than the other paintings	◯	◯	◯	◯	◯	◯	◯	Excellent match, they both evoked the same emotion.

What sort of painting would you expect or prefer to see with this song? Select all you see fit

- [ ] The one I chose
- [ ] With people
- [ ] Without people
- [ ] With dark colours
- [ ] With light colours
- [ ] With warm toned colours
- [ ] With cool toned colours
- [ ] With winter landscapes
- [ ] With summer landscapes
- [ ] With autumn landscapes
- [ ] With spring landscapes

The Way I Am - Eminem

Listen to at least the first 50 seconds of the song The Way I Am by Eminem.
https://open.spotify.com/track/23wfXwnsPZYe5A1xXRHb3J?si=a5ef09d2cd0f4831

Looking at the figure below, in what quadrant would you place the song? *

- 1
- 2
- 3
- 4
- None of them

Which image do you think matches the mood of the song best? *

A

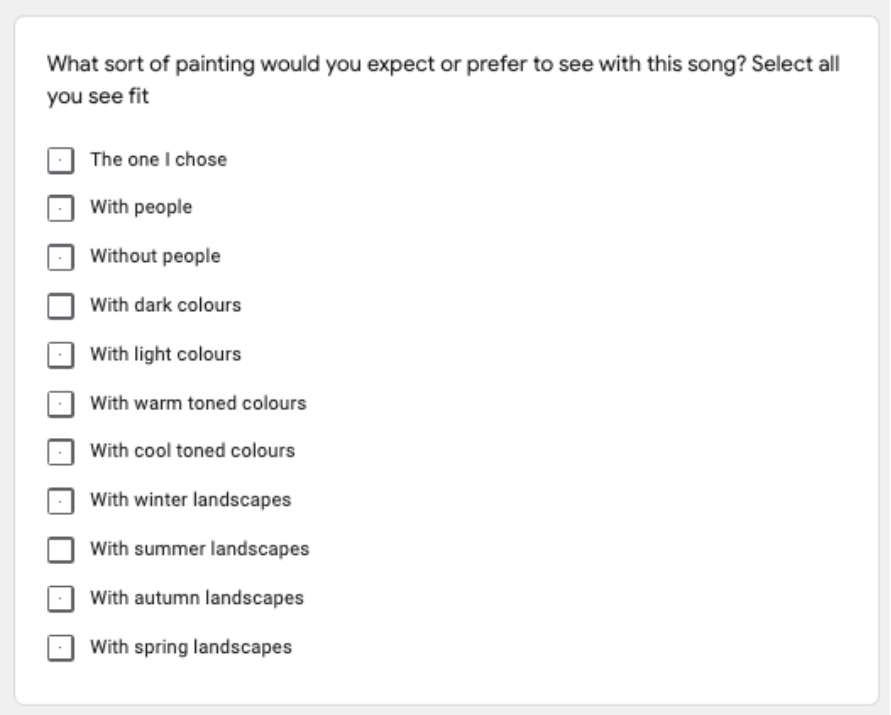
B

C

D

On a scale of 1 to 7, how well do you think the painting you chose above matches the song? *

	1	2	3	4	5	6	7	
Bad match, but better than the other paintings	◯	◯	◯	◯	◯	◯	◯	Excellent match, they both evoked the same emotion.


What sort of painting would you expect or prefer to see with this song? Select all you see fit
The one I chose
With people
Without people
With dark colours
With light colours
With warm toned colours
With cool toned colours
With winter landscapes
With summer landscapes
With autumn landscapes
With spring landscapes

Rehab - Amy Winehouse

Listen to at least the first 50 seconds of the song Rehab by Amy Winehouse.
https://open.spotify.com/track/3N4DI1vuTSX1tz7fa2NQZw?si=c8327e88876048e9

Looking at the figure below, in what quadrant would you place the song? *

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ None of them

Which image do you think matches the mood of the song best? *

A

B

C

D

On a scale of 1 to 7, how well do you think the painting you chose above matches the song? *

	1	2	3	4	5	6	7	
Bad match, but better than the other paintings	○	○	○	○	○	○	○	Excellent match, they both evoked the same emotion.

What sort of painting would you expect or prefer to see with this song? Select all you see fit

- [ ] The one I chose
- [ ] With people
- [ ] Without people
- [ ] With dark colours
- [ ] With light colours
- [ ] With warm toned colours
- [ ] With cool toned colours
- [ ] With winter landscapes
- [ ] With summer landscapes
- [ ] With autumn landscapes
- [ ] With spring landscapes

D. Survey Answers

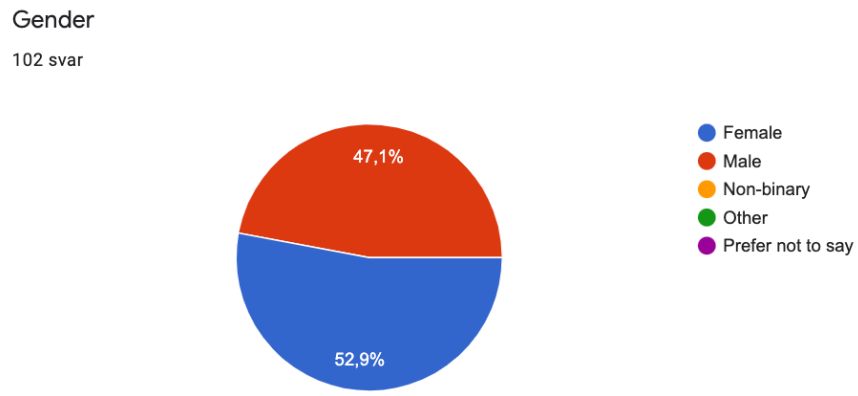


Figure D.1.: Survey results: Participants' gender

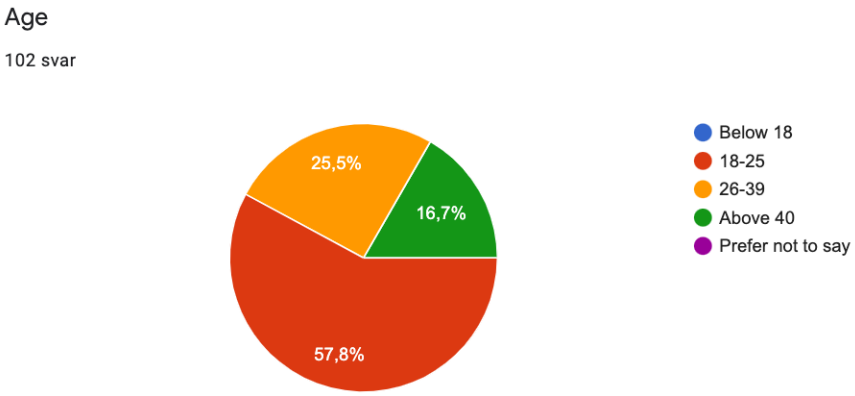


Figure D.2.: Survey results: Participants' age

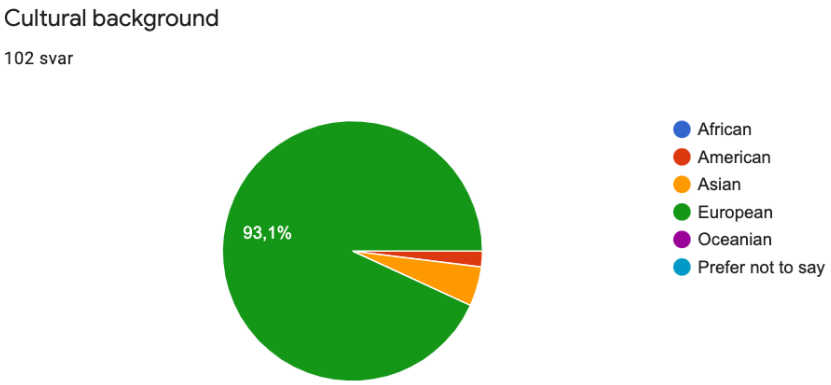


Figure D.3.: Survey results: Participants' cultural background

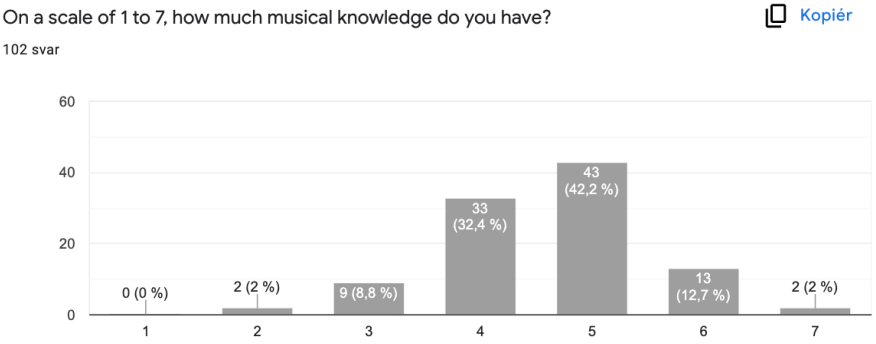


Figure D.4.: Survey results: Participants' musical knowledge

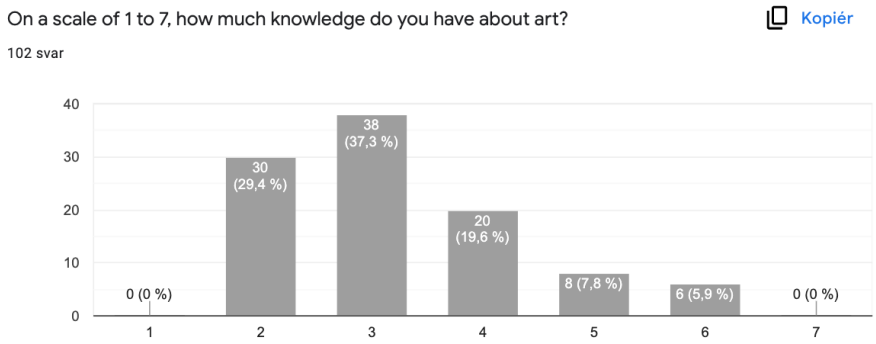


Figure D.5.: Survey results: Participants' artistic knowledge

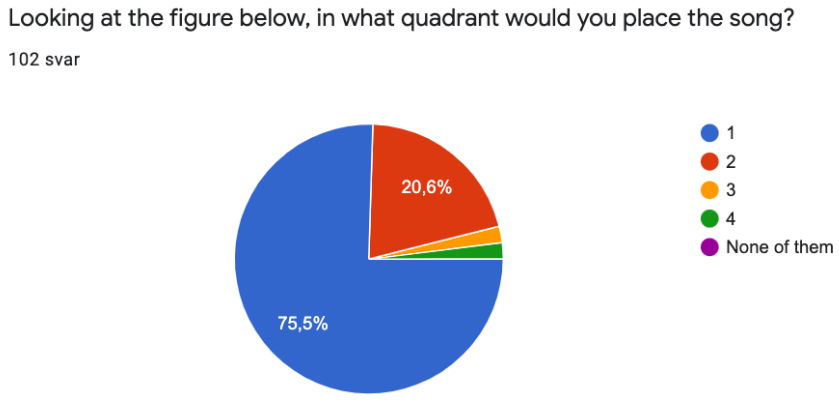


Figure D.6.: Survey results Thriller: Quadrant

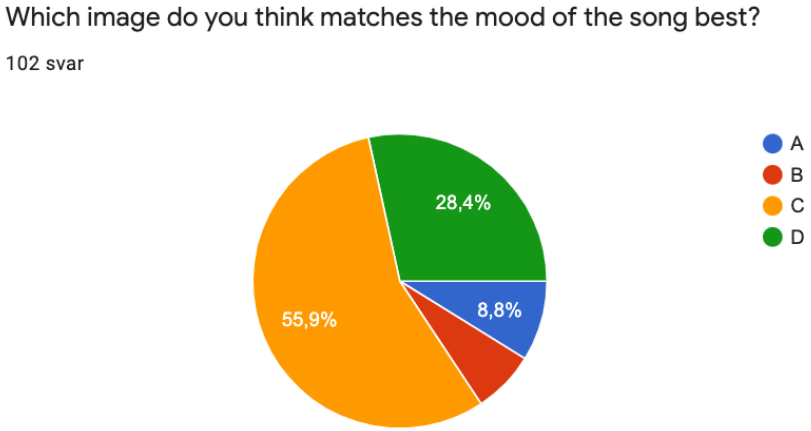


Figure D.7.: Survey results Thriller: Image

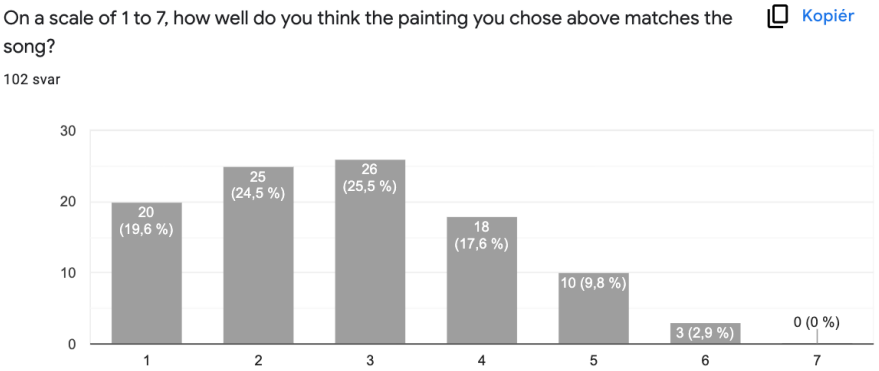


Figure D.8.: Survey results Thriller: Image match rating

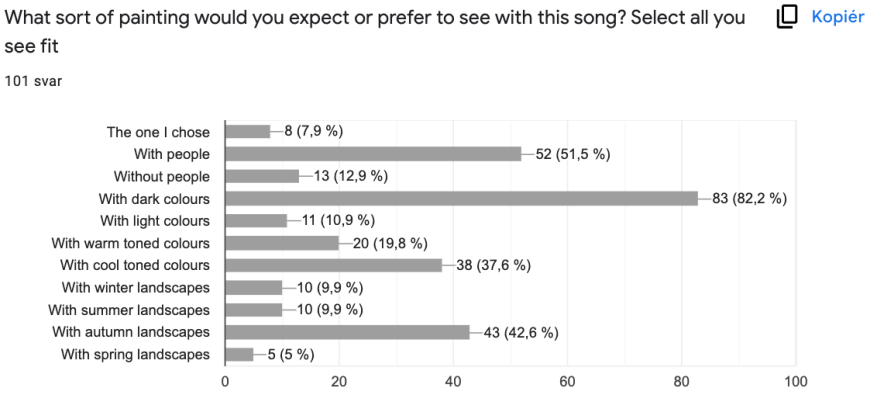


Figure D.9.: Survey results Thriller: Image preferences

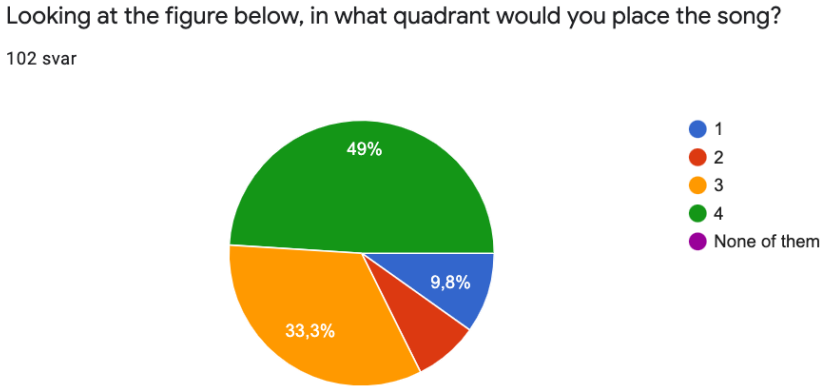


Figure D.10.: Survey results Dangerously in Love: Quadrant

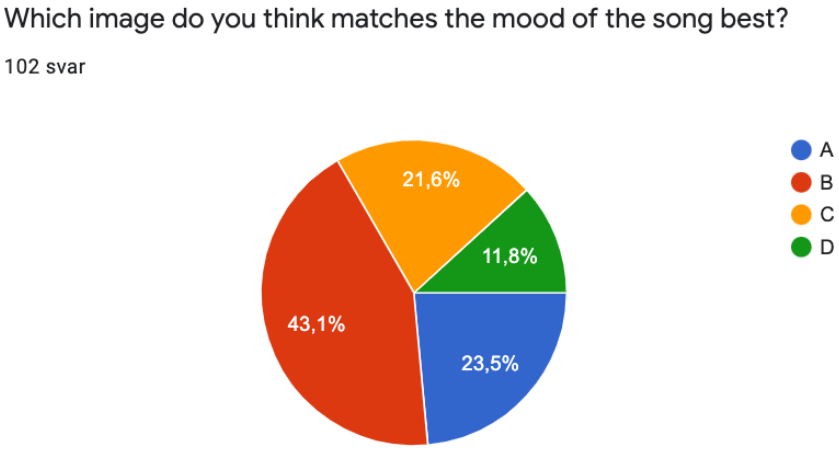


Figure D.11.: Survey results Dangerously in Love: Image

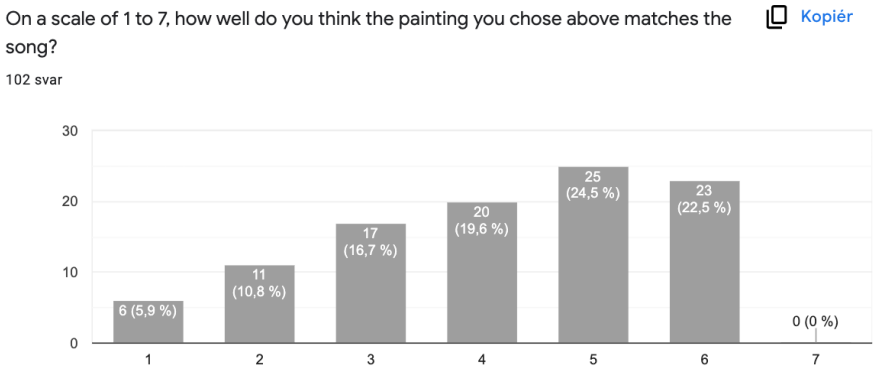


Figure D.12.: Survey results Dangerously in Love: Image match rating

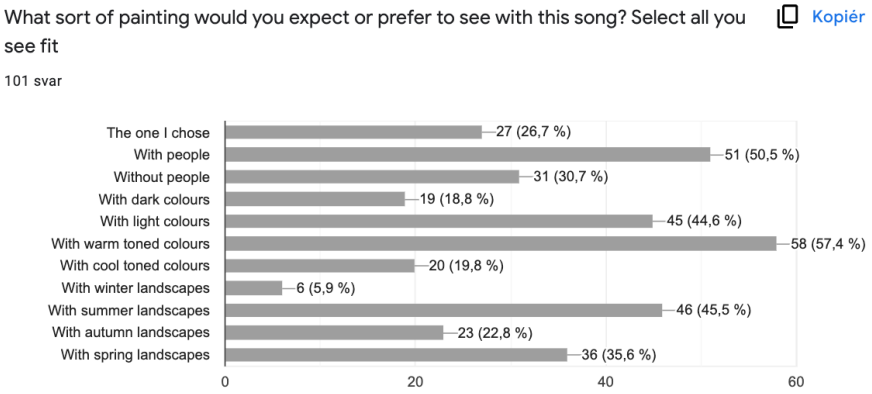


Figure D.13.: Survey results Dangerously in Love: Image preferences

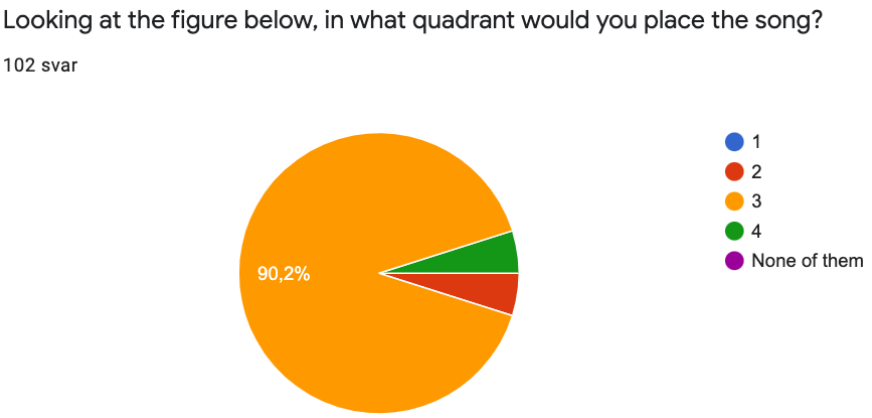


Figure D.14.: Survey results Hurt: Quadrant

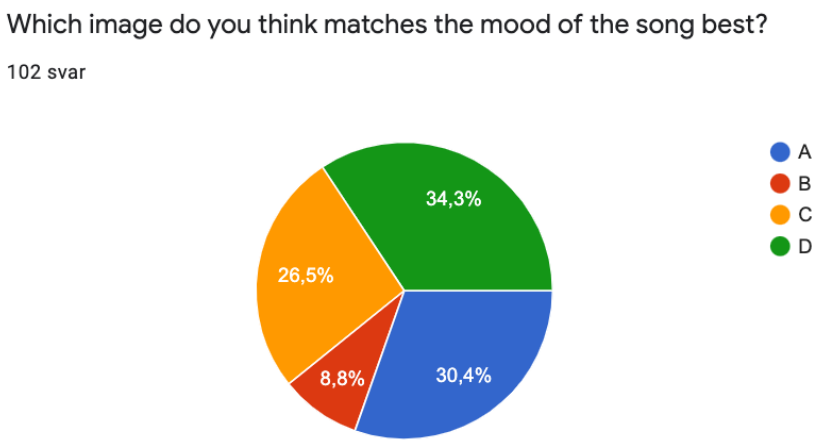


Figure D.15.: Survey results Hurt: Image

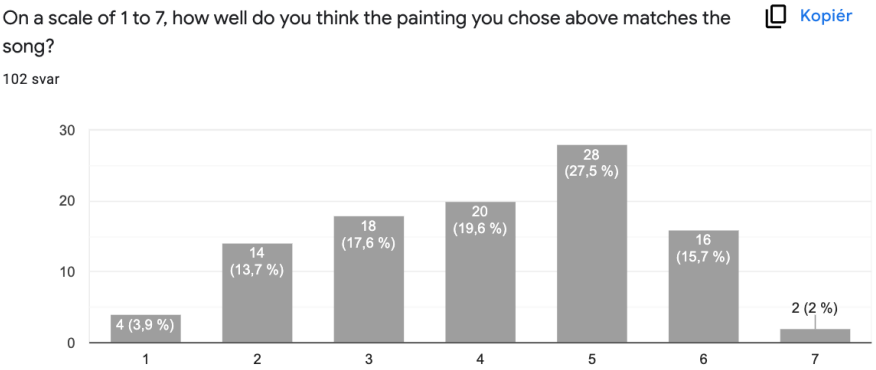


Figure D.16.: Survey results Hurt: Image match rating

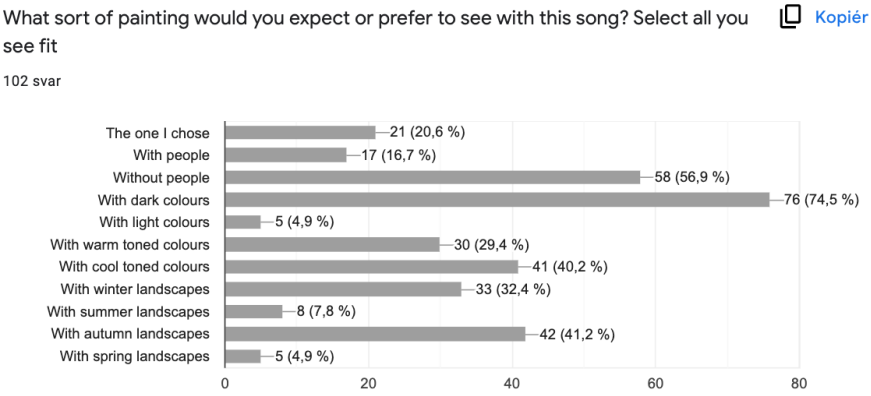


Figure D.17.: Survey results Hurt: Image preferences

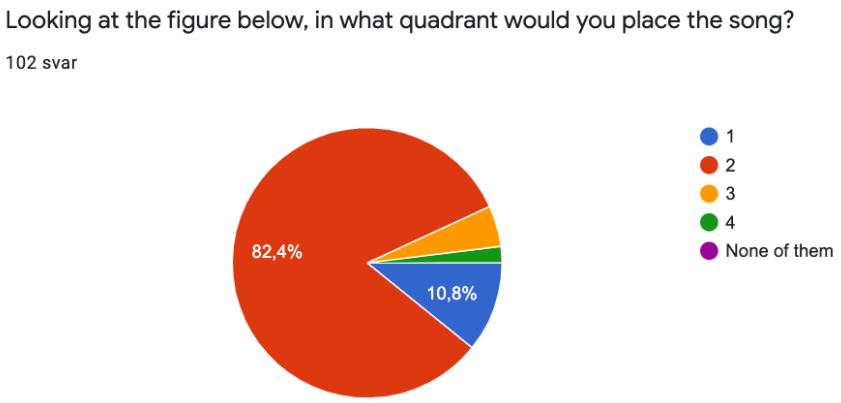


Figure D.18.: Survey results The Way I Am: Quadrant

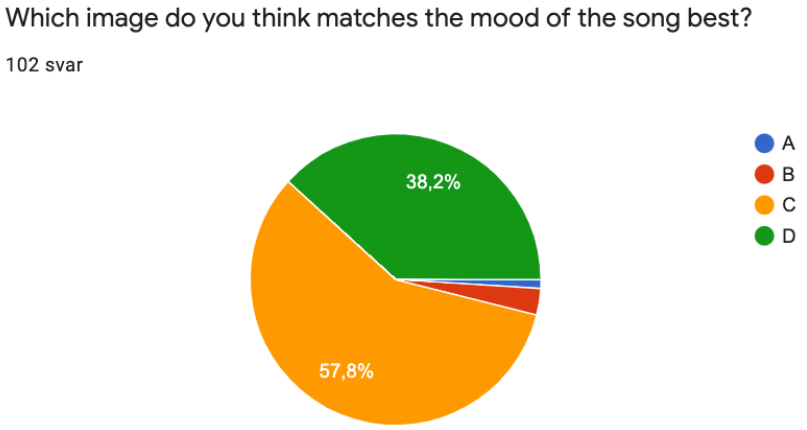


Figure D.19.: Survey results The Way I Am: Image

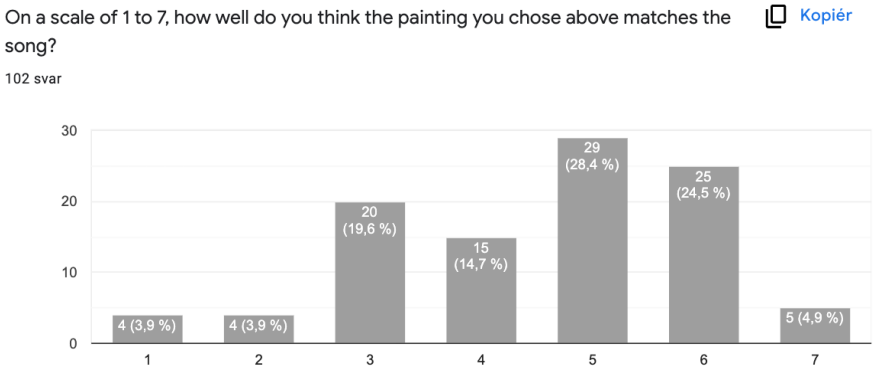


Figure D.20.: Survey results The Way I Am: Image match rating

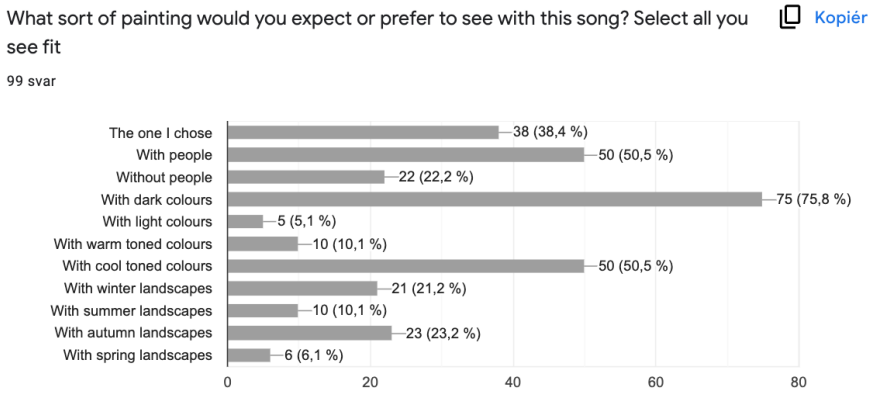


Figure D.21.: Survey results The Way I Am: Image preferences

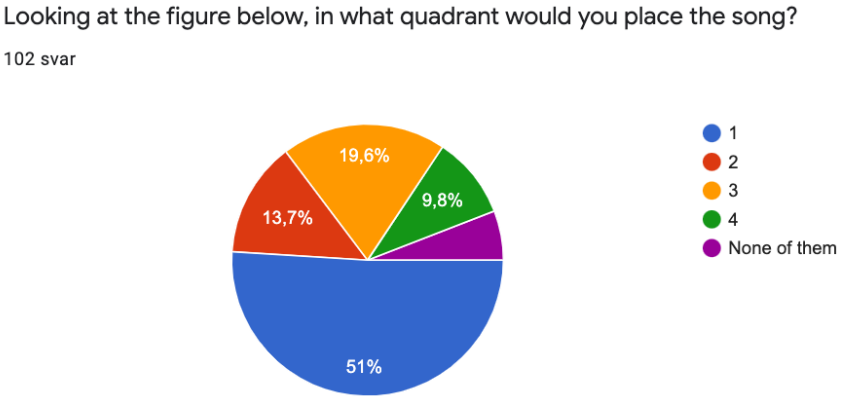


Figure D.22.: Survey results Rehab: Quadrant

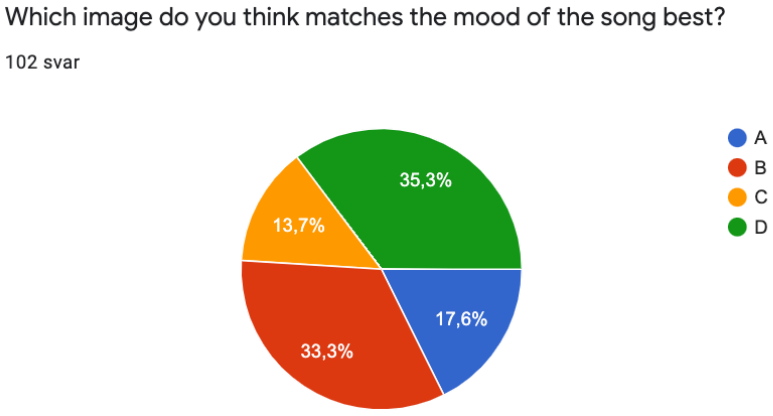


Figure D.23.: Survey results Rehab: Image

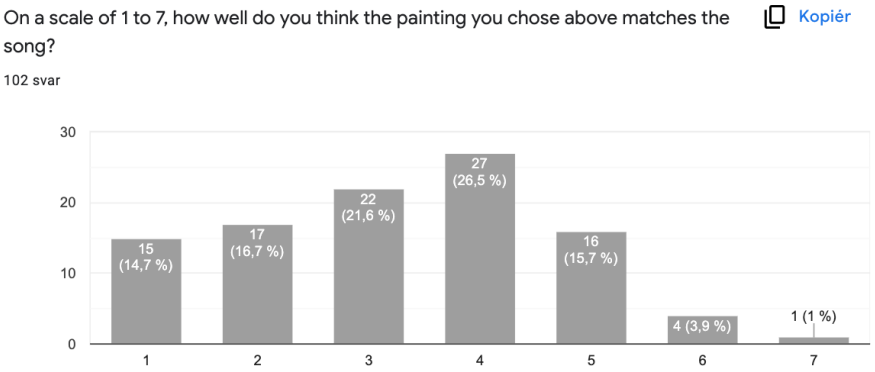


Figure D.24.: Survey results Rehab: Image match rating

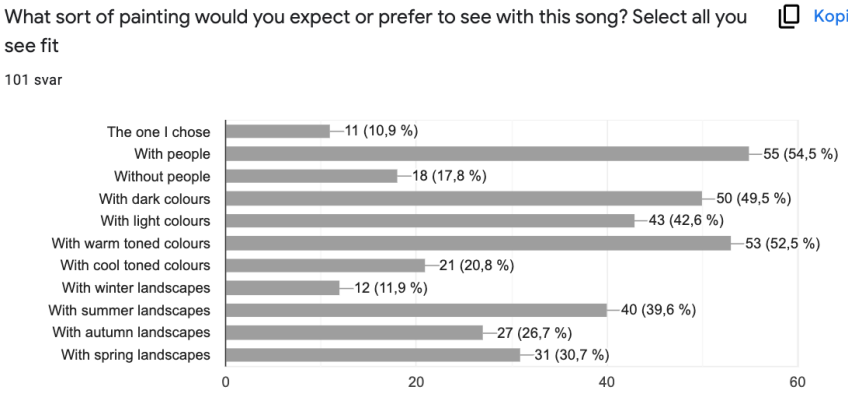


Figure D.25.: Survey results Rehab: Image preferences

Hedda Mathilde Sæther Langvik

Uniting Music and Painted Art Using Emotion Categories and Metadata

NTNU
Kunnskap for en bedre verden