NATIONAL UNIVERSITY OF IRELAND, GALWAY

DOCTORAL THESIS

Art and Analogy Through Evolutionary Computation

Author:
Aidan Breen

Supervisor:
Dr. Colm O'Riordan

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Computational Intelligence Research Group
College of Engineering and Informatics

March 28, 2019

Contents

Declaration of Authorship

I, Aidan Breen, declare that this thesis titled,
"Art and Analogy Through Evolutionary Computation"
and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the National University of Ireland, Galway (NUIG).
- No part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution.
- No professional editorial services have been employed in the production of this thesis in accordance with NUIG policy.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- The work in this thesis is not part of a group project.
- Where the content within this thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

Art and Analogy Through Evolutionary Computation

by Aidan Breen

In this thesis the development of a computer system capable of generating art through the process of analogy is presented. A visual display is generated analogous to a given piece of music. This analogy is based upon the aesthetics of both music and visuals whereby an aesthetic piece of music should prompt an aesthetic visual display with similar characteristics.

The aesthetics of both music and visuals which form the core of the proposed system remain difficult to capture. A novel approach is presented, designed to capture and evaluate aesthetic experience for use in computational systems using a graph based ranking approach. A number of experiments were conducted to analyse the performance of this approach, and a number of studies were conducted using this approach to gather aesthetic data.

Representing an analogy within a computer system requires codifying the connections between two domains. Mapping expressions are introduced to represent these connections. Together with the aesthetic data gathered, these mapping expressions are evolved using a computational evolution approach. The output of this process is an abstract representation of an analogy. This representation is then used to create a visual display.

Further studies have been carried out to evaluate the effect of these generated visuals on the aesthetic experience of a viewer.

Acknowledgements

I wish to thank Dr. Josephine Griffith, Dr. James Cruickshank and Dr. Conn Mulvihill for their support, guidance and precious time over the last number of years as my Graduate Research Committee. A special thanks to Dr. Colm O'Riordan, my supervisor for his insight, expertise and guidance throughout my research, and for originally encouraging my pursuit of a doctorate.

This work could not have been completed without the hard work of all the staff in the Department of Information Technology, with special thanks to Tina Earls and Mary Hardiman.

Finally, to my family and friends who have encouraged and supported me throughout my studies, thank you.

To my beautiful bride-to-be.

Chapter 1

Introduction

Creativity is part of how we define ourselves as human. It is a phenomenon that has occurred naturally with all the intricate complexities one would expect a biological system to hold. While procedural systems — capable of churning out infinite variations of repeated patterns — may be created with relative ease, a truly creative computer system, following the same reasoning and generative processes as the human mind, is still beyond the reach of even the most advanced Artificial Intelligence systems.

Analogy has long been regarded as a core process of human communication, problem solving and indeed creativity. In combination with computational intelligence approaches such as evolutionary computation, it may lead the way to computational creativity. The research presented in this thesis explores this possibility and asks: *Can a computationally intelligent system be created which is capable of making useful analogies between separate domains such as music and visuals?*

The following paragraphs introduce the work conducted in answering this question, with formal research question and hypotheses defined in Section 1.2.

The first step in answering this question is to define what exactly is meant by analogy. In general the process of analogy involves making comparisons between otherwise unrelated domains. In the literary sense, analogy can be a strong descriptive tool but it can also be an effective method of communication and problem solving. To take a relevant example, a musician may be required to play a piece of music loudly, often by use of the sheet music dynamic symbol ***f***, or *Forte*. It is also common for a visual art piece to be described as loud — often referring to bold contrasts, bright colours and striking patterns — even when loudness is an aural phenomenon. This example of linguistic analogy suggests that there is some relationship between a loud piece of music, and a loud visual, even though the individual art forms are fundamentally different.

In this work, an aesthetic analogy is adopted such that the identified domains should be compared, or connected, by some attribute that affects the aesthetic appeal of objects in either domain. Aesthetic analogy was selected due to the strong role that aesthetic appeal plays in the creative process, especially in artistic domains.

The use of analogy requires the identification of specific domains to be compared. The domains of music and visuals are a natural choice. There is a clear connection between the two domains, with music often performed with an accompanying visual display such as a music video, stage lighting or some generated visual display. Approaches to pair music and visual scores have been numerous, ranging from abstract art with musical motifs or musical scores to accompany film, to physical devices to allow the performance of a light show similar to how a music piece would be performed. Indeed the combination of music and visuals permeates modern culture with visual displays at musical concerts, music videos accompanying recorded songs, and music visualisation software built in to media players on most home computers. The connection between the domains is undeniably strong, yet the physical attributes that convey this connection remain elusive.

The potential application of machine learning and Artificial Intelligence in live performance of music is a further motivating factor. Currently, AI techniques are revolutionising many industries with notable recent applications being self-driving cars [3, 167], natural language translation [164] and imaging [40, 98]. While these applications are obviously pushing the boundaries of technology in their respective domains, there have been relatively few applications of Artificial Intelligence used as creative tools, aiding the user with direct human interaction.

The particular computational intelligence techniques employed were selected based on a number of criteria. The approach would have to be suitable to efficiently traverse a vast search space, while allowing the flexible representation and analysis of a given analogical structure. Further, the approach should allow the evaluation and execution of analogies online — or in real time — using reasonably provisioned hardware, which will likely require the search and generation of analogies to be conducted offline. Evolutionary computation was therefore selected as a suitable candidate.

Evolutionary computation aims to mimic the natural process of evolution and the Darwinian theory of Survival of the Fittest. Individuals represent algorithms or digital artefacts. These individuals are represented digitally with chromosomes which describe the appearance of the artefact or the operation of the algorithm. Groups of these individuals form a population and the fitness of each individual is calculated by the use of some fitness measure. The chromosomes of the fittest individuals are then spliced together in the *crossover* process. This process is similar to the mating of two or more individuals, producing the next generation by combining their chromosomes. In this way, the process aims to take fit individuals and produce a fitter offspring, with all the best traits of the parents. The process repeats until an individual of a suitable fitness is discovered.

The process of evolutionary computation has been adopted in many varied domains with success. Recent applications include the study of financial loopholes [180], real-time environment control [163], robotic control [11], game theory [125], and many other tasks.

The process of evolutionary computation was particularly suited to the challenge proposed in this thesis. Creativity and analogy at their very core are the results of an evolutionary process. While computational evolution may be a relatively weak facsimile of the natural phenomenon, it is the most likely computational model to produce the desired outcomes in a similar fashion.

In this work, an analogical structure is proposed which enables the comparison of domains using a mapping expression. The structure of expressions can be formalised and codified by the use of a grammar. Using a specific application of evolutionary computation called Grammatical Evolution, we can use the defined grammar to evolve analogical mapping expressions. By the use of this approach, we can evolve testable analogies and explore their effectiveness as a creative tool.

The application of the work presented in this thesis is not restricted to the creation of creative analogies. The creation of effective analogies may have separate applications in the study of human cognition. As a logical reasoning skill, computational analogy may have applications in the creation of an Artificial General Intelligence. The specific musical and visual analogies created may have further applications in User Interface design whereby interfaces may be enhanced by visual or audio aids. Subliminal enhancements could ease fatigue and enhance cognition speed.

Supporting work including the Graphsort algorithm may have applications in psychological studies where subjective feedback is required. The Graphsort algorithm is also useful in certain sorting or ranking applications where the comparisons of sortable objects are not reliable.

1.1 Open Questions

The work in this thesis is oriented towards computational intelligence, however, many of the motivating questions behind this work come from other areas of research.

1.1.1 Aesthetics

Neuropsychology focuses on behaviour, emotion and cognition in relation to brain function. Within this domain there is particular interest in human subjective responses, the responses of an individual when presented with some stimuli. The obvious challenge here is in understanding the mechanisms behind a particular response, as well as the variation in responses between people. This interest extends to aesthetic response where an art piece can be enjoyable to some and detested by others. Often, it is not clear why some people respond negatively to an art piece and there is insight to be gained in understanding this phenomenon.

Open Question 1 If the generation of an art piece can be controlled, can an aesthetic response be affected in a predictable way?

Open Question 2 Are there some factors that have stronger effect on aesthetic response than others? Is it possible to use a computational system to test this?

1.1.2 Analogy

Analogy is of interest in many domains including cognitive psychology, philosophy, logic, linguistics, artificial intelligence and machine learning. The use of analogy is known to be valuable and an understanding of the mechanics of analogy may be useful in the creation of a true artificial general intelligence. While much has been written on the topic, the specifics of this mechanism remains an open question. The work in this thesis aims to explore this area with specific focus on the following question:

Open Question 3 Can a computationally intelligent system be used to create useful analogies?

1.1.3 Synesthesia

Synesthesia is a condition whereby certain signals within the brain are intermixed, often combining the senses and even providing a sort of physical sense of abstract concepts. Common symptoms include the ability to *see* sounds and music, to *taste* numbers, or to *see* the days of the week as a colour.

In many ways, this condition encapsulates the goals of the work in this thesis. Indeed, the simple existence of the condition reinforces the role of analogy within the brain. As humans, we naturally make connections between concepts and senses and in the case of synthesthesia, this is more than a conceptual mechanism, it is a literal experience. Many accounts by those with synesthesia report the ability in a positive sense, enhancing the experience of music and being useful as a mnemonic. This leads to the question:

Open Question 4 If the combination of senses is useful and positive in the case of synesthesia, can a computational system be designed to create a positive experience by building analogies between artistic domains?

1.1.4 Artificial Intelligence, Machine Learning and Creativity

Recent advances in Artificial Intelligence have gathered a great deal of public attention with systems which have learned to play games without any hard coded rules [156]. While examples of Machine Learning applied to games like Checkers can be found as early as the 1950's [152], only recently have such systems been capable of challenging top tier players in games that are thought to require a certain level of creativity, such as Go. These games are considered a more difficult challenge than classic example of Checkers and Chess, as top tier players reportedly often rely on gut instinct rather than the computation of future moves. However, the most powerful systems capable of these feats are based on deep neural networks, with a very clearly defined fitness metric: to win the game. The learning process involves a system playing hundreds of millions of games against itself to reinforce its behaviour — a function only possible with a concrete fitness metric. And while a system may be capable of winning in what is considered by humans to be a creative task, the processes that occur within the system are essentially a black box, revealing no insight into the process used.

Open Question 5 While current computational systems may be capable of performing well in what we consider a creative challenge, we gain no insight into their processes. Is it possible to use a creative computational system that provides insight into the creative process?

Introducing human interactivity into computational intelligence systems has been achieved in many ways. Supervised machine learning — perhaps the most popular approach — involves a great deal of human input in the gathering, structuring and verifying training data. In a great deal of applications however, this data is prohibitively expensive to gather or may simply be unreliable or worse, incorrect. For artistic and creative work, it can be even more difficult to gather data for supervised systems as responses are often subjective and vary rapidly and unpredictably. Interactive Divergent Evolution has been employed to attempt to use human feedback in a less restrictive manner, but this approach relies heavily on a single human participant to remain focused and interested in the task at hand.

Open Question 6 Supervised and interactive machine learning techniques are successful but the cost of gathering training data and of interacting with the system is a major pitfall. Can a system be designed that minimises human interactions and thus human fatigue while remaining robust to unreliable input and training data?

1.2 Research Questions and Hypotheses

The research questions and hypotheses are formally identified and defined in this section. The primary research question is first outlined, followed by subquestions which serve to answer the primary question in part. Hypotheses are defined for each subquestion.

1.2.1 Primary Research Question

Can a computationally intelligent system be created which is capable of making useful aesthetic analogies between separate domains such as music and visuals?
This is the overall goal of the work presented in this thesis. The research question contains two main areas of interest: computational analogy in the context of the aesthetics of music and visuals, and the usefulness, affect, or impact of the created analogy. The following subquestions and hypotheses address these individual concerns.

1.2.2 Sub Question 1

Can a computationally intelligent system be built to create aesthetic analogies?
This subquestion addresses the creation of an aesthetic analogy using computational intelligence. The hypotheses related to this subquestion are focused on the practical implementation of aesthetic computational analogy, including the design of the analogy system, the evaluation or execution of the system and the gathering of data to be used within the system.

H1 Aesthetic data can be gathered in a way that minimises participant effort.

H2 Aesthetic attributes in the chosen domains of music and visuals can be objectively measured.

H3 Aesthetic data can be gathered in the chosen domains.

H4 An analogical structure can be created using the gathered aesthetic data.

H5 The analogical structure can be evaluated in real-time.

1.2.3 Sub Question 2

Does the analogy generated affect aesthetic response?
This subquestion aims to verify that a generated aesthetic analogy will impact the aesthetic response of humans viewing the artistic material.

H6 The use of aesthetic analogy has an impact on the enjoyment of an observer in artistic material

H7 The use of aesthetic analogy has an impact on the interest of an observer in artistic material.

H8 The use of aesthetic analogy has an impact on the fatigue of an observer while viewing artistic material.

1.3 Methodology

Most of the proposed hypotheses can be tested using an empirical approach. However, certain underlying hypotheses, such as **H2:** *"Aesthetic attributes in the chosen domains of music and visuals can be objectively measured"* cannot be addressed solely by experimentation. In the case of **H2**, previous foundational literature was used to first build evidence in support of the hypothesis. The evidence gathered was then tested by conducting an experiment and independently gathering data to verify the reproducibility of reported results. **H2** and **H3** were tested using this approach.

In the case of purely empirical work, we first introduce the content and methodology followed by the results and discussion providing a detailed analysis of the results obtained. Contributions that rely on foundational literature will present relevant references and previous results. In the some cases, where there are many individual results, the presentation of results and the discussion are presented together for the benefit of the reader.

A high level description of the approach taken throughout this research is now presented.

The use of aesthetic analogies necessitates the use of human interaction due to the strongly subjective nature of aesthetics and in particular the aesthetics of music and visuals. This challenge is made more difficult by the strong effect of fatigue on aesthetic responses. Any data that is to be collected must be collected using the absolute minimum number of participant responses. The nature of human feedback of this type is also subject to unreliable responses. Asking a participant their opinion of the same piece of music or image twice may provide two conflicting answers.

To begin tackling this challenge, the Graphsort algorithm was designed. Using a graph based ranking algorithm, subjective responses are gathered by asking the minimum number of questions to obtain a partial ranking. The output of the algorithm is a ranking of all objects for each participant. These individual responses are then combined to find an average ranking across a population. The successful implementation of the Graphsort algorithm allows the collection of aesthetic data in the domains of music and visuals. Details of the Graphsort algorithm are presented in Chapter 3

As supporting work, the Graphsort algorithm was compared to a number of other ranking and sorting algorithms. The focus of this study concerned the performance of each algorithm when comparisons between objects was not reliable. This unreliability is called noise and can take a number of forms, known as noise patterns. The performance of each algorithm was measured in two ways.

Firstly, the number of comparisons made by each algorithm before it reached a sorted, or partially sorted result were measured. This value gave in indication of the evaluation time of each algorithm and was particularly important when evaluating the performance of Graphsort, where the absolute minimum number of comparisons is favourable.

Secondly, the accuracy of the output was measured as the similarity of the output list to a fully sorted list by use of Spearman's Footrule algorithm. This value indicated how well each particular algorithm is capable of ranking elements, given the various noise patterns. The value was less important when evaluating the performance of Graphsort, but did provide insights into the expected accuracy of the algorithm.

The analogy structure requires an attribute by which the two domains can be compared. Harmony was selected as the first attribute to be tested as it has a strong effect on the aesthetic appeal of both music, in the form of musical consonance, and visuals, in the form of colour harmony.

Within the musical domain, musical harmony was measured using the Graphsort algorithm. A small group of participants were asked to rank note pairs in terms of musical harmony. The survey produced results strikingly similar to previous studies in the area. This work was published in the Sound and Music Computing Conference 2015 [20] and is discussed in detail in Chapter 4.

Using the same approach, similar data was gathered within the visual domain with participants ranking colour pairs in terms of visual harmony. The colours presented were varied only in hue, and not saturation or lightness. Interestingly, the data gathered did not agree with previous studies which varied hue alongside saturation and value. Our analysis of this result can be found in Chapter 4.

The next challenge was to show that the data gathered could be used together with the proposed analogy structure to create analogies. Following the iterative nature of human creativity, Evolutionary Computation was adopted as a machine learning approach. The analogy structure proposed uses a set of Mapping Expressions to map an attribute value in one domain to an attribute value in another domain. These Mapping Expressions may take the form of a mathematical expression, which in turn can be represented trivially by a grammar. Following this, Grammatical Evolution was adopted which uses an Evolutionary Computation approach with a predefined grammar to evolve expressions.

The data from each domain was used, along with a defined grammar and a fitness approach which favoured accurate mappings of values from one domain to the other. This approach proved successful with analogical mappings been evolved over a number of generations to a high fitness. This work was published in Evolutionary Computation Theory and Applications 2016 [22] and is described in detail in Chapter 5.

One of the desired outcomes of this process was to create a system capable of real time operation. In this respect, the evolved analogical mappings could be generated at any time, but should be evaluated in real time - converting values from one domain to another with little or no delay. The expressions produced by the system were tested by measuring the time taken to read live musical input, preprocess it to fit the expected data structure, evaluate the expression and pass the output to a visualisation server and update a visual display. The results suggest that this can be done within realtime thresholds.

The final step is to evaluate the actual impact of the analogical mapping on human viewers. A study was conducted using the visualisation system described above to generate realtime visual displays from a piece of music. A number of mapping expressions were used of various fitness to produce visual outputs. Study participants were presented with the music and accompanying visuals and asked to give feedback on their enjoyment and interest. The study found that there was no significant difference in performance between any mapping expressions indicating a negative result.

1.4 Contributions

This thesis presents a number of individual contributions that were required to validate the primary research question. These contributions have been reported in various publications as detailed below:

1.4.1 Design and Analysis of Graphsort

The Graphsort algorithm was devised to gather subjective feedback and to form an objective model based on unreliable input. The design and implementation of the Graphsort algorithm is described in detail in Chapter 3. The algorithm was initially presented in:

- Aidan Breen and Colm O'Riordan. "Capturing and Ranking Perspectives on the Consonance and Dissonance of Dyads". In: *Sound and Music Computing Conference*. Maynooth, 2015, pp. 125–132

Analysis of the performance of the Graphsort algorithm, together with a comparison of the performance of the algorithm to several other ranking and sorting algorithms, is also presented in Chapter 3. This work has been reported in:

- Aidan Breen and Colm O'Riordan. "Capturing Data in the Presence of Noise for Artificial Intelligence Systems". In: *Irish Conference on Artificial Intelligence and Cognitive Science*. Dublin, 2016, pp. 204–216

1.4.2 Aesthetic Data

In Chapter 4 the gathering of aesthetic data using the Graphsort Algorithm is presented.Two separate studies are presented which gather musical and visual data respectively. The data gathered is used in further work and serves to confirm previously reported results. Part of this work is previously reported in:

- Aidan Breen and Colm O'Riordan. "Capturing and Ranking Perspectives on the Consonance and Dissonance of Dyads". In: *Sound and Music Computing Conference*. Maynooth, 2015, pp. 125–132

1.4.3 Analogical System: Mapping Expressions

In Chapter 5 the design of an analogical system is outlined. The proposed system defines an analogical structure composed of mapping expressions which connect separate domains. The domains of music and visuals are used as a motivating example. Mapping expressions are generated using Grammatical Evolution together with a fitness function which makes use of the data gathered using the Graphsort algorithm. This work has been reported in:

- Aidan Breen and Colm O Riordan. "Evolving Art Using Aesthetic Analogies: Evolutionary Supervised Learning to Generate Art with Grammatical Evolution". In: *8th International Conference on Evolutionary Computation, Theory and Applications*. Porto, Portugal, 2016, pp. 59–68

1.4.4 Analogical System: Realtime Analogies

The analogy system outlined in Chapter 5 has been implemented, demonstrating that mapping expressions can be generated using Grammatical Evolution and that the analogical structure can be used with a live audio input to generate real time visual displays. The work has been reported in:

- Aidan Breen, Colm O Riordan, and Jerome Sheahan. “Evolved Aesthetic Analogies to Improve Artistic Experience”. In: *6th International Conference on Computational Intelligence in Music, Sound, Art and Design.* Amsterdam, 2017

1.5 Thesis Structure

The following chapters of this theses are organised as follows. Chapter 2 provides a detailed literature review of relevant related research. Chapter 3 introduces the ranking of subjective preferences and presents the details and performance of the Graphsort algorithm. Chapter 4 presents the results of two studies conducted using the Graphsort algorithm to gather aesthetic data. Chapter 5 introduces the analogical system which evolves mapping expressions based on the data gathered in previous chapters. Finally, Chapter 6 concludes the thesis with a summary of work, and further discussion including a revisit of the contributions made, the limitations of the work and a brief discussion on potential future work.

Chapter 2

Background

In this chapter, the relevant foundational areas of research are introduced, beginning with a brief discussion on the challenges of creativity and artificial intelligence in Section 2.1, before introducing the field of Computational Analogy in Section 2.2. This is followed by an introductory discussion of aesthetics in Section 2.3, in particular the aesthetics of music and visuals (Sections 2.3.2 and 2.3.3 respectively), with which the work in this thesis is concerned. The gathering of data regarding the aesthetics of music and visuals is introduced separately in Section 2.4. Existing and historical examples of visual displays based upon musical sources are presented in Section 2.5. Finally, the artificial intelligence and machine learning techniques used in this work are introduced in Section 2.6 before the chapter is summarised in Section 2.7.

2.1 Introduction

Many advances in Artificial intelligence have been made in recent years. We have seen computer systems master games like Checkers [152], Chess [121] and Go [156], appear human by passing the Turing test [183] and perhaps even more impressively, compete and win a trivia game show on live television [52].

While these advances are clearly exceptional feats of engineering and computer science, there are always new challenges to face in the field. One outstanding challenge that remains far from any clear solution is creativity. As human beings we value our creativity and those of us who demonstrate an ability to be creative in a particular field are held in high regard. Unfortunately, even the most gifted individuals struggle to explain what exactly it is that drives their creativity. Perhaps this is most evident when that spark is lost, even temporarily. What came so naturally only hours or days before, may now seem like an unsurmountable challenge. The difference between the state of effortless creativity and oppressive rut may be subtle, but is certainly indescribably powerful. This lack of understanding into the creative process makes the challenge of building a creative AI system incredibly difficult.

While a spontaneous creative outburst might not be possible, or even desirable, for an AI system, it might still be possible to make use of the various insights into creativity that we do have. One insight in particular is the use of metaphor and analogy, building a set of connections between otherwise unrelated domains. Throughout history, painters have often been inspired by the beauty of nature, transferring observed colours, patterns and textures to canvas. The resulting work reflects the emotional aspects of the original scene in a new medium. Similarly, musicians and poets have been inspired by personal relationships, using literal metaphor in their work. Even further inter-inspirational events have occurred where visual artists are inspired by a music piece or visa versa. Comedians often make strong use of the mechanics of analogy by setting up two seemingly unrelated domains and connecting them in an unexpected way. A classic example of this technique can be found, appropriately, in a joke told be John Cleese during his 1991 speech, On Creativity:

"The old one about a woman doing a survey on sexual attitudes who stops an airline pilot and asks him, amongst other things, when he last had sexual intercourse. He replies 1958. Now, knowing airline pilots, the researcher is surprised and queries this. Well, says the pilot, it's only 2110 now."

Cleese uses two domains of understanding and builds an unlikely connection between them. In the context of the surveyor and the understanding of the audience, 1958 is clearly referencing a year. The punchline relies on the audiences understanding of the famed promiscuity of airline pilots, but overlooking the use of military time which would be more popular, we suppose, in the parlance of a pilot. When the unexpected connection is made in the punchline, the humour is obvious. A clear use of structural analogy as a creative means.

Clearly the use of analogy and metaphor is an important aspect of creativity. The work in this thesis aims to investigate the use of analogy as a tool which may lead to the creation of an AI system capable of creativity. The achievement of this goal relies upon a great deal of interdisciplinary knowledge and builds upon a large body of work. The following sections in this chapter aim to introduce this work.

2.2 Computational Analogy

Analogy is an important part of creativity, underpinning language, art, music, invention and science [61].

Computational analogy as a research domain has been active since the late 1960s, accelerated in the 1980s and continues today. The aim of computational analogy research has been to gain some insight through computational experimentation into what has been deemed one of the most important and powerful actions our brains perform: analogy making. It is worth noting that in this sense, analogy and metaphor are considered “ubiquitous within our cognitive experience” [69] which, when we consider the importance of metaphor in our understanding of creativity, underlines the importance of computational analogy in research into aesthetics and artificial intelligence.

The subject has roots in a number of areas such as artificial intelligence, philosophy, neuroscience and cognition, and fine art which all make contributions to the area to a certain degree. A subject with as broad a background as computational analogy, taking influence from some seemingly unrelated areas, leads to many different experimental approaches and research directions. However, in the early years of analogy research, Gentner formed a view “that analogy entailed finding a structural alignment, or mapping between domains,” [61] an observation that has become central to almost every computational approach to analogy making today.

There are many different ways to classify computational analogy systems however the most common classification is based on the underlying architecture of the system. The three major categories of systems are:

- **Symbolic systems** which follow certain common symbolic AI principals such as the use of logic, symbols, planning, means-ends analysis, and search heuristics.

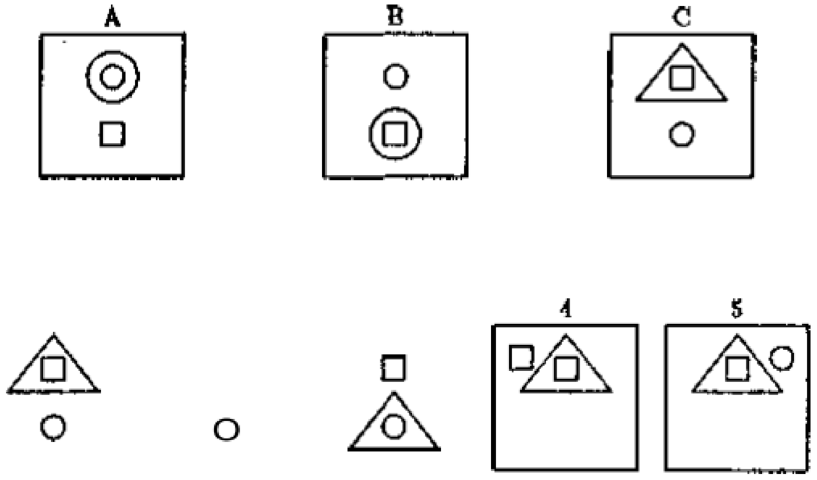


FIGURE 2.1: **A is to B as C is to...?** An analogy based question typical of standardised tests of the 1960s.

- **Connectionist systems** which attempt to build a network between frameworks or domains (concepts or particular 'things' we want to make analogies between) using nodes, weights and spreading activation, back propagation etc.
- **Hybrid systems** which fall somewhere between the two and tend to be larger, agent based systems.

2.2.1 Symbolic systems

Many early attempts in the area of computational analogy making were made in an attempt to demonstrate how a computer system could express intelligence in a way we might normally qualify it in a human. At the time, standardised tests were a common method of gauging an individual's intelligence and many of these tests based their questions on analogical reasoning. Researchers therefore turned to analogy as a means of programmatically solving these problems, making use of the logical and structural techniques of which computers at the time were capable of utilising. The following systems all fall within the category of symbolic systems.

One of the earliest systems built with this challenge in mind was Evans' ANALOGY system [46], designed to solve geometric analogy problems. The system solved problems similar to that shown in Figure 2.1, where three example figures were shown: A, B and C. An analogy is assumed between A and B such that one or more 'rules' were followed which turned A into B. An answer 1, 2, 3, 4 or 5 is then to be chosen based on the analogy between C and the answer, where the analogy between C and the correct answer most closely resembles the analogy between A and B.

Evans' system works by comparing objects within the figures. An object, in this sense, is any particular shape with a given location. The system keeps track of what objects are deleted, added, or mapped between figures. The list of mapped objects also kept track of euclidean transformations. ANALOGY preferred simple over complex euclidean transformations which lead to some seemingly obvious solutions being overlooked due to diagonally moved objects.

The system correctly answered the problem shown in Figure 2.1, and is described as having performance "roughly comparable to median performance in a population of high school students." [69].

The next major development in computational analogy making was introduced by Becker's JCM system [9]. Becker developed the notion of *working memory* and *long term memory* which allowed systems to store large amounts of information which may or may not be relevant to the current situation in *long term memory*, and extract pertinent information to *working memory* to solve the task at hand. In this way, systems could begin to "learn" or at least build up a knowledge base without drastically decreasing performance. This paradigm would prove to be integral to many future systems and indeed in the design of computer systems in general as we know them today.

Where Evans' approach was practical, limiting itself to a very specific problem set, Richard Brown's 1977 Ph.D. thesis entitled "Use of analogy to achieve new expertise"[26], explored a much more theoretical and abstract approach which could be applied in a variety of circumstances. Brown's approach, as outlined in a later review, described a system of reductions "where a hard problem in one domain is reduced to a familiar and more easily solved problem in another domain" [69]. Brown's aim was to transfer problem solving "expertise" across domains that did not map directly to each other.

"Expertise" in Brown's view, was any knowledge that made a problem easier and consisted of a “domain description written as axioms in first-order logic, problem solving plans that determine predicates or functions in the description, and LISP programs that carry out portions of plans.”[69] Brown’s actual analogy making process consisted of three steps: "Map, Solve and Lift"[26].

First, the system maps a target difficult problem into a source problem in which we have some expertise. Next, it finds a solution to the source problem. Finally, it then maps the solution and its justification back into the target domain for evaluation. This approach has a major benefit in that the process of solving a problem and creating an analogy generates a significant knowledge base about the target problem. In essence, the system can learn more about the target problem by attempting to solve it by analogy.

Around the same time, McDermott published research regarding his ANA system: “an implemented production system that uses analogies when solving problems in a simulated environment” [115]. The simulated environment here is a paint shop in which a user can ask the system to perform tasks like painting, spraying or washing objects that exist within the virtual paint shop. The system will then attempt to carry out those tasks within the environment. The system may not have been explicitly told how to perform the given task but it was capable of making analogies between the given tasks actions and a previously executed tasks actions.

The system works by building up a working memory of goals and sub-goals, as well as preconditions which can verify if the potential solutions are valid. In the case of an error occurring while attempting to decide on a possible action, for example if the system finds it needs more information from the user, or the task the user has asked of the system is deemed impossible given the state of the environment, the system will ask the user to give it more information. The ANA system has been proven to be “effective on a small sample of tasks in a constrained experimental environment.” [69].

In more general applications, case based reasoning refers to a similar process of solving unseen problems by referring to the solutions of similar past problems. MEDIATOR [157] is credited with being the first computational analogy system to officially use case based reasoning as a core function of operation.

A later system, Prodigy/Analogy [177] combined case based reasoning with means-ends analysis in an architecture specifically designed to scale to larger problem sets.

In 1978, transfer frames were introduced by Winston who further developed the concept in 1980 and 1982 [186, 185, 187]. In this approach, a source and a target object were described as similes. The most salient properties of the target were compared to those of the source and mapped accordingly, producing a rule which could be abstracted as an analogy.

The structure and approach of transfer frames anticipated the later work of Dedre Gentner introducing structure mapping theory [60]. Gentner's work in this areas has been hailed as "unquestionably the most influential work to date on the modelling of analogy making" [57].

Structure mapping theory outlines an approach to analogy making where two domains, the target and the base (previously referred to as the source) consist of node like objects and predicate-like relations. Domains are compared by the overlap of their structure, rather than the overlap of particular attributes of the containing objects. There are four major categories of *similarity results* obtainable from this approach, illustrated in the source text using the example of an analogy between batteries and reservoirs:

- **Literal similarity** where two domains must have an "overlap in both object attributes and inter object relationships". This concept is similar to the Object Oriented Programming concept of Object Equivalence. Two domains are *literally similar* in the way that two instances of the same object type are similar. Two separate reservoirs are literally similar as they perform the exact same role of storing water as potential energy, and possess the same attributes (*holds water* for example) even though those attributes may be different such as the volume of water they hold, or the dimensions of the reservoir.

- **Analogical similarity** is present if there is an "overlap in relationships but not object [attributes]". A battery is *analogically similar* to a reservoir as they both store energy and allow it to be used to do work. Yet, a battery and a reservoir have entirely different attributes; a battery stores electricity while a reservoir stores water.

- **Appearance similarity.** "Overlap in object-attributes but not relationships is seen as a mere appearance match." This case would describe two domains appearing similar but holding no practical or functional similarity. For instance, a reservoir would be similar in appearance to a drinking glass as they are both objects that hold water, yet their relationship with how they use the water is distinctly different.
- **Anomaly.** Finally, a comparison where neither object overlap nor relational overlap is present is seen as an anomaly.

Gentner also mentions how the relationship between an analogical similarity and a literal similarity is not a dichotomy, but rather a continuum which allows some form of flexibility in it's analysis.

The structure mapping theory was put into practice in the "structural mapping engine" [49] and has also been extended in many other implementations [48, 62, 56, 51] and later [99].

Finally, there have been a number of different approaches taken more recently that should be acknowledged in order to show the full gamut of research in this area.

The Incremental Analogical Machine (IAM) maps pieces of a base domain to the source incrementally, building up a gradual interpretation based on select components of the domains [88, 89]. The initial research, with IAMs processing being totally serial in nature, was initially criticised due to it's restricted scalability [55]. These criticisms were addressed by the Incremental Structure Mapping Engine (I-SME) which is, in the most part, a combination of SME and IAM, taking the analytical approach of SME together with the incremental architecture of IAM, bringing parallel processing to the original IAM system [55].

Finally, ACT-R is a system developed in order to investigate "a unified theory of metaphor understanding, semantic illusions and text memory" and to model invention by analogy [5, 27]. The system was ambitious and far reaching but failed to ignite any significant discussion.

2.2.2 Connectionist systems

In his review of computational analogy systems, French explains the position of connectionist models within the broader context of computational analogy: “Symbolic systems are generally well equipped to model relational structures involving situations represented as objects and relations between objects. For this reason, these models held the high ground for many years in the computational modelling of analogy-making. However, due largely to recent advances in their representation techniques, connectionist models have taken their place alongside symbolic models of analogy-making”[57].

The connectionist systems described in this section are structurally similar to what we now refer to as neural networks. Many recent advances have been made in the use of neural networks, particularly deep neural networks, as a machine learning toolset with frameworks such as TensorFlow currently in use in various academic and enterprise applications [1]. The recent advances in neural networks are partly due to the ongoing improvements in computational power and distributed computation required to run these systems at scale. For this reason, while connectionist systems showed great potential, early implementations were severely restricted.

The first example of a connectionist model, which used a neural network like system to solve analogies was ACME [74]. The analogy was an emergent result of the states of the network with “hypothesis” nodes which represented possible relations between a base and a source domain. Excitatory and inhibitory nodes within the structure cancelled themselves, or strengthened themselves based on semantic similarity, structural similarity and pragmatic importance. In this way, contradictory nodes would reduce the strength of the analogy while consistent nodes mutually benefited and strengthened the analogy. The system evaluated all possible mappings in parallel and produced an output of the best mapping and the set of strong hypothesis nodes.

LISA [77], a descendant of ACME, which has been described as ”one of the most ambitious connectionist models of analogy making” [57] was a more flexible system built upon more reasonable association constraints such as partially distributed representations of concepts, selective activation and dynamic binding instead of one to one mappings. The system uses a method of binding only nodes which "oscillate in synchrony" which is crucial in allowing both long term memory and working memory to interact during both retrieval and mapping.

The first distributed connectionist model, STAR-1 [68] was designed to solve proportional analogies and made use of the tensor product connectionist models by Smolensky [159]. It has since been followed up by STAR-2 which was designed to investigate the emergence of analogy making capabilities of children [184].

The use of holographic reduced representations, "a method for representing more complex compositional structure in distributed representations" [138] has been put into use by DRAMA [44] which makes use of a fully distributed representation of concepts and includes mechanisms to deal with both semantics and structure of the base and target domains.

Adaptive Resonance Theory [31] was put to use by a system described in 2000 which used synaptic triads to be explicitly appealing in a neurobiological sense [81]. This particular approach and it's focus on strict neuroscience demonstrates the potential application of connectionist systems in combination with previously gathered data.

2.2.3 Hybrid systems

Hybrid models consist of those systems that attempt to make use of both symbolic representations with logical processing, and subsymbolic connectionist properties such as spreading activation, and excitatory and inhibitory nodes in neural network like models. These systems were often agent based and more abstract than connectionist systems. Though these systems typically require more complex architectures, they have not demonstrated much more flexibility in their performance.

Hofstadter proposed a hybrid system which has been utilised in various implementations such as COPYCAT , TABLETOP, Letter Spirit, and METACAT [119, 58, 116]. The proposed system consists of agents, which Hofstadter refers to as codelets, which work in a simulated parallel architecture to build up or break down connections in a simulated world.

Agents work in one of two separate ways. Firstly, some agents work in a "top down" fashion, which is functionally similar to the use of working memory in symbolic systems, using and building problem specific data and structures. Secondly, a second set of agents work simultaneously using a "bottom up" approach, similar to the long term memory of symbolic systems using data from cross domain problem sets and deeper structures. The main benefit of this system is how the agents are able to build up their own representation of the source and target domains as well as the connections between them without any predefined structure, based only upon a small set of predefined rules. The system's simulated parallelism allows partial relations between the domains to influence the outcome of the system as agents continuously influence each other. The lack of true distributed computing and parallelism however was a major constraint.

As the potential benefits of spreading activation within a connectionist system were emerging, Hendler explored the potential hybridisation of the various schools of thought in the area [72, 71]. Primarily, the impacts of the marker-passing model, which focused on the spread of symbolic information throughout a large "associative knowledge representation", and local-connectionism which focused on the spreading of sub-symbolic numerical information through smaller generalised connectionist networks. Hendler concludes that the hybridisation was successful, but only in a very limited implementation. Nevertheless, it began a long running discussion on the hybridisation of various connectionist techniques.

Further work by Lange et al. explores the various challengers and opportunities associated with generalised connectionist systems [101, 100]. Analogical retrieval and short-term memory retrieval are among the potential opportunities investigated with both distributed and localised connectionist system architectures explored and discussed.

DUAL [94, 96], and its successors AMBR [95] and AMBR-2 [137] became early examples of agent based systems making use of fully hybrid architectures. They are similar to Hofstadter's model in that they consist of small agents working in parallel representing small fractions of information or knowledge about the analogy. DUAL places a strong emphasis on the context sensitivity of the system, a trait that was introduced by Hofstadter's codelets. In contrast to other existing hybrid systems at the time, DUAL is described as a "a hybrid system at the micro level rather than at the macro level - it consists of hybrid microagents and so both aspects of the architecture take part in every stage of every cognitive process."

The later versions, AMBR and AMBR-2 extended the system to make use of the emerging parallel processing technologies at the time for recollection, mapping and transfer of information between domains.

Serious research efforts in computational analogy reduced shortly after this time with the more pragmatic and achievable benefits of hybrid connectionist systems were given more attention. Today, we see the remnants of this work in deep neural networks which achieve hybridisation through the use of many layers of sub-symbolic networks, with symbolic stops - often the result of human interaction and tuning - punctuating the structure.

2.3 Aesthetics

We begin this section with an introduction to the philosophical and scientific approaches to aesthetics and human perception. Following this, we discuss the specific aesthetic knowledge bases in the domains of music and visuals.[1]

Aesthetics is a branch of philosophy concerned with the appreciation of beauty. It can be more rigorously defined as a study of sensory and emotional perception. Classical work in aesthetics often involves a critical appreciation and reflection of art, design and nature. However, the concept of aesthetic value — some measure by which an object can be judged as pleasing to the eye — is not restricted solely to fine art and nature. Aesthetics plays a large role in our interaction and perception of everyday objects. In data visualisation, "Empirical studies have shown a correlation between perceived aesthetics and usability" [53]. This effect is only growing stronger as we spend

---

[1] The work in this thesis is concerned with producing a visual display from some musical input. In contrast to the musical input, our visual display is simplified and represented as a pair of stage lights which have a brightness, and colour which can change over time. This simplification reduces the complexity of the output by removing any notion of line, shape and two-dimensional pattern and allows us to investigate the practical applications of the system without being restricted to these common pitfalls of computational art in two dimensions. In a stage performance, a simple lighting setup of this type is common and can produce powerful aesthetic responses. Simplification of the musical input to this degree was not possible.

more and more time in front of computer screens of all sizes, upon which every pixel is analysed and jostled around in search of the optimum User Experience. User Experience is, of course, just another name for the practical aesthetics of a computer interface.

In this respect aesthetics have become less of a philosophical exercise, but rather a set of practical tools with real-world consequences and business applications. Unfortunately, while the importance of aesthetics is being recognised, Ekart et al. reported in 2012 that aesthetics remain "one of the top ten unsolved problems of information visualisation" [43], a fact that remains true today.

There is a "gap that separates the sciences and the humanities - the two cultures" [65] and by extension, there is a gap between the pragmatic aesthetics of User Experience and the romantic aesthetics of philosophy. While both approaches have made countless contributions and furthered the overall understanding of aesthetics, neither approach will prove entirely successful in isolation. Indeed, much of the common knowledge in one domain may have even been discovered as a result of work in another. As Huang notes in a 2009 article on the topic of the neuroscience of art, historically, "Although not always with reproducible experiments or verifiable results, [artists have] intuitively unlocked the secrets of the eye and the visual brain." [76].

This suggests the benefit of an approach to aesthetics which recognises the contributions from both philosophical and scientific domains. Similar to the computational analogy systems mentioned above that make use of both top down (working memory) and bottom up (long term memory) approaches to building understanding of a new problem — it is conceivable that a philosophical approach to understanding aesthetics could compliment a scientific one.

2.3.1 Neuroscience and Neuroaesthetics

The study of neuroscience is concerned with understanding the way the human nervous system — the eyes, the ears, the brain and all the innumerable connections between them — perceives sounds and images. With this understanding of perception comes an understanding of the perception of beauty and aesthetics. More specifically, the field of neuroaesthetics, a sub-field of neuroscience, focuses directly on our nervous system and it's perception of aesthetics.

A number of studies attempting to understand the science behind aesthetics have been carried out with varying success. Perhaps some of the earliest formal work in the area was conducted by G.D. Birkhoff who undertook a programme to capture aesthetics with mathematics, by defining the "aesthetic measure" formula $M = O/C$, the ratio of order (O) to complexity (C) [16, 15]. Birkhoff's work has been often criticised as over-simplistic but it did begin an important discussion on the definition of aesthetics.

More recently, Ramachandran has introduced "Eight laws of artistic experience' — a set of heuristics that artists either consciously or unconsciously deploy to optimally titillate the visual areas of the brain" [143]. This work aims to combine a number of concepts in contemporary neuroaesthetics and provide a framework for artistic aesthetic appreciation. One of Ramachandran's eight laws is that of contrast: the magnitude of difference between features of a domain. Ramachandran defined contrast as a function of aesthetics, but interestingly, this concept has been tackled in a very different way by Sagi, (working with Gentner, mentioned previously in Section 2.2.1) in the context of analogy.

Sagi et al. have recently carried out research on the topic of how we perceive and define the differences between two images. Starting with the rather obvious statement that "detecting that two images are different is faster for highly dissimilar images than for highly similar images." Making the distinction between images faster suggests it is somehow easier for us, and slower would suggest a level of effort needed. "Paradoxically, we showed that the reverse occurs when people are asked to describe how two images differ—that is, to state a difference between two images." [151] The research would suggest that a person presented with two greatly dissimilar images would find them easy to comprehend until they are pressed to further examine the images, leading to more effort than initially encountered.

This observation is not restricted to the contrast of images alone. The same phenomenon has been observed in areas like Typography. In a recent web site article, Mayer discusses the importance of highly contrasting typefaces in the aesthetics of a piece of text. When faced with the task of using more than one typeface, he notes: “keep it exactly the same, or change it a lot — avoid wimpy, incremental variations.” Mayer continues to make his point with an analogy between typefaces and coins. He states that placing two identical coins together will look pleasing due to their similarity, and placing two extremely different coins together will also look pleasing due to how dissimilar they are. However, if an American dime is placed beside a coin from another currency, of a similar size and colour, it creates “an uneasy visual relationship because it poses a question, even if we barely register it in on a conscious level — our mind asks the question of whether these two are the same or not, and that process of asking and wondering distracts us from simply viewing.” [114]

This statement would agree strongly with work of Sagi and further strengthen Ramachandran’s view that contrast is an extremely important feature of aesthetics.

The work of Sagi et al. could be said to cover two of Ramachandran’s laws: contrast, described above, but also perceptual problem solving. Perceptual problem solving refers to the fact that in some contexts, an observer being challenged, solving a problem, can add to the aesthetic quality of a piece of art. In his article “Is Art Lawful” Tyler discusses these principals and Ramachandran’s approach. He notes “of the eight principles delineated, the one that may be of most interest to artists is that of perceptual problem solving (where the subject matter of the artwork can be extracted only with some effort rather than being immediately obvious).” [171]

This would suggest that some aspects of aesthetics, complexity for example, might indeed be more influential than others. While Birkhoff’s model might have been over simplified, some combination of these aesthetic aspects might prove to be a good indicator of aesthetic value after all.

2.3.2 Musical Aesthetics

Defining and measuring the aesthetic qualities of a piece of music is a remarkably difficult challenge. Even when taking a simplified definition of musical aesthetics to be ’the qualities that makes a piece of music enjoyable to listen to’ we are met with a huge number of potential factors. Another dimension is added when we consider that the salient factors that determine the 'great' from the merely 'good' pieces of music change from year to year, and this change is not uniform across genres.

In general we know that what makes a successful piece of music pop music: a familiar hook, a steady beat and common, familiar chord progressions. Even still, the specific attributes of the melody may vary from year to year and from performer to performer, along with trends in beat patterns and melodic hooks.

Rather than tackling this massive problem, the work in this thesis focuses more on a higher level notion of aesthetics, using more general attributes such as complexity, intensity, and harmony.

Musical Harmony

Consonance and dissonance capture a notion of pleasant or unpleasant harmonies respectively and are fundamental to our understanding of musical aesthetics. The concept is greatly influenced by cultural differences and is largely subjective across individuals. It is also largely influenced by timbre which varies between sound sources and may introduce beatings between close harmonics. This makes it difficult to build a strong representational model of consonance or dissonance that holds across cultures and over time.

Since Pythagoras, consonance has been observed to correlate to the frequency ratios of notes. Galileo was the first to draw the connection between these ratios and the operation of the eardrum. He noted that pairs of musical notes with simple integer frequency ratios tended to sound more pleasant than those with more complex ratios, with the eardrum "kept in perpetual torment" [59].

The idea of "roughness" was first introduced by Helmholts [179] to describe the auditory phenomenon of harsh sounding signals due to amplitude fluctuations or a "beating" effect. It has been hypothesised that the roughness of a signal has a strong equivalence to the consonance or dissonance of that signal, at least within western music, however, the link between this measure and some notion of aesthetics is unclear [176, 83, 178].

There have been numerous approaches to modelling the consonance or roughness of musical intervals both mathematically and experimentally based on trials involving human subjects.

Helmholtz, wrote at length about the topic of roughness [179]. His observations, though often referenced and strikingly similar to more modern revisions, were not much more than personal observations based on musical theory. However his model of roughness provided a starting point for further research to expand upon.

Musical Complexity

"In general, an object's complexity reflects the amount of information embedded in it." [155] The complexity of a given piece of music may be measured in many ways, with rhythmic complexity, tonal or timbral complexity, and melodic complexity, representing completely separate and distinct dimensions.

The most basic representation of musical complexity is the rhythmic complexity. In this context, we do not consider variations in pitch, volume, or timbre, but consider a single temporal pattern without variation in pitch or timbre, repeated over time. The length of time the note is played for is the only variable in this instance. We now have a minimal representation of rhythm.

Pressing[142] discusses similar rhythmic patterns and proposes three distinct types of complexity therein:

- **Hierarchical Complexity** The existence of multiple structures or patterns in a music piece.
- **Dynamic Complexity** The change and difference of a music piece over time.
- **Generative Complexity** The Kolmogorov complexity, or the length of the shortest possible computer program capable of generating a given object, or music piece.

In an attempt to bridge the gap between the technical complexity measures proposed by Pressing and human perception of complexity, Shmulevich and Povel [155] propose three separate measures aiming to match the human perception as close as possible:

- **Root Pattern Elaboration** Based on Tanguiane's Artificial Perception approach [166], "taking the maximum number of root patterns, over all possible structural levels, required to generate the rhythmic pattern in question."[155].

- **Dynamic Complexity** The change and difference of a music piece over time.
- **Generative Complexity** The Kolmogorov complexity, or the length of the shortest possible computer program capable of generating a given object, or music piece.

The timbral complexity of music, representing the complexity of a sound at any instant is reflected in the timbre of the notes played. Variations in note pitches, volume, and the combinations of instrumental timbres all have a strong impact on this measure. This measure is captured by the harmony of sounds and has been discussed separately in Section 2.3.2 above.

Finally, beyond the rhythmic and timbral complexities, melodic complexity is perhaps the most salient. A melody may exist independent of rhythm, as a set of successive notes without any temporal information, but in practice, the melody that a person perceives is a combination of notes and rhythm together.

Non-rhythmic melody, or *melodic contour*, is often represented by Parsons Code [135] which disregards rhythm, absolute pitch, and even the intervals between pitches by representing a melody using just 3 symbols: $\boldsymbol{u}$, $\boldsymbol{d}$ and $\boldsymbol{r}$, or **Up** from the previous note, **Down** from the previous and **Repeat** of the previous note. This remarkably simple representation was theorised to be useful in song identification, and indeed at first glance it seems to create distinct symbol strings for separate songs quite well, however, later studies have shown that the use of melodic contour in song retrieval is relatively impractical with queries of at least 30 symbols required for reasonable success [172].

The perception of melodic complexity seems to be related to how predictable that melody is. Predictability may be based on previous melodic patterns, either in the song itself or from genre or stylistic examples, as well as expectations of resolution. According to Eerola, "melodies which create expectancies that are clearly structured in terms of their tonal, intervallic, and rhythmic properties tend to be easier to reproduce and recognise, and are also judged by listeners as being less complex." [42]

Musical Intensity

In this work we consider intensity as a single dimension, as the intensity of the music being performed independent of any particular emotion.

In a study investigating perceptions of musical intensity conducted by Brittin et al., *intensity* in its everyday use has been described as “an affective perception of music that conveys strong, ardent, or concentrated emotion”[24]. This is in contrast to the acoustic definition of the phenomenon related directly to the volume, or loudness of a sound. With this distinction in mind, the same study used classical musical pieces with strong changes in “loudness, tempo, rhythm, melody, harmony, timbre, and pitch height that were likely to effect changes in the apparent intensity conveyed”.

It is not clear what specific aspects of a song create intensity, or even the relation between various aspects in this regard. Some evidence has been reported of specific musical features provoking emotional responses — such as appoggiatura, an extra note leading into a phrase, provoking a tearful response — however variations between individuals was too great for results to be of significance [158]. Indeed it is possible that individuals experience intensity differently, however there is a reasonable likelihood that the various measurable aspects of a music piece all contribute to various degrees.

2.3.3 Visual Aesthetics

As we have previously discussed, the work in this thesis is concerned with producing a visual display in the form of two LED stage lights. These lights can vary in brightness and colour over time. By taking this approach we avoid the common pitfalls of any graphical computational art system such as line, shape and two-dimensional pattern which are attributes associated with a two-dimensional image. We introduce visual aesthetics in this context.

Visual Harmony

Research suggests that the colours of any object, be it a stage light, an advertisement or a teapot, may directly influence the effect it has on the people interacting with it [104]. Further work has used this fact in reverse — not just to observe the effects of colour, but to generate colour palettes to affect the aesthetics of physical objects [170].

Studies attempting to formalise colour harmony face a number of challenges. Similar to musical perception, the aesthetics of colour are influenced by context, cultural bias, personal preferences and the mood of the viewer. The strength of these effects are shown by a 2001 review of 12 separate books concerning colour harmony which: "failed to demonstrate a general acceptance of any ranked list of terms" [28] used to describe or define colour harmony. The difficulty in measuring and building a model of colour harmony is compounded by historical colour harmony models which "appear to be simplistic or overly generalized in nature; or seem to be based on unfounded opinion or unsubstantiated claims"[126].

If the creation of a reliable model of colour harmony remains beyond our reach, we can turn to the structure of the eye and human physiology to help build an understanding of colour perception. The retina of the human eye contains two types of cells that are sensitive to light: rods and cones. Rod cells are more sensitive to light and play a large role in peripheral, and low light vision, but are not believed to be sensitive to colour. Cone cells are sensitive to colour and are found in three varieties in the retina, each sensitive to different wavelengths of light. Cone cells are not distributed uniformly in the eye, and the sensitivities of cone cells to various wavelengths of light may also vary even between individuals with so-called "normal colour vision" [18, 123].

The perception of colour is not directly related to the sensitivities of cone cells, but rather a complex process including differential outputs of cone cells processed by the visual cortex. It does, however, provide a starting point, demonstrating that colour perception is based on the combinations of wavelengths of light.

Colours can be represented in many ways, called colour models. There are many colour models used for different purposes, including lighting, print and device independency. We now introduce some relevant colour models.

- **RGB** The simplest and most common colour model is the RGB model, which stands for red, green and blue. RGB is an additive colour model, meaning the colour is produced by the addition of various strengths of three primary additive colours, red, green and blue.

- **HSV, or HSL** These models are a mapping of the RBG model which better represents the features of human colour perception, such as colour hue, saturation, and value, or lightness.
- **CMYK** A subtractive colour model typically used in print media where colours are created by mixing cyan, magenta and yellow pigments. Most colours can be obtained with these three pigments, however, a black pigment is required to produce some dark colours.
- **CIE** This model was designed to map closely to the human perception of colours based in the wavelengths that cone cells are sensitive to. The model was named for the International Commission on Illumination, and is based upon experiments originally carried out in 1931 [47]. Variants of this colour model are still in use today.

The harmony of colours is affected by each dimension in a given colour space - but is most often associated with colour hue. While we have shown that historical models of colour harmony may not be entirely correct, commonly used paradigms of colour harmony may still be used as a guide.

Practical colour harmony tools often make use of a colour wheel, a ring of colours where hue is varied in 360 degrees. Colour palettes are then made up of various hues with specific hue offsets around the colour wheel. Typical examples include:

- **Complimentary** 2 Colours offset by 180 degrees apart.
- **Analogous** 3 Colours beside each other, 30 degrees apart.
- **Triad** Three equally spaced colours, 120 degrees apart.

Many other combinations and approaches to combining colours in harmony are possible and have been discussed in great depth [66], including some work attempting to directly relate the aesthetics of musical intervals to colour hue intervals [109], though the objective quality and saliency of any one approach remains questionable.

Visual Complexity

The representation of a visual display adopted in this work ensures that any questions of two-dimensional visual complexity are avoided. In the absence of two-dimensional complexity, we are left with temporal complexity. Temporal visual complexity bears a strong resemblance to the temporal aspects of musical complexity, discussed above in Section 2.3.2.

Rhythmic complexity is essentially identical across both domains, where musical rhythmic complexity disregards tone and pitch, visual rhythmic complexity disregards colour and brightness. In both cases we consider only a temporal beat pattern, which may be analysed similarly to musical rhythmic patterns.

Another parallel can be drawn between musical melody — a variation in pitch over time — and the colour hues of our stage lights, which also vary over time. With this approach, a visual melody can be created and similarly represented and analysed in terms of predictability and "expectancy", which have been discussed in terms of musical complexity above.

Visual Intensity

In a two-dimensional image, the intensity of a visual is a function of strength of line, sharpness of edges, strength of patterns, and boldness of colours among other aspects. Representing physical phenomenon such as intensity in a graphical format has been the topic of many artistic pieces including the work of Kandinsky [84], and Klee [92].

Once again, however, the representation used throughout this work removes many of the complications of two-dimensional graphics. We are left with the brightness of colour as the only dimension which reflects the intensity of the visual at any one time.

2.4 Gathering Aesthetic Data

As we have discussed in the preceding sections, the research into the aesthetics of music and visuals is wide ranging, but historically based on anecdotal evidence or personal opinion, which can often be misleading. It is therefore quite important to base any further research into the aesthetic connections between music and visuals on empirical evidence.

However, gathering empirical evidence in this area is not straightforward and there are many opportunities for human error and bias to creep in. This section is dedicated to discussing the various methods used to gather aesthetic data, the challenges faced, and the solutions to those challenges.

There are two main approaches to gathering aesthetic data that are covered here. Firstly, by conducting a study asking human subjects about their aesthetic preferences and responses to various stimuli. This approach is discussed first in Section 2.4.1. Secondly, aesthetic data can be gathered by analysing aesthetic artefacts such as paintings, music pieces or other artistic objects which have been categorised aesthetically. This is discussed in Section 2.4.2 below.

2.4.1 Psychological Studies

The overall aesthetic value of any art piece is entirely subjective, and therefore a useless measure in any practical sense. However, the individual aspects of art piece — the loudness, the harmony, the contrast — can be evaluated somewhat objectively. For example, if a set of people are asked to report the beauty of a piece of music, each subject will likely give a different answer. The beauty of a piece of music is objectively unmeasurable. If a similar set of people are asked to report the loudness of a piece of music, however, they are likely to give similar answers. The loudness of music is objectively measurable.

If the loudness of a piece of music is one aspect of its overall aesthetic value, we can use it to obtain some notion of objective aesthetic value - even if aesthetic value itself is unmeasurable.

Musical Aesthetics

The loudness of a piece of music is a physical attribute that is easily measured, but it does not give us any insight into the human perception of musical aesthetics. The harmony of music [2] on the other hand, is an aspect of the overall aesthetic quality of a music piece,

---

[2] In this context, musical harmony is meant specifically in terms of consonance and dissonance - or the perception of how enjoyable a set of notes sound when played together.

similar to loudness, but is not simple a physical attribute. It is a human aesthetic perception that is objectively measurable.

Between 1913 and 1915, Malmberg published a study which tested 1045 subjects in three distinct groups for their preferences of consonance [111]. Subjects were played two pairs of tones – dyads, "or two-clangs" in Malmberg's words – on a piano or violin and asked to select which they preferred in terms of consonance. A strong effort was made to ensure all subjects understood the concept of consonance as Malmberg noted "the fundamental reason for the great divergence in the ranking by experts and the consequent disparagement of the ranking of consonance and dissonance has been due to the failure to take common ground in the definition of these terms".

Malmberg's study was conducted on a scale that has not since been reproduced. However, he did not build a ranking solely based on the responses of his subjects. Malmberg's tests aimed to compare the responses of his 1045 subjects – the "empirical ranking" – to a "norm", a ranking which had been developed based on the responses of "eight observers [who] were carefully selected on the basis of their training and fitness for the work". The eight observers were originally required to produce independent rankings, of which the averages would become the "norm". Malmberg notes that after an initial conference, the "discussion and mutual criticism was so stimulating and interesting that all the observers agreed to sit again and continue by the same method until all should agree and a unanimous verdict could be handed in as in the case of a jury". The results of this study appear unprincipled and highlight potential difficulties in obtaining agreement in this domain.

Though the results obtained from Malmberg's observers did not directly affect the responses of his empirical study, the order of presentation and the particular dyads compared by subjects were based on his initial flawed observations which may have affected the final results.

Later, Plomp and Levelt described their approach to investigating the effect of critical bandwidth — the sensory limitations of hearing — on consonance by having subjects rate pairs of tones (generated by sine-wave oscillators) on a seven point scale, consonant to dissonant [139]. They mention that some subjects had to ask for a definition of "consonant" which they provided as "beautiful and euphonious" as it had be ascertained that "consonant, beautiful and euphonious are highly correlated for naive subjects." Obviously Plomp and Levelt conducted these experiments with a relationship between consonance and aesthetics in mind, which serves to highlight a point made by Malmberg regarding a failure to adopt a common ground on the definition of consonance and dissonance leading to disagreement in the domain.

Though the experimental set up that Plomp and Levelt used was rather inelegant — equipment had to be readjusted by hand between each test — their tests have a number of aspects in common with future studies. Firstly, subjects were not all musically trained. Secondly, the tone pairs generated for their tests were composed of simple sine waves which reduced the tonal complexity of the resulting sound presented to the subject. Finally, their subjects rated tone pairs on a graduated scale.

Plomp and Levelt generated tone pairs about a set of mean frequencies (125, 250, 500, 1000 and 2000 Hz) and used a separate sample group for each. The sample groups were reduced in size from 19, 22, 18, 11 and 18 to 11, 10, 11, 10 and 8 respectively when they removed subjects who displayed incoherent responses.

Kameoka and Kuriyagawa conducted an independent study on the absolute and relative consonance of dyads [83]. Like Plomp and Levelt, their sample groups included two groups and their sample tones were generated using sine-wave oscillators. In contrast however, subjects were not asked to rank consonance and dissonance absolutely, but rather in a relative manner. Subjects were presented with two dyads (A and B) and asked to provide an answer on a five point scale, -2, -1, 0, 1, or 2 "according to the subjective distance in consonance between A and B. If B is more consonant than A, a plus sign and if more dissonant, a minus sign was given" [83].

The groups chosen by Kameoka and Kuriyagawa for their experiments are rather puzzling. The first, "audio engineers, who are regarded as ordinary people" and the other, "mixers of ... the Japan Broadcasting Corporation", later referred to as "specialists" [83]. The authors proceed to mention that their "experiments were carried out with audio engineers, and the results in this paper should be interpreted as for ordinary people", a contradiction in itself.

It should be noted that Kameoka and Kuriyagawa used exclusively Japanese speaking subjects. This resulted in the concepts of consonance and dissonance being defined in an entirely different language, adding a further layer of complexity to Malmberg's notion of the definitions of these concepts and how that may affect a subject's response. Additionally, there may also have been a cultural influence which might have had a strong effect.

Hutchinson and Knopoff later produced a "formalism" for calculating the consonance of a pair of notes based on the work of Helmholtz, Malmberg and Plomp and Levelt [79]. They discuss a number of formulas which serve to produce absolute values of consonance for different intervals. The comparison of their values to the results produced by Malmberg seem to show a large degree of correlation, however they fail to provide a direct assessment of their values by means of experiment.

It has been noted that there are issues with the reproducibility of both studies carried out by Kameoka and Kuriyakawa, and Hutchinson and Knopoff [113] suggesting that the methods they used were simply not informative enough to accurately describe the behaviour of their subjects.

Plomp and Levelt had their subjects rank note pairs in an absolute 7 point graduated scale [139]. Malmberg based his ranking on the unanimous decisions of hand picked experts and then tested on large groups of subjects all at once, with a common test pattern [111]. The work of Plomp and Levelt, and Malmberg has been the basis of other papers attempting to produce a formulation of consonance, such as Hutchinson and Vassilakis [79, 176]. None of these studies have taken into account the contextual effects of juxtaposing particular note pairs or ranking intervals absolutely and the effects these approaches may have on results.

For example, the first note pair a subject is presented with is the only pair that does not follow a previous pair and thus has no context. The subject knows the pair must fall somewhere on the scale of dissonance provided but without any reference point, their first response is arbitrary. Alternatively, if a subject is given a number of note pairs in succession which would otherwise have been ranked on one extreme of the scale, they may be more likely to spread their ranking across the scale. Furthermore, after these similarly ranked note pairs, if a new note pair is given that, in another context would be close to the center of the scale, the subject might subconsciously rank this pair further to the other extreme of the scale. It could be argued that successive dyads may be presented to subjects with a large enough interval to ensure short term memory does not effect the response. Indeed Plomp and Levelt adopted this strategy with a 4 second interval [139], however other studies do not follow this paradigm, for example Kameoka and Kuriyakawa who mention an interval of only 0.5 seconds [83].

The problem here is context. Human beings are heavily influenced by context which can lead to incoherent responses. Studies conducted like those described in this paper depend on each test being free from contextual influences from previous tests, which is simply not possible given the experiment design in the studies mentioned earlier.

There are three main solutions to this contextual conundrum. Firstly, to present pairwise comparisons rather than an absolute ranking which reduces the effect of a user subconsciously dispersing their rankings. This approach was adopted by Kameoka and Kuriyakawa with success [83]. Secondly, to test subjects individually and present comparisons in a random order, which will reduce the effect that the first note pair presented without context has on the result across a population. Finally, some studies use test intervals in a training session to establish a relative scale for a subject. However there are issues with this approach as different studies my train subjects to different degrees, by different methods, or indeed over train subjects and introduce fatigue.

In many earlier studies this solution was not achievable. Malmberg went to great lengths to find reproducible tones, testing pianos, organs, violins and bottles filled with wax among other instruments. If the subjects had been tested individually, it would have been arduous on the musicians and unreliable in terms of producing identical note pairs due to the fluctuations in timbre and pitch over time due to fatigue and temperature or humidity changes. With modern computer systems, it is possible to produce identical tones over and over again, as well as randomizing test orders fairly and handling the resulting data.

Another problem we have seen is how researchers may simply disregard incoherent, or contradictory responses from subjects [139]. We believe there is information to be gathered from contradictory responses and that rather than being disregarded, contradictions may actually provide a greater insight into the subjective preferences of test subjects, whether it be consonance, dissonance, aesthetics or any other subjective – or noisy – domain. Once again, it may have been the technological constraints of the time that forced researches to reduce the complexity of their analytical processes, but with modern technology it should be achievable to handle these cases.

Color Harmony

The harmony of music notes is a well studies area with a great deal of music theory backing up theoretical approaches. The same cannot be said for colour harmony. Colour theory in general is far less well structured than music theory and paradigms are based on less empirical evidence.

A simple example of this might be the difference between colour vision and the colour wheel, where visible colours exist on a linear spectrum of electromagnetic frequencies from infra-red to ultraviolet, yet the colour wheel manages to be continuous and circular with red and violet merging seamlessly. Even this cursory investigation demonstrates that the two systems are only loosely linked. Nevertheless, a number of studies have been conducted investigating perceptual responses to colours.

Chuang et al. investigated the relationship between 46 pairs of points within a CIELAB uniform colour space and their effect on perceived colour harmony. The study found subjects to be extremely unreliable in their responses. However, they still report that "the color interval of lightness may be the dominating factor with respect to the influence of (perceived) color difference" [36].

Schloss et al. conducted a number of experiments testing the aesthetic responses to colour combinations with particular focus on the distinction between a subject's preference for a colour combination (whether they liked the colours or not) versus the colour harmony of a combination (whether they felt the colours went well together) [154].

In their first experiment, focusing on preference, Schloss et al. find that "pair preferences are highest when the figure and ground have the same hue (but differ in saturation and/or lightness levels)". This finding is repeated in their second experiment focusing on colour harmony, with the relation between hue offset and harmony being more pronounced. The variance between experiments is attributed to the influence of a subject's preferred colours which may be present in a colour pair. With subjects instructed to rate harmony rather than preference, the effect is reduced.

Szabó et al. developed a series of mathematical models of colour harmony for both two-colour and three-colour combinations [165]. Of particular interest here is the influence of colour hue for two-colour pairs where "hue difference was ... found to be a relevant factor of colour harmony impression".

2.4.2 Algorithmic Approaches

The results of psychological studies of aesthetics are often skewed due to the influences of inadequate stratification, cultural differences, historical biases, human fatigue, and personal preferences as well as many other factors. An alternative approach is to use statistical and machine learning techniques on large datasets to determine the qualities of artistic objects in an entirely objective, data driven manner.

Machine Learning approaches to colour harmony

Nishiyama et al. utilize a support vector machine to categorize images into aesthetic groups. They make use of a public database of images and rankings (DPChallenge [32]) with crowd-sourced meta-data [124].

Lu et al. make use of a large colour harmony evaluation dataset (CHE-dataset), consisting of 29,844 images with associated aesthetic meta-data. They use a latent Dirichlet allocation and Gaussian mixture model to discover a selection of categorized colour palettes. These palettes are then used to identify aesthetic images [108].

The work of Nishiyama et al. and Lu et al. seems to suggest that aesthetic images tend to make use of common colour palettes, although they do not provide any insight into the specific relationships between those colours.

One potential downside to machine learning techniques is a reliance on the underlying dataset. For example, the aesthetic visual analysis dataset (AVA [122]), of which the CHE-dataset is a subset, provides 250,000 images along with rich meta-data including multiple aesthetic scores for each image. The information available in this dataset is incredibly useful in training AI systems, however the authors provide no information on the source of the meta-data, when it was obtained, and from what cultural context. The DPChallenge dataset proves more problematic as it is continuously changing based on public use of the the website. Making the comparison of different algorithms difficult.

2.5 Music Visualisation

Sight and sound are perhaps the most important senses we use to perceive the world. The senses excite deeply interconnected networks within the brain [103, 17, 91] and form a complex relationship that gives rise to a fascinating interplay of perception [117]. There is evidence to suggest that some connections between visuals and audio might be the result of learned behaviour [70], but as we will discuss in the following sections, a number of examples of innate audio-visual connections within the human brain have been observed. This in-built connection between audio and visuals has been exploited by artists, inventors and researchers throughout modern history with many attempts being made to formally bridge the gap between the two domains.

2.5.1 Synaesthesia

Obvious and intriguing examples of audio-visual interplay within the human brain have been observed in the many varied reports of synesthesia. The condition might be described as a crossing of the senses, with synesthetes — those who report experiencing synesthesia — reportedly hearing colours or seeing music, among other sensory combinations. The phenomenon is introduced by Oliver Sacks in *Musicophilia*:

For most of us, the association of color and music is at the level of metaphor. "Like" and "as if" are the hallmarks of such metaphors. But for *some people one sensory experience may instantly and automatically provoke another. For a true synesthete, there is no "as if"—simply an instant conjoining of sensations.* [150]

Synesthesia is not considered a disease, but is often thought to be a gift or simply just part of the way a person perceives the world. Because it is not a physical trait, direct measurement or observation of the condition impossible. Reports of the prevalence of synesthesia vary wildly between 2% [181] and 25% [147] with many other observations in-between [146, 112, 50]. These observations are complicated further by a potential 1:6 male-to-female bias [8].

While formal studies on the prevalence and mechanics of the condition suffer from a difficulty in reliably verifying the authenticity of cases, a number of specific bona fide cases have been identified [39].

One case reported by Sacks details the experiences of contemporary musical composer Michael Torke [150] who has produced a number of musical pieces inspired directly by his synesthesia such as *Bright Blue Music*, and *Ecstatic Orange* [168]. Torke's experience of synesthesia is strongly connected to musical key as he reports "G-sharp minor, for example, has a different 'flavour' from G minor" and "Each key, each mode, for him, *looks* as distinctive (and as 'characteristic') as it sounds". Torke also claims to have *perfect pitch*, an ability to identify musical notes exactly without any reference or context. Indeed, he also reports that his so-called *key synesthesia* is dependant on his perfect pitch.

2.5.2 Music Visualisation in Art

A number of artists have attempted to create visual art based on musical pieces. Russian painter Wassily Kandinsky is one such person, credited with creating the first true abstract art pieces and was also a synesthete, discovering his ability to see music as coloured shapes at an early age during a Wagner opera [174].

In reference to his own paintings he would often call quick studies "improvisations" and larger pieces "compositions", a direct reference to musical composition. Numerous pieces such as his 1926 piece *Three sounds* explore the direct relation between sound and visuals proving to be excellent examples of analogy between the two domains in classical art.

Kandinsky's work "clearly shows the artists deliberate use of musical metaphor and analogy" [162] and his purposeful abstraction is further demonstrated by writings on the philosophy of abstract art in which the concepts and theory behind representation of abstract concepts are described in great detail [84].

Paul Klee, another synesthete — who taught alongside Kandinsky at the German Bauhaus school of art, design and architecture — may be one of the greatest examples of analogy between music and visuals. His paintings, for example *Polyphon gefasstes Weiss*, aimed to convey sound directly to the viewer and to achieve this he "translated elements of music into pictorial equivalents" [82].

During his time in the Bauhaus, Klee produced what is possibly the most important document in relation to analogy making between sound and visuals. The Pedagogical sketchbook is a step by step guide to constructing visuals that convey energy such as sounds and motion. It is an intriguing look into the inner workings of his mind, a glimpse at his synesthesia and his approach to art. The book, though somewhat disjoint, covers multi dimensional representations, lines, balance and stroke weights [92].

Len Lye, an artist primarily known for his kinetic sculptures and experimental films worked on a number of pieces based on musical scores being brought to film. His 1935 piece *Kaleidoscope*, a vibrant explosion of colour and motion, and the 1958 black and white animation *Free Radicals*, are two excellent examples of his work. *Kaleidoscope* is set with a high energy "old time-y" salsa style tune and reflects this with splashes of colour and fast, jagged motion. *Free Radicals* however, set with African rhythmic drumming is black and white with smooth animation and well defined abstract shapes. It seems to focus more on the beat and the rhythm of the music, rather than the texture that *Kaleidoscope* conveys [86].

Scott Snibbe, an American multimedia artist that has produced a number of pieces that are intended to be performed in a similar way to, or alongside music. In this way, he invites the audience to reflect upon the connections between visuals and music and allows the user, or performer, to visually express sensations that are not naturally visual. [162].

Snibbe's Motion Phone (1989-1996) was the first networked device that allowed multiple people to work together on one artistic composition. The motion phone creates abstract geometric shapes that move about the screen following the users mouse movements in dynamic ways. Though the system does not generate music itself, it has been exhibited alongside music and is used a method of creating visuals that go with the music or at least flow in a very musical way [161].

Snibbes Bubble Harp is a more recent piece that creates music by drawing Voronoi graphs around recorded finger strokes on a touchscreen. The user can move the seed points which then move in a similar manor to the motion phone. From this, the cells are generated creating a strikingly organic looking animated visual and musical display [160]. The piece could be seen as a precursor to Boundary Functions, exploring personal space using Voronoi graphs in real time.

2.5.3 Visualisation Machines - Colour Organs

The combination of music and visuals has been explored since ancient times with Plato [34] and Aristotle [6] both discussing tone and harmony in relation to visual aesthetics. Later, upon discovering the ability of a prism to split white light into various component wavelengths, Sir Isaac Newton split these wavelengths into 7 separate colours, analogous to the notes in a heptonic musical scale, the most common scale in western music [93].

Experimentation in light performance, similar to musical performance, has taken many forms. Early systems used the term *colour organ* to describe a device, or instrument, used in the creation of these performances. Colour organs often took a piano-like form with keys that would control the production of different colours of light. An early example of a colour organ of this type was proposed by Louis-Bertrand Castel in the 1730's. Castel, a French jesuit monk who created the *Ocular Harpsichord*, utilised glass panes behind curtains that could be raised by the press of a key to alternately shine or block light through the coloured glass [93].

Further examples can be found in D. D. Jameson's 1844 pamphlet 'Colour Music', and Bainbridge Bishop's 1887 electric arc colour painting device [136]. Similarly, Alexander Rimmington, who "believed that sound and colour were both due to vibrations that stimulated the optic and aural nerve endings" [10] created a device that projected light onto a curtain or screen so that those present could more closely perceive what a composer intended. The device is detailed in Rimmington's 1911 book "Colour-music: the art of mobile colour" [145].

Thomas Wilfred's *Clavilux* created bulbous visual displays akin to lava lamps in the 1920's and by the 1930's he had produced 16 *Home Clavilux* and *Clavilux Junior* units [13].

As manufacturing techniques improved, making the production of such physical devices more realistic, numerous systems were designed creating many different optical effects to accompany musical performances. However, few systems added much to the discussion of colour music beyond the artistic concepts that have been discussed in Section 2.5.2 above. More in-depth reviews can be found by Behravan [10], Peacock [136], and Klein [93].

2.5.4 Visualisation Technology

In the mid 20th century, one might have argued that as the manufacture of physical colour organs would become more feasible, the use and application of such devices would grow. However, the advance of digital technology — and the ease and flexibility of its use — enabled the creation of dynamic visual display systems, quickly superseding the need for any specialised device.

In lighting, industry standards have taken over with common lighting protocols such as DMX512, which has become an ANSI standard, allowing the interoperation of many different lighting devices. Lighting devices themselves have also developed a great deal, with LED lighting now available that creates extremely effective and efficient lighting. LED lighting panels can display a wide range of colours, changing colour dynamically with undetectable delays. Modern lights are energy efficient, producing light with very little energy loss through heat.

Beyond simple fixed position stage lights, modern musical performances employ complex and dynamic devices. Moving stage lights have been in use for decades, but more recently, the use of fully addressable LED displays has become standard for stadium and festival headline performances. Displays of this kind can be used in a similar way to a stage-sized television screen, displaying high definition images and animations in time with the performance. More impressive still is the ability to incorporate these displays directly into the design of the stage itself, bringing a 3 dimensional aspect to the performance.

Stage design of this calibre, while intensely captivating, is extremely expensive and out of reach of all but a very few performers.

In everyday life there are other applications of visualisation technology, the most recognisable of which might be the *media player visualiser* — a feature found in many media player applications found on personal computers during the 1990s and early 2000s — with 2 dimensional displays are generated based on the music being played. These displays were often abstract and simplistic 2 dimensional forms - varying randomly in colour, but varying in size, shape or brightness based in the loudness, or frequency spectrum of the music. These systems were simplistic, and have fallen out of fashion in recent years.

2.5.5 Procedural Generation

Beyond the many endeavours to combine visuals and music within specific artistic circumstances or contexts, a number of systems and approaches have been developed to generate music and visuals with more general applications in mind. These algorithmic approaches have become more popular as computing becomes a ubiquitous part of modern life, but as computing becomes more advanced and techniques become more powerful, the instances where these techniques are in use become less obvious, or visible to the layman.

Generating media based on algorithms, as opposed to creating media by hand, is typically referred to as procedural generation. Often, a procedural generation system will take some set of variables, introduce some randomisation and produce an output based on some iterative or recursive algorithmic pattern. By far the most common and obvious use of procedural generation is to be found in game development. The particular algorithms and formulae that

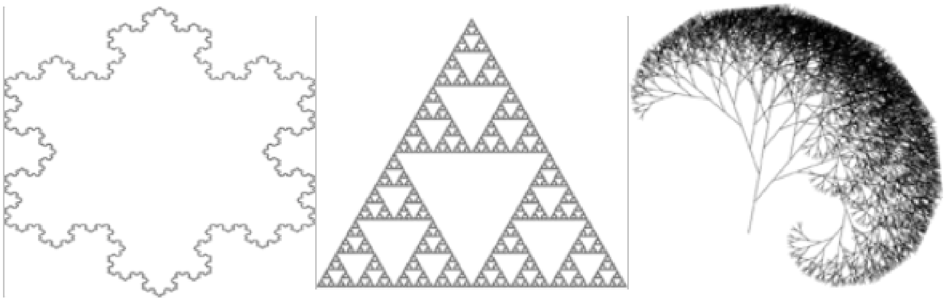


FIGURE 2.2: **Lindenmayer Systems** Examples of fractals produced using L-systems. Koch Snowflake (left), Sierpinski Gasket (centre), Willow (right).

are used are also extremely useful in other areas such as natural language processing, image processing and computer simulation.

Today, it is often less effort intensive and computationally faster to generate vast virtual environments on the fly instead of creating them by hand and loading them from memory, even if this relies on some further processing to actually generate these environments. Sandbox, adventure, and real-time-strategy style games such as Everything, No Man's Sky, Civilisation VI, and many others where content is not strictly scripted, use procedural generation to create environmental objects, maps, and other aspects of the game that are constantly changing as the player progresses.

Importantly, the overall effect created by procedural generation can be extremely impressive and immersing. The recent viral phenomenon Minecraft (2009) is based entirely on algorithmically generated landscapes and has become one of the most successful independently developed games of all time, making approximately 23 million euro only 2 years after early release. [129]

Grammars play a large role in computer graphics and procedural generation as a whole with both formal grammars and context sensitive grammars in common use. These grammars, classically implemented as language generators, have also been used in parsing and recognition, that is, deciding if a particular text string is a valid utterance of a defined language [35]. Indeed, the work in this thesis relies heavily on the application of grammars in Evolutionary Computation.

Lindenmayer systems, often abbreviated to L-systems, are a type of formal grammar known as a rewriting system and are used to

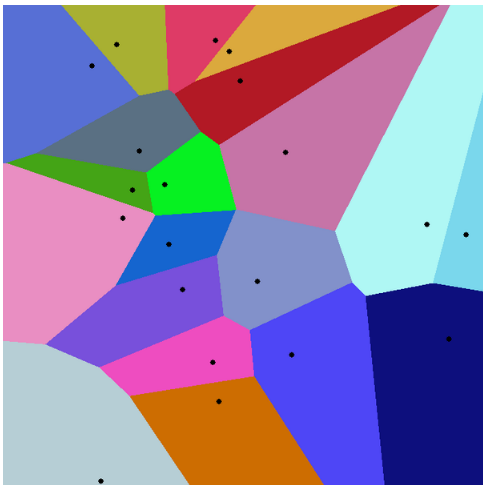


FIGURE 2.3: **Voronoi Graph** in 2 dimensions showing seed points (black) and cells (coloured sections).

build fractal strings based on a particular set of rules. L-systems consist of an alphabet of symbols — letters, words, strings, actions — and a set of production rules which expand symbols and an axiom, the seed symbols that are the starting point of the system [105]. The system was initially created by Hungarian biologist Aristid Lindenmayer and used to study and model the growth of organic systems like plants. They have since been found to be exceptionally useful in computationally generating organic looking flora as shown in the examples in Figure 2.2.

Voronoi graph's are a strong example of computer generated content that has been widely considered artistic and used as part of artistic installations. A Voronoi graph, shown in Figure 2.3, also known as a Voronoi tessellation or Voronoi partition is a method of splitting a space up into sections. The system works by starting with a set of seed points and dividing the space up in such a way at each "cell" of space consists of all points closest to that cells seed point [127]. Using different methods of distance calculation, the output of the system can look quite varied, ranging from very organic, bubble looking cells (Euclidean distance) to grid like patterns (Manhattan distance) which have been used to procedurally generate landscapes divided into natural looking fields [173].

2.6 Artificial Intelligence and Machine Learning

Machine Learning (ML) is a field of computer science within the broader domain of Artificial Intelligence (AI). ML enables a computer system “to adapt to new circumstances and to detect and extrapolate patterns” [149]. Often, a trail and error approach is used to continuously improve the system by splitting data into training and test sets. In contrast to an optimisation approach which aims to find a single optimum solution to a specific problem, ML approaches attempt to find generalised solutions that may be applied to many instances of a problem.

This thesis aims to demonstrate the potential application of AI and ML as tools to enable computational creativity. AI and ML are relatively new technologies, with their practical uses growing as computing power improved within the last number of decades. Early systems such as Samuel Arthur’s checkers playing system [152] demonstrated that some form of learning can be achieved by a computer system, and since then the field has quickly grown. The practical applications of AI and ML have been demonstrated numerous times in both academic and industrial applications with contemporary ML frameworks such as TensorFlow [1], Theano [12] and Amazon Machine Learning [4] among others in wide-spread industrial use.

With the massive potential applications of AI and ML now being realised, the range of AI and ML techniques has also grown. It would be far beyond the scope of this text to cover all of the possible AI and ML approaches that might produce similar results within the problem domain.[3] Therefore a brief introduction to the relevant areas of AI and ML that have been used in this work is now presented.

---

[3] As is often the case in computer science, there is rarely a perfect or canonical solution to any specific problem. It may be possible to use different existing ML techniques and achieve similar results with greater computational efficiency. However, the technologies used in this work have been selected based on their suitability for gaining an insight into the problem domain, rather than producing optimised results. It is hoped that the techniques developed in this thesis will guide future work which may include applying other ML techniques to a similar task.

2.6.1 Evolutionary Computation

The 1950s and 1960s saw the independent creation of a number of systems utilising a similar approach to machine learning based on biological evolution. Lawrence J. Fogel's Simulated Evolution [54], John Holland's Genetic Evolution [73], and Ingo Rechenberg's Evolution Strategies [14] all developed in isolation for approximately 15 years until a generalised term of Evolutionary Computation was adopted in the early 1990s.

These systems use a population based approach where individual solutions, represented by a pseudo genetic code, are evaluated for fitness. A combination of mutation, crossover and selection — known as genetic operators — are then used to form a new generation, mimicking the process of natural selection. The process of evaluating fitness and producing a new generation is repeated until an acceptable solution is discovered, or some other halting threshold is reached. The process allows sub-optimal solutions to be recombined and mutated in order to find incrementally better solutions over time.

Evolutionary Computation can be seen as a search function which is capable of efficiently searching a high dimensional search space using stochastic and hill climbing heuristics. The approach can produce highly optimal solutions, and by making use of the stochasticity of mutation and selection operators, the search space is explored producing potentially distinct and novel solutions.

Evolutionary Computation also benefits from the ability to evaluate the fitness of individuals in a population in parallel. This allows an Evolutionary Computation system to be distributed across multiple computing nodes with relative ease.

The broad definition of Evolutionary Computation together with the many potential methods of application have lead to a number of specific implementation varieties which have developed into distinct bodies of literature. Relevant examples of such distinct implementations include Genetic Algorithms and Genetic Programming. Further examples include Particle Swarm Optimization [140], Evolution Strategies [14] and evolutionary programming [54], however, these implementations diverge from relevance to this thesis and therefore further introduction is not provided.

Genetic Algorithms

The Genetic Algorithm (GA) is perhaps the most broadly defined implementation of Evolutionary Computation as we know it today and is the basis for most later implementations. It is described in general terms, often using a simple binary genome, without domain specific optimisations or performance enhancing additions. This has lead it to be described as the "plain vanilla GA" or "Elementary GA" [41].

In this implementation, the solution we are attempting to discover must be encoded as a binary array. Each digit in the binary array represents a gene, with the array as a whole representing the genome. A fitness function is defined which decodes the binary genome producing a potential solution. The output of this fitness function is the fitness, or effectiveness of the decoded solution.

To begin, a population of individual genomes is generated at random. The fitness function is then used to calculate the fitness of each individual in the population. The genetic operators are then applied as follows.

The mutation operator stems from biological mutation and operates similarly, by randomly "mutating", or changing genes at some rate. Mutation is an extremely effective method of avoiding premature optimisation and allows the population to vary and explore solutions beyond potential local optima.

The selection operator simply choses individuals that are to make up the next generation. Selection can be achieved in many ways, with both stochastic implementations using roulette wheel, or "Monte Carlo" style selection, and deterministic selection based on fitness values being useful to varying degrees.

The crossover operator is applied after selection of multiple individuals. Crossover is analogous to sexual reproduction in biology where the genes of two individuals are combined or "crossed over" to product one or more offspring. The goal here being to product offspring the possess the successful genes of both parents thereby producing a fitter individual. Crossover in a GA differs from biology in that any number of parents may be involved in the process, any number of offspring may be produced, and any number of methods of combining genes may be used.

As previously mentioned, this process is then repeated until an acceptable solution is discovered, or some other halting threshold is reached.

Genetic Programming

Genetic Programming is a variation of Evolutionary Computation designed to directly evolve computer programs. [97] In this case, the solution is not encoded in genes, but rather directly encoded such that the program itself is the gene. There are many methods of achieving this encoding, which require specific implementations of the genetic operators. Determining an optimum method of encoding is still an active area of research, however existing methods have achieved useful results.

Many approaches have been investigated, including Cartesian Genetic Programming [118], whereby programs are encoded on a cartesian graph, or the more common approach: the tree based representation. [141]

Using the tree based approach, a program is represented as tree with each tree node representing an operator and each leaf node, or terminal node representing an operand. A tree can be evaluated recursively. Using this approach, the initial population is generated as a random set of program trees.

Mutation can be applied by changing the values of operands, swapping operators with other valid operators or changing sub-trees within the individual itself.

Crossover is often achieved by swapping sub-trees. The recursive nature of the tree representation allows this to occur by simply selecting nodes in each parent at random as crossover points. The size of the sub-trees that are swapped will not impact the validity of the program.

One of the major drawbacks of this representation is that the operators described above may require significant computational power, limiting the speed at which evolution can occur and a solution can be reached. A second drawback of this approach is that the size of program trees has no fixed limit and through mutation and crossover, individual program trees can grow beyond reasonable limits. Finally, moving away from the elegant simplicity of representing solutions as genes results in an approach that moves away from the biological basis of Evolutionary Computation and results in a great deal of development overhead.

With these limitations in mind however, the concept of encoding a computer program to be evolved is an exciting proposition.

2.6.2 Grammatical Evolution

Grammatical Evolution (GE) is a further specialised form of Genetic Algorithms, "that can evolve complete programs in an arbitrary language using a variable-length binary string."[128] proposed by O'Neill et al. which introduced a number of novel and useful techniques such as *degenerate genetic code*, and *wrapping*. GE encapsulates many of the advantages of Genetic programming techniques, such as variable length genomes and the production of diverse output programs, while maintaining a structure and approach that is strongly rooted in biological evolution, similar to Genetic Algorithms.

In GE, a grammar is used in a genotype-phenotype mapping. The application of this genotype-phenotype mapping enforces a stronger resemblance to biological systems where DNA codons are used to create proteins of a particular shape. In both systems, a many-to-one relationship occurs where many genotypes may produce one particular phenotype, which introduces a natural robustness while still allowing crossover and mutation to take effect. The process of structuring genetic code to achieve this robustness is termed *degenerate genetic code*.

By defining a grammar, we can take any genotype and guarantee a grammatically correct phenotype. The GE genotype to phenotype mapping begins with a gene consisting of a fixed length binary string. Each 8-bit section of the gene is transcripted to an integer value between 0 and 255, known as a codon. The grammar consists of terminals, non-terminals and a starting non-terminal which dictate the structure and content of an evolved program. Terminal expressions represent fixed pieces of the output program, such as an input variable, a constant value or a mathematical operator. Non-terminal expressions represent pieces of an intermediate program that are recursively replaced by other terminals or non-terminals.

According to the supplied grammar, each non-terminal will define a set of possible replacements. Beginning with the first codon in a gene, the value of a codon is then mapped to one of the possible replacements for a non-terminal. In this way, varying the value of the codon can vary the selected replacement and the structure of the output program. The recursive replacement continues until either a complete legal program (phenotype) is created, or a halting threshold reached. It is common for a program to require more codons than are present in the gene and in this case, we simply begin at the start of the gene again in a process termed *wrapping*.

GE provides a number of distinct advantages. Primarily, a GE system suits the creation of executable programs which can be easily defined by a relatively small and simple grammar. In comparison to other Evolutionary Programming systems, the implementation is strongly based on evolutionary models and observed biological processes. The output of the GE system can be parsed and executed simply and efficiently. Using a grammar provides the ability to include useful 'pre-baked' expressions like sin, cos and log functions, as well as application-specific expressions which can save time and help to guide the system to find optimal solutions using domain knowledge. Finally, the output is a textual, human readable program which can be easily stored for later evaluation, analysis or debugging.

2.6.3 Music and Art

Artificial Intelligence and Machine Learning systems have found practical applications in industry with modern systems overtaking the performance of classical algorithmic approaches in many areas. However, in the context of this work, we focus on computer systems in music and art, domains which have seen numerous research endeavours. While these art and music systems do not always make use of classical machine learning techniques, many systems have aimed to produce output mimicking human creativity, firmly placing them within the category of artificial intelligence.

George Lewis' Voyager [102] was built to imitate an improvising musician by listening to phrases of music they played. It has been recorded with and performed since the early 1990's. The system builds a knowledge base of the key, notes, speed, accidentals and other features that describe the music it heard. The system is then capable of generating music that resembles a free jazz improvisation simply by outputting notes and phrases that are similar to those already played. Lewis designed the system to embody “African-American aesthetics and musical practices” using a “a nonhierarchical, interactive musical environment that privileges improvisation”.

A similar example of a knowledge-based system used in computational music and creativity is the PACT, or Potential Action system introduced by Pachet in 1991 [131]. In this system, a knowledge-base of potential actions is built which allows a system to choose an outcome based on a probabilistic model.

The system was put into practice by Ramalho in his work “Simulating Creativity in Jazz Performance” [144]. In this work, PACTS represent a performers potential actions and may consist of individual notes, strings of notes, not playing, or other chains of actions. Each action is then activated based on the context, resulting in musical notes being played. PACTS can be built up from the current piece, or from recordings resulting in a large library of potential outputs, and allowing the system to change over time.

Linson takes this approach further with Odessa [106]. Odessa is designed to operate in a similar way to Voyager with the architecture of the system being the major advancement. The system adopts a subsumption architecture [25] designed originally for robotic systems required to physically interact with the world. Using this architecture, actions are layered in a hierarchy where higher ranking actions may consist of lower ranking actions which all receive sensory input and all produce output. This approach is designed to be flexible and allows systems to be agent based or distributed. The resemblance of the system structure to biological systems also allows for outputs that resemble biological systems.

A number of systems have been created to generate music using evolutionary algorithms, with drumming or beat pattern generating systems being a popular application. EvoDrummer is an example of such a system, using an evolutionary algorithm to generate novel drum beats based on a reference beat [134]. The system provides an interesting look at how a divergent approach can be used to generate novel output in this domain, however the limitation of providing a reference drum beat makes the system inflexible and unsuited for live performance.

Similarly, Ostermann et al. present a system for evolving drum beats using an evolutionary system and Evaluation Rules which allow evolution to occur in real time [130]. Their system dynamically evolves drum beats to match musical inputs, such as a piano or guitar performance. Evaluation Rules codify various aspects of the input music, mapping a set of input variables to one variable representing an aspect of the music being played. Loudness, Staccato and Randomness rules are used alongside rules with attempt to capture the playing styles of iconic jazz performers. This approach allows the user or performer to control the impact of each rule providing control of the velocity of evolution to produce output more closely matching what is desired.

Beyond the limited applications of producing beat patterns, further efforts have been made to make use of machine learning for melodic music.

Functional Scoffolding, based on evolutionary computation [75] has been used to produce melodic accompaniments "represented by a special type of neural network ... which produces harmonies by elaborating on and exploiting regularities in pitches and rhythms".

Autonomous agents with Long Short-Term memory neural networks trained on chord progressions from thousands of jazz compositions have been used for real-time composition by Hutchings et al. [78]. In this system "agents take turns in leading the direction of the composition based on a rating system that rewards harmonic consistency and melodic flow".

Grammatical Evolution in particular has also been used with success with work by Loughrane et al. making use of GE, and tonality-driven fitness functions to create a population of piano melodies [107]. The authors note that the evolved melodies are "interesting and unpredictable" while containing "discernable patterns and motifs". However, while the melodies created might be unique, it is also mentioned that they are not particularly "noteworthy or pleasant sounding". This is mainly attributed to the fitness function which aimed to capture tonal statistical validity, rather than any measure of creativity or "musical goodness".

Machine learning has also been used to produce visuals with many examples demonstrating the potential artistic possibilities.

Cohen et al. have made use of genetic algorithms to mimic the iconic visual style of the abstract modern artist Mondrian [38]. Their work focused on the "balances, color symmetries and composition" of Mondrian artworks produced between 1922 to 1932 and produced images that are identical in style to the source material. Reproducing any artwork to such a degree is, of course a success, however the particular artistic style of Mondrian was purposefully restricted to an extremely small set of rules. All of Mondrian's paintings, for example, followed the same pattern of a white background, non-overlapping rectangles of bold red, blue or yellow colours separated by bold black lines. While this work demonstrates the potential for similar systems, it also highlights the limited scope of existing technologies.

While imitating the style of an artist is possible with some large restrictions, the creation of wholly new art is not currently achievable. Perhaps the closest we have come in recent years to a truly creative and artistic application of AI is the result of the accidental discovery that Neural Networks trained to classify images of an object, contain the information required to generate images of that object.

The application of Convolutional Neural Networks (CNN) to image classification is proving to be powerful tool, however the main criticism of Neural Networks is in their lack of operational transparency. These systems typically pass input through several layers of artificial neurons, gradually identifying higher and higher-level features until the image can be classified. The problem lies in the fact that the output of each layer is not understandable to a human. While the output classification can be verified, the specific method by which the network reaches a conclusion is not easily known.

DeepDream began as a tool to explore this problem [120]. The system works by taking a network that has been trained to identify a particular object, a banana for example, and work backwards to produce an image. "Start with an image full of random noise, then gradually tweak the image towards what the neural net considers a banana." By including some simple rules that force the output to represent natural images, such as neighbouring pixels having similar colours, the output can produce remarkably striking images. The system can be further tweaked to amplify certain features, such as strokes or lines, which produce painterly, stylised images.

2.7 Chapter Summary

This chapter has introduced and discussed the main research themes that are presented in this work. Computational Analogy, a core concept throughout this work, has been introduced in detail with discussion on the various approaches and paradigms used in recent history. As the goal of this work is to use analogy to create art with computer systems, the area of Aesthetics was introduced. Particular focus has been placed on the aesthetic domains of music and visuals which are used throughout this work as motivating examples. The gathering of aesthetic data for scientific and computational use has been introduced with a focus on techniques used to gather visual and musical data specifically. An introduction to musical visualisation, a relatively general topic that spans a number of artistic and academic areas has been provided covering a number of historical and existing techniques. Finally, a brief introduction to the general areas of Artificial Intelligence and Machine Learning and their applications has been provided with a more detailed description of Evolutionary Computation and Grammatical Evolution, core technologies used throughout this work, which will be explored in more detail in later chapters.

Chapter 3

Ranking Subjective Preferences

This chapter is based upon research that has been published in part in the following conference proceedings:

- Aidan Breen and Colm O'Riordan. "Capturing and Ranking Perspectives on the Consonance and Dissonance of Dyads". In: *Sound and Music Computing Conference*. Maynooth, 2015, pp. 125–132
- Aidan Breen and Colm O'Riordan. "Capturing Data in the Presence of Noise for Artificial Intelligence Systems". In: *Irish Conference on Artificial Intelligence and Cognitive Science*. Dublin, 2016, pp. 204–216

3.1 Introduction

In this chapter, the task of ranking subjective preferences is introduced followed by the description of a novel algorithm, *Graphsort*, which has been designed to efficiently rank subjective preferences. The performance of *Graphsort* is then compared to the performance of similar ranking and sorting algorithms in a number of experiments. These experiments are intended to demonstrate the effectiveness of *Graphsort* in the capacity for which it was designed.

Chapter Layout

The layout of this chapter is as follows. Subsection 3.1.1 describes the relevant research questions and hypotheses that are tackled in this chapter in detail. Subsection 3.1.2 describes the task of gathering subjective data in detail with focus on existing systems and the challenges faced by these systems.

Section 3.2 describes a system designed to efficiently gather subjective data while retaining the benefits of existing systems.

Section 3.3 describes a set of experiments carried out to test the efficiency of the aforementioned system. Subsection 3.3.2 introduces the concept of noise as implemented in these experiments. Subsection 3.3.3 discusses our approach to tuning the amount of noise present in each experiment. Following this, section 3.4 briefly introduces our methodology followed by section 3.5 which presents a description of each experiment ordered by the noise pattern used. Each experiment subsection includes a graph of the data obtained, a description of these data and a short analysis. In section 3.6 we discuss the results observed ordered by each algorithm before finally concluding the chapter in Section 3.7.

3.1.1 Research Questions and Hypotheses

The reader is reminded of the first research question of interest in this work, described in Section 1.2.2 as follows: *Can a computationally intelligent system be built to create aesthetic analogies?* Of particular interest in this chapter are the first two hypotheses associated with this question:

H1 Aesthetic data can be gathered in a way that minimises participant effort.

H2 *Aesthetic attributes* in the chosen domains of music and visuals can be *objectively measured*.

Aesthetic attributes refer to the qualities or features of an object that are thought to have an impact on the overall aesthetic appeal of that object. Objects of interest in this work may be musical or visual. A musical object might be a song or melody. A visual object might be a lighting display or a two-dimensional image. The subjectivity of the aesthetics of music and visuals dictates that individual aesthetic attributes are subjective in nature.

Objectively measuring an aesthetic attribute refers to the ability to reliably measure the value of that subjective attribute, regardless of an individual's personal or subjective opinion. Conversely, measuring the opinion of an individual would be a subjective measurement.

By example, loudness may be an aesthetic attribute of music. Loudness can be physically measured in terms of decibels quite reliably, but two individuals who hear the same music piece may have conflicting opinions on whether it is loud or not. If we find a correlation between objective measurements of loudness (decibels) and the subjective responses of individuals, we can say that loudness, as an aesthetic attribute, can be objectively measured.

A counterexample might be the aesthetic attribute of happiness. A major key is often associated with happy songs, and identifying whether a song is in a major key is trivial. However, in practice, there are many examples of unhappy music written in a major key. In this case, happiness is not an objectively measurable aesthetic attribute.

In order to demonstrate that any aesthetic attribute is objectively measurable, it is therefore necessary to compare an objective measurement of that attribute to a collection of subjective measurements, or subjective *aesthetic data*.

3.1.2 Gathering Subjective Data

Gathering subjective measurements is a difficult task. As we have discussed in Section 2.4.1, there have been a number of approaches used in the past to gather subjective data in psychological studies with varying success. While studies utilising these approaches often report reasonable results, common challenges are often overlooked. When gathering aesthetic data, these effects are further exaggerated.

In a 2012 meta analysis of visual aesthetics studies, Palmer et al. discuss the various methodologies used. For empirical studies, where data is gathered as human subjects complete tasks, it is noted that "in most respects the optimal task is 2AFC, in which observers indicate which of two simultaneously presented visual displays they 'like better' (prefer aesthetically) for all possible pairs." [132].

A 2AFC, or two-alternative forced-choice study is conducted as follows. A subject is asked to compare two samples and provide a preference for one or the other. The subject is not allowed to skip a choice. In a full 2AFC study, the subject compares all samples to each other, resulting in a comparison matrix from which a ranking of samples can be made.

The 2AFC approach has been widely used [153, 154] and addresses some of the common challenges faced by psychological studies. However, while aspects of the 2AFC approach are useful, it does not provide an ideal solution for gathering subjective aesthetic data. We now introduce the common challenges of psychological studies and how we have taken each issue into account in designing a system suitable for gathering subjective aesthetic data.

Contextual Challenges

Many studies adopt the popular *Likert Scale* experimental design which involves rating each object in random order on a scale. With this common approach in mind, a number of observations can be made. The first object shown to a subject is viewed in isolation, given no context, and so the subject's response is rather meaningless. If a number of objects naturally fall to one end of the scale — with a score of 0, or Strongly Disagree for example — the subject may be tempted to spread out ratings to show a difference, however small, between those objects. This will skew the actual meaning of their response. Now consider, directly after a number of these low ranking objects have been presented, a middle ranking object is presented; the subject is likely to rate this object much higher than they otherwise would.

In this case, directly comparing two objects using a 2AFC style approach removes any issues with context or juxtaposition of objects.

Cognitive and Memory Load

When asking a subject to provide responses on their aesthetic experience, any distraction or complication may impact their response. It is therefore important to reduce the cognitive and memory load on subjects where possible.

A complex set of instructions given to the subject, an unfamiliar setting, or the comparison of many objects at once might increase this load and affect the quality of responses. The 2AFC approach is simple to explain to subjects and requires very little cognitive load throughout the experiment. The subject can also directly compare two objects requiring no memory load. These two effects in combination should increase the speed at which a subject provides a response and decrease the mental load required to proved this response.

Fatigue and Boredom

Even with a low cognitive and memory load, Palmer et al. note certain drawbacks to the 2AFC approach. The upper bound of choices comparing $\boldsymbol{N}$ objects is $\frac{\boldsymbol{N^2-N}}{\boldsymbol{2}}$. This becomes a major issue for studies with any reasonable granularity. For example, a study of colour harmony conducted in 2001 by Chuang et al. [36] tested 46 colours in pairs similar to a 2AFC approach with 15 repeated test pairs resulting in **1050** total individual colour pairs. Each experimental session lasted approximately an hour, with subjects showing very poor reliability. The authors recognise “1050 judgments [sic] to be made in this experiment are too many for subjects to focus their attention throughout the complete experimental session”.

Considering the alternative approach of simply asking a user to rank each object on a scale which can be conducted in linear time with just 46 choices, it is easy to see why 2AFC would be unsuitable. Possible solutions to this include tournament ranking (See [33] for an in-depth review), or using random sampling to gather as much data as possible until the user becomes fatigued [153]. The downside to classical tournament ranking algorithms is that if a user does become fatigued or bored before they complete their tasks, we may not be able to retrieve any useful data. While random sampling approaches are better in this regard, there is no way for a subject to know whether they have provided enough information or not, leading to the subject continuing the experiment and becoming fatigued.

Design Considerations

While the 2AFC approach handles context and cognitive and memory load well, it is not particularly suited to reduce subject fatigue. An alternative approach is therefore necessary.

The act of ranking (or sorting by preference) by making comparisons between just two samples at a time is functionally identical to a typical sorting algorithm or tournament ranking system. In this sense, assuming the choices that a subject makes are somewhat reliable, it should be possible to obtain a full ranking of samples with far fewer number of comparisons, reducing the effort required of subjects, reducing subject fatigue.

However, each of the challenges we have presented impact the reliability of subject responses to varying degrees. If the subject is confused, overwhelmed, tired, bored or acting maliciously, their responses become unreliable. The reliability of subject responses is a large factor impacting the effectiveness and usefulness of any ranking or sorting algorithm. Later in this chapter we discuss how the reliability of subject responses is modelled using what we call noise patterns. These noise patterns can be applied to a ranking or sorting algorithm to simulate various ways subject responses might be unreliable.

Based on these considerations, we have chosen to explore the use of a directed graph to form the basis of a ranking algorithm.

Using directed graphs to rank or sort objects is not a new idea. There are numerous topological sorting, or “toposort”, implementations using directed graphs. Topological sorting is achieved either by performing a reverse postorder depth first sort, or by pushing the first node found with no incoming edges onto a stack — the output — removing that node from the graph and repeating the process until all nodes have been stacked [133, 2, 67, 87]. These approaches, however, rely on the graph to be acyclic, which is not guaranteed when dealing with subjective responses.

The field of graph theory is extensive, providing many tools for the representation, transformation and computation of digraphs. Further approaches within this domain have been described to rank players in a tournament [148] or candidates in an election using both weighted and unweighted digraphs [33, 7]. These approaches, while capable of handling cycles within the graphs, are restricted to semi-complete digraphs — digraphs where every two vertices are connected by at least one edge.

While these approaches may not be useful to us in most cases, some aspects, such as the Copeland score method[1] have been useful as a basis to form our own algorithms.

3.2 The Graphsort Algorithm

We now introduce an algorithm designed to combine strengths of a 2AFC style study with the efficiency of a tournament or sorting algorithm. To rank a set of objects based on one subject's feedback, we first treat each object as a node on a graph and each response from the subject as a directed edge. The graph is built as described in Algorithm 1.

Algorithm 1 Graph Rank Algorithm: Overview

```
setup:
Calculate rank for all nodes
loop:
Select any two tied nodes $i$ and $j$, and present them to the user.
Prompt user to choose between $i$ and $j$
if preference for $i$ then
    Add directed edge $i \rightarrow j$
end if
if preference for $j$ then
    Add directed edge $j \rightarrow i$
end if
Calculate rank for all nodes
if Any 2 nodes have the same rank then
    goto loop
end if
```

---

[1] The Copeland Score, or Copeland Method [33] is a common and easily understood ranking method for nodes on a directed graph. Nodes are ranked by their out-degree, or alternatively their out-degree minus their in-degree. This method, while easily understood, often leads to ties and bears no relation to the overall structure of the graph.

The efficiency of the algorithm is the result of the strategy used to pick what question to ask the subject next. This is determined by the current rank of all nodes on the graph. The ranking of the graph and selection of questions are now described in detail.

Ranking Nodes

Consider the nodes A, B and C where $A \rightarrow B \rightarrow C$. In this example, B is said to be forward of A (and C forward of B). A *reachable* node is any node that is connected to another node by any number of edges or other nodes. For example, A is *reachable* from C. A *forward reachable* node is a node that is reachable only by following the direction of the edges. In the example, C is *forward reachable* from A, but A is not *forward reachable* from C, as to get to A from C, one must follow edges against their direction.

The set of forward reachable nodes from any node n is defined as δ_n^f. In our algorithm, we define the rank of any node n as

$$R(n) = |\delta_n^f| + 1$$

That is, the rank of any node n is the number of forward reachable nodes from n, plus 1. A node with no outgoing edges will thus be ranked 1.

As an example, consider the graph in Figure 3.1. Initially, the rank of node Y would be 1, as it has no outgoing edges. Z also has a rank of 1. X has two forward reachable nodes: Y and Z. The rank of X is therefore $(1 + 1) + 1 = 3$.

The rank of any node n can be calculated algorithmically by traversing all forward nodes of n, keeping track of all visited nodes, including n. Once all forward reachable nodes have been visited, we simply return the number of nodes in the visited collection. This procedure is formally presented in Algorithm 2.

Adding Edges

To build the graph, we begin with a node for each individual object the subject will be presented with. We decide which two objects to compare — or where to add an edge — by selecting any two tied nodes as described in Algorithm 1. After each comparison, we re-rank the graph and repeat until there are no more ties.

Algorithm 2 Graph Rank Algorithm: Ranking a Node

```
setup:
visited ← empty collection
function RankNode($currentNode, visited$)
    Insert $currentNode$ to visited
    for each child node of $currentNode$ as $n$ do
        if $n$ is not in visited then
            RankNode($n$,visited)
        end if
    end forreturn Number of nodes in visited
end function
```

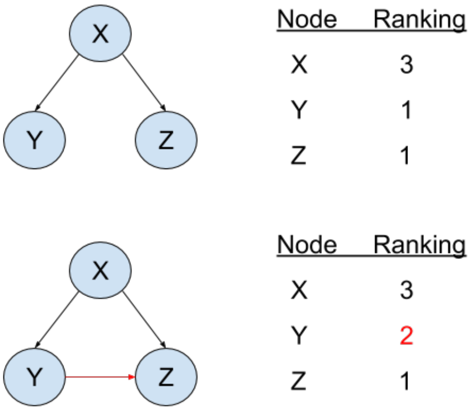


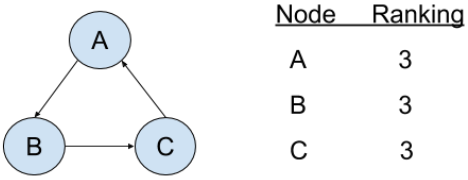
FIGURE 3.1: Top: A graph with a tie. Bottom: A new edge breaks the tie.

Consider a graph such as the one shown in Figure 3.1. We have three nodes, X, Y and Z. In this case, the subject has already shown preference for X over the other two nodes. Nodes Y and Z are tied with a rank of 1.

To break the tie between Y and Z, we ask the subject for their preference. In this case, the subject has decided that they prefer Y. We add a new edge between these two nodes and recalculate the rank. We find that X remains ranked 3, however Y is now ranked 2 and Z 1. As no nodes are tied, we declare the graph fully ranked.

This is obviously a contrived example with an overly simple graph, however, the same principles apply to more complex graphs. In larger graphs multiple ties will occur — indeed before any data is gathered, even the simple graph in Figure 3.1 will have a three-way tie. In this case, we simply choose any two tied nodes of the same rank at random.

Ties which are the result of cycles are ignored when adding new edges. By following the tie breaking strategy stated above, we do not offer the subject an opportunity to create cycles within the graph. For some studies however, with some pre-determined preferences for example, it may be possible to introduce cycles. A cycle such as the one shown in Figure 3.2 will result in tied nodes, each with a defined rank. We believe this is a strength of our algorithm as it accurately captures a structure of equality that the subject cannot express using other methods.



Node	Ranking
A	3
B	3
C	3

FIGURE 3.2: A graph containing a cycle produces three tied nodes.

3.3 Experiments

This section describes a set of experiments conducted to test the efficiency and effectiveness of the *Graphsort* algorithm in comparison to other sorting and tournament ranking algorithms, as well as the robustness of these algorithms to noise.

Sorting and tournament ranking algorithms share a common goal: to take as input a set of items, and produce as output an ordered list of each item in that set using some comparison operator. Sorting algorithms typically use a comparison operator based on alphanumeric order. In contrast, tournament ranking algorithms use matches — or games, bouts, fights — as a comparison operator.

A vast amount of research has been conducted in the areas of sorting noisy data and tournament ranking. Representing a tournament as a directed graph, the minimum feedback arc set (FAS) problem on tournaments formalises this challenge and is NP-complete [85]. Recent work by Braverman [19], Kenyon-Mathieo and Schudy [90] and Wauthier [182] explore this area and propose possible approximation approaches. However, the FAS problem focuses mainly on a fair ranking after a large number of arcs have been created. Instead, the focus of this research is to investigate the calculation of a rank without the need for these arcs in the first place.

In the following experiments, tournament ranking algorithms are used to sort arrays of integer numbers as if each number was a competitor in a tournament. Matches are henceforth referred to as comparisons and competitors referred to as elements. In this way we can fairly compare tournament ranking algorithms to sorting algorithms in terms of performance by number of comparisons, and accuracy. The accuracy of any sorting or ranking algorithm may be measured in terms of how close the output of that algorithm is to a truly perfect ranking. Spearman's footrule has been used in a number of studies to measure this value and is adopted here [45, 182, 29].

3.3.1 Omission of rating algorithms

The algorithms selected for comparison to *Graphsort* in this section may all be described as instances of a common algorithm type: ranking and sorting algorithms. However, there exists a functionally similar, but distinct type of algorithm, the rating algorithm, that also imparts an order upon a set of elements or a group or competitors. Where ranking algorithms intend to provide a ranking of elements at one moment in time, rating algorithms aim to take into account the variable performance of competitors — or value of elements — over an indefinite period of time.

A particularly notable example of a rating algorithm in common use would be the "glicko", or more recently "glicko-2" rating algorithms [64, 63]. These algorithms were originally designed for chess rating, but have also seen use in many other areas such as computer games, sports and other ranked activities.

The algorithms intend to rate players based on the games they play against other rated players. After a completed game, a player's rating ($\boldsymbol{r}$) is updated based on their current rating, and the rating of their opponent. Their rating deviation ($\boldsymbol{RD}$, the standard deviation of their rating), and in the case of glicko-2, their rating volatility ($\boldsymbol{\sigma}$, a measure of the consistence of their performance distinct from standard deviation, and calculated iteratively) are also taken into account.

Glicko and glicko-2 allow players to obtain a rating which may change over time to reflect changing performance. One of the benefits of this approach is that the rating become more accurate with time. These algorithms also allow some players to participate in many more games than others, while still providing a fair rating.

However, algorithms of this type are not comparable to *Graphsort* in the context of the following experiments. This is primarily due to the continuous nature of the systems. In the following experiments, we assume the ranking and sorting algorithms have a defined end state, when the algorithm has returned a ranking. After this point we use the number of comparisons made between elements as a major factor of comparison. In contrast, to determine the end state of a continuous rating algorithm, we would need to employ some method of identifying when a steady complete rating has been reached.

This steady state is impossible to determine for two reasons. First, when the noise patterns described in the next subsection are applied, a steady set of rankings may not be produced. Ratings may continuous vary thus making a steady state impossible. Second, the rating deviation and volatility values will continuously change as more games are played. It is impossible to say whether these values will reach a steady state when noise is introduced.

The unsuitability of these algorithms to tournament style ranking is evident in the fact the they are not used for competitive chess tournaments, where the Swiss style tournament is preferred [45].

3.3.2 On the introduction of noise

For any ranking or sorting system to operate, it must compare two or more items and if these comparisons occasionally produce incorrect results, we say the comparison operator is subject to noise.

In most approaches and domains, these comparisons are trusted to be true. When sorting an array of numbers, for example, the comparison of $1 > 2$ will always return false. However, in many real world applications, the comparison may not be so simple. Of particular relevance is the possible noise introduced by human fatigue or error, introduced previously in Section 3.1.2. For instance, an individual in a study may click a button without thinking, click a button in error, or simply just feel like not answering the question correctly.

The reliability of an individuals responses can be effected by contextual issues, cognitive and memory load, or fatigue and boredom.

Other potential sources of noise could be physical sensors that are influenced by varied power supply or other atmospheric conditions. In this case, the comparison may be correct, but the underlying values may be incorrect. Finally, it may be advantageous to introduce noise into a system in order to overcome premature optimisation in areas such as genetic algorithms. In this case, an artificial noise may be applied to a particularly rugged fitness landscape in order to explore potential solutions that would otherwise be unreachable.

With these applications in mind, we introduce four noise patterns used in the experiments presented later in this chapter. The various algorithms used are summarised in Table 3.1.

3.3.3 Tuning noise patterns

While the amount of noise introduced by any artificial noise pattern can be controlled by input parameters, not all noise patterns are created equal. For two of the four patterns used in this work, the amount of noise that is introduced to a system is affected by the size of the data and the operation of the system itself. For example, with noise that increases over time, a system that requires 10 comparisons will be introduced to a disproportionately small amount of noise than a system that requires 100 comparisons. Alternatively, two systems that require the same number of comparisons with different noise patterns cannot be fairly compared unless we ensure the amount of noise introduced by each pattern is similar. Tuning the input parameters for each noise pattern is therefore necessary.

In order to estimate the noise introduced by a noise pattern using particular parameters, we introduce the Noise Value function. This is demonstrated as the ratio of *false comparisons* to *true comparisons*. A comparison is the boolean result of a relational operator for two values. A *true comparison* produces the result we would expect if neither value was influenced by noise. We consider $\boldsymbol{n(1) < n(2) = true}$ to be a *true comparison*, where the function $\boldsymbol{n}$ applies noise to the values $\boldsymbol{1}$ and $\boldsymbol{2}$. A *false comparison* produces a result contrary to what we would expect if neither value was influenced by noise, and thus we consider $\boldsymbol{n(1) < n(2) = false}$ to be a *false comparison*.

During the tuning phase, comparisons were made with and without noise; while the noisy result is used in the algorithm, the value calculated without noise is stored and the Noise Value calculated. The input parameters of each noise pattern were then tuned resulting in a similar Noise Value for each pattern.

It should be noted that due to the stochastic nature of the noise patterns, Noise Value is simply an estimation and tuning may not produce exact Noise Values.

3.3.4 Calculating Accuracy

In our results we compare the accuracy of different algorithms. The basis of our accuracy measurement is Spearman's footrule, a measure of the similarity of two arrays of the same elements. It is calculated as the sum of the absolute difference of position between the elements of the two arrays. The footrule value, f, for an algorithm is calculated as the Spearman's footrule of the output of the algorithm and a fully sorted array.

Normalisation is necessary due to the non-linear nature of f. This can be demonstrated by taking a number of arrays, swapping values at random, and calculating the footrule between the original sorted array, and the new augmented array. In Figure 3.3 we have calculated an average value for f after each swap for 50 arrays with 500 elements. After 1000 swaps, f begins to plateau at approximately 83300. This curvilinear plot may be estimated with a power function.

To normalise f (non-linear), we take the estimation function and calculate an approximate number of swapped values (linear).

The normalisation function, solved for *number of swapped values*, S, is

$$S = cLog_2(\frac{b}{a - f})$$

where $a = 83761.92$, $b = 83754.69$ and $c = 178.8318$. Finally, for convenience, S is scaled to plot both the number of comparisons made and the accuracy on the same graph.

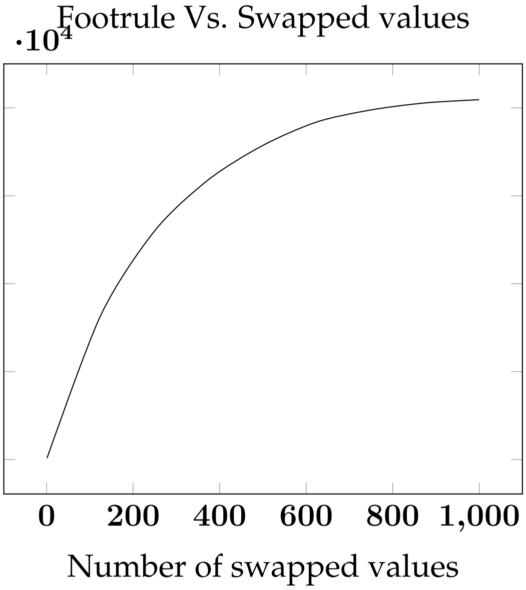


FIGURE 3.3: Footrule increases non-linearly with random values swapped.

3.4 Method

Eight tournament ranking and sorting algorithms were tested using four noise patterns. A control experiment was conducted without any noise present. The noise patterns are described in detail in section 4.3.2.

Before the experiments were carried out, a tuning phase was carried out to ensure noise patterns introduced a similar level of noise to each execution of an algorithm. The details of this tuning phase are outlined in section 3.3.3 above.

A set of 50 randomly ordered arrays of 500 unique integer elements were generated. Each algorithm was executed with each noise pattern to sort all 50 arrays. Each algorithm was implemented with an abstract comparison operator enabling noise to be introduced without altering the implementation of the algorithm. Each algorithm was altered to store the number of times the comparison operator was used while sorting an array.

After each array was sorted, it was compared to a correctly sorted array of the same elements and the Spearman's footrule was calculated and stored. Footrule values are normalized and scaled as described in section 3.3.4 above. This normalized value is displayed as "Accuracy" on the figures below.

TABLE 3.1: Tournament Ranking and Sorting Algorithms

Algorithm title	**Abbr.**	**Average Complexity**
Graphsort - Random tie breaking	GSR	$nlog(n)$
Graphsort - High tie breaking	GSH	$nlog(n)$
Graphsort - Low tie breaking	GSL	$nlog(n)$
Bubble Sort	BBL	n^2
Insertion Sort	INS	n^2
Quick Sort	QCK	$nlog(n)$
Swiss Tournament*	SWS	$log(n)$
Round Robin Tournament	RBN	n^2

*The Swiss Tournament only achieves a partial ordering.

TABLE 3.2: Noise Patterns

Pattern	**Description**	**Parameters**
Control	No noise. Comparisons and values are always correct.	None
Percent noise	Values vary by a percentage of the total range of values.	Value Range, Percent
Flip noise	Comparison result is inverted with a predefined probability.	Probability
Increasing noise	Flip noise with an increase in probability of an incorrect result each time a comparison is made.	Init. Probability, Increment
Tail noise	Percent noise with an increase of noise for larger values.	Value Range, Percent Range

TABLE 3.3: Common Experiment Parameters

Parameter	**Value**
Array size	500
Arrays for each noise pattern	50
Noise patterns	5
Tuned to noise value	0.9

3.5 Results

The following sections describe each noise pattern in detail and present the results of experiments with the average number of comparisons made to sort each of the 50 arrays, along with the average Spearman's footrule for each algorithm.

The algorithms used in each experiment are referred to by their abbreviated name in order to avoid any confusion between similar algorithms, such as Graphsort with each tie breaking strategy. These abbreviations are presented in Table 3.1.

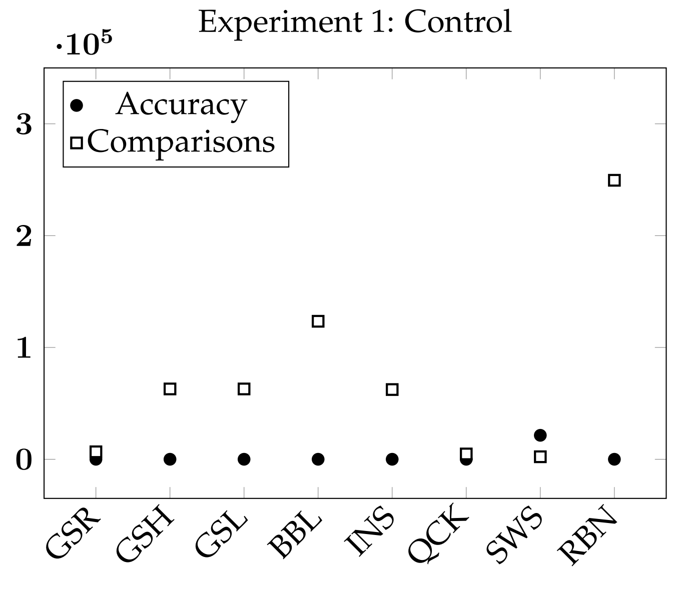


FIGURE 3.4: Accuracy and performance of each ranking algorithm under typical conditions.

3.5.1 Experiment 1: Control

The control noise pattern introduces no noise to the system. As shown in Figure 3.4, arrays sorted using this pattern showed a Spearman's footrule of 0 as expected, with the exception of SWS, which produces a partial ordering[2]. The number of comparisons for each algorithm is indicative of typical performance under normal conditions.

Analysis

SWS produces a partial ordering, with a low number of comparisons as expected, resulting in the only Spearman's footrule above 0. The aim of this algorithm is to identify the strongest competitor in a tournament in the minimum number of matches, while allowing the other competitors to continue competing against competitors of

---

[2] A Swiss style tournament begins by randomly pairing competitors. After each round, competitors are given a score — typically 3 for a win, 1 for a draw, 0 for a loss — and paired again but only with competitors with the same, or similar score. The tournament continues until a clear winner is decided. Assuming no draws, a clear winner is decided in the same number of rounds as a knockout tournament, which is the binary logarithm of the number of players rounded up. This reduces the number of comparisons required to finish the tournament but does not produce a full ranking of other players who have not won (or lost) every game.

a similar quality to build experience. A side effect of this approach is that the weakest competitors are also identified.

As expected, RBN results in the largest number of comparisons, followed by BBL. Both algorithms have an average case complexity of $O(n^2)$. Interestingly, INS performs as well as both GSH and GSL. This is because of the high and low tie breaking strategies which force the graph to be sorted from top to bottom — or bottom to top. In these cases, the elements at one extreme are ranked quickly, at the expense of elements at the other extreme which are sorted in a fashion similar to BBL. In this sense, these strategies share a number of features with SWS.

Finally, GSR and QCK perform the best of all algorithms that produce a full ordering with the lowest number of comparisons.

TABLE 3.4: Noise pattern parameters: Percent Noise

Parameter	Value
Minimum value	0
Maximum value	500
Percent	15
Tuned to noise value	0.9

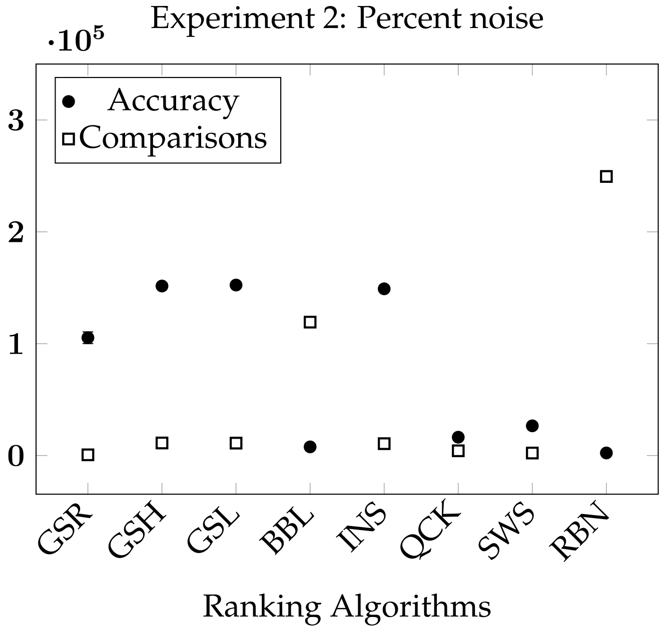


FIGURE 3.5: Accuracy and performance of each ranking algorithm subjected to `Percent noise`.

3.5.2 Experiment 2: Percent noise

This noise pattern introduces a random variance to the values being compared. For any array of values and an input parameter of x percent, each value may be varied by at maximum x percent of the range between the highest and lowest values in the array. For example, for an array of 500 elements, 1 to 500, and a percent parameter of 15, any value being compared may vary by plus or minus 15% of 500 (75).

This has the effect of producing unreliable comparisons for elements that are similar, and reliable comparisons for elements that are very dissimilar. The comparison $110 < 120$ should typically be true. With `Percent noise` applied however, we may actually observe the result of the comparison $(110 \pm 75) < (120 \pm 75)$ which could be $170 < 103$, which is of course false. On the other hand, $30 < 400$ will always produce the correct result as the difference of the values is larger than the sum of any potential noise.

Analysis

As shown in Figure 3.5, in terms of the number of comparisons needed to reach a partial ranking, round robin and bubble sort have a very poor performance. A by-product of the large number of comparisons is that the partial ranking produced by both algorithms is much closer to the correct ranking. This is expected as larger numbers of comparisons may reduce the effects of perturbing the data according to a randomly distributed noise pattern.

The performance of the other algorithms is generally much better, with less extreme variation. INS, GSL and GSH occupy the middle range with more comparisons and a larger footrule than the higher performing QCK, SWS and GSR. In terms of the number of comparisons, GSR is capable of achieving a partial ranking with the best performance. However, this is at the expense of accuracy where QCK and SWS have the advantage.

TABLE 3.5: Noise pattern parameters: Flip noise

Parameter	Value
Probability*	0.105
Tuned to noise value	0.9

*The likelihood of a comparison to produce the wrong result.

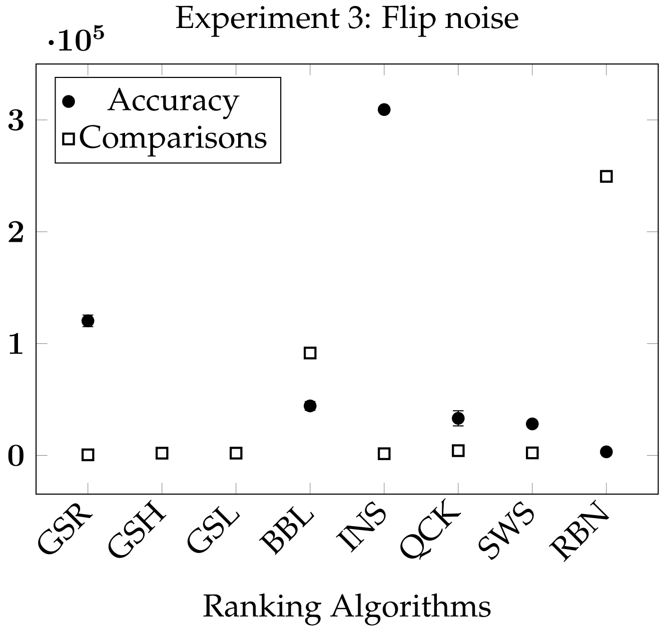


FIGURE 3.6: Accuracy and performance of each ranking algorithm subjected to `Flip noise`.

3.5.3 Experiment 3: Flip noise

This noise pattern represents the comparison operator producing an incorrect result — regardless of the values being compared — with a specific probability. It is similar to the previous pattern in that it introduces random noise to the system. However, this pattern is not sensitive to the actual values being compared, as `Percent noise` would be. The effect of this noise pattern is analogous to a human participant who may give the wrong answer to a question through error.

Analysis

The results of this experiment, shown in Figure 3.6, are similar to those generated using `Percent noise`: RBN and BBL requiring the most comparisons to produce relatively accurate partial rankings, and GSR, QCK and SWS producing rankings with much fewer comparisons at the expense of accuracy. This is perhaps unsurprising due to the similarities between the two noise patterns. However, there are some variations that may be a result of the lack of sensitivity to the actual values being compared.

QCK was relatively unaffected by the change in noise in terms of comparisons made, but suffered greatly in terms of accuracy with an increase in footrule ($\boldsymbol{f}$) of $\mathbf{82.587\%}$ relative to experiment 2. This is easily justified as the divide and conquer nature of quick sort would lead to large portions of the result being out of place, even if these out of place portions were independently partially sorted.

BBL seems to also have suffered greatly in terms of accuracy with a $\mathbf{356.956\%}$ increase in $\boldsymbol{f}$ relative to experiment 2. Again, this is easily because of BBL attempts to bubble each element into place and one incorrect comparison — which can now occur at any point, rather than close to the correct point in the previous noise pattern — can result in an element being placed incorrectly. One incorrect placement like this may then have a further influence on subsequent elements.

GSR actually sees an increase in performance here with the number of comparisons decreasing by $\mathbf{20.553\%}$ relative to Experiment 2. Interestingly, this increase in performance does not come with a large decrease in accuracy, with $\boldsymbol{f}$ increasing by only $\mathbf{6.067\%}$ relative to Experiment 2.

TABLE 3.6: Noise pattern parameters: Increasing noise

Parameter	**Value**
Probability increase*	0.0000045
Initial probability	0
Tuned to noise value	0.9

*The increase in probability for each comparison.

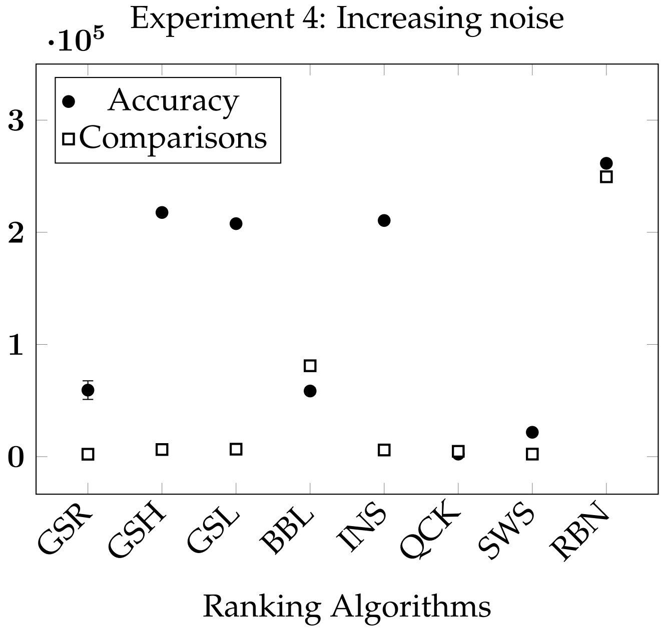


FIGURE 3.7: Accuracy and performance of each ranking algorithm subjected to `Increasing noise`.

3.5.4 Experiment 4: Increasing noise

This noise pattern also represents the comparison operator producing an incorrect result with a certain probability, similar to `Flip noise` above. However, the probability of an incorrect result increases with each comparison made. The effect is perhaps analogous to a human participant becoming more fatigued over time, which will increase their probability of answering questions incorrectly. Again, the noise in this pattern is not sensitive to the actual values being compared in contrast to `Percent noise`.

As the amount of noise present in this pattern is dependant on the number of comparisons made, it is not possible to fairly achieve the same noise value for each algorithm. While this fact introduces another variable for each experiment, it also accurately reflects the performance of each algorithm in a similar manner, and in doing so, moves towards reproducing the effects of human fatigue.

Analysis

The results of this experiment are displayed in Figure 3.7.

BBL sees a large decrease in accuracy relative to experiment 2 which is most likely due to a similar effect seen with previous noise patterns, where one out of place element may have knock-on effects for further elements.

RBN also sees a drastic decrease in accuracy. This decrease, together with the decrease in accuracy we observe with BBL is likely exaggerated due to the increase of noise over time. Both RBN and BBL require the largest number of comparisons which will result in a much greater amount of noise being introduced.

GSL, GSH and INS perform similarly with slightly lower numbers of comparisons and slightly decreased accuracy than experiment 2.

SWS remains relatively unaffected with a slight increase in accuracy of 15.851% relative to experiment 2. This reflects how the algorithm begins by pairing up similarly ranked values, which should achieve a good partial sorting quickly before the noise value increases resulting in values only slightly out of place.

GSR shows an increase in accuracy of 28.026% relative to experiment 2, but as with previous experiments, this increase in accuracy requires an increase in the number of comparisons of 226.774%. While the increase in comparisons may seem rather large, it should be noted that this remains the lowest number of comparisons of any algorithm in this experiment.

Finally, QCK shows the best accuracy and the third lowest number of comparisons — greater than GSR and SWS by 120.733% and 110.036% respectively. Compared to experiment 2, QKC also shows a marked increase in accuracy.

TABLE 3.7: Noise pattern parameters: Tail noise

Parameter	Value
Minimum value	0
Maximum value	500
Minimum percent	0
Maximum percent	120
Tuned to noise value	0.9

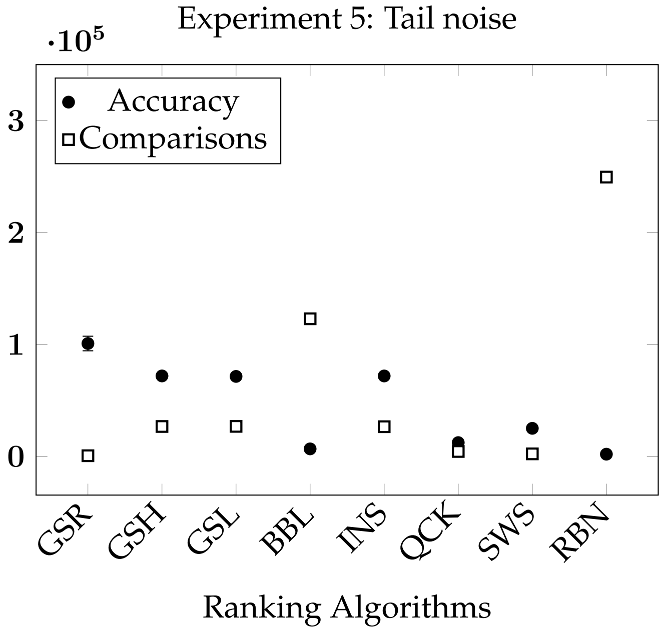


FIGURE 3.8: Accuracy and performance of each ranking algorithm subjected to `Tail noise`.

3.5.5 Experiment 5: Tail Noise

This noise pattern, similar to experiment 2, introduces a random variation of the values being compared. However, in this implementation, the variation of the values is dependant on how large those values are. This noise pattern aims to mimic the effect of highly ranked competitors who maintain a relatively steady performance between matches versus lower ranked competitors who may have wildly varying performance between matches.

A 'best performing' value, in this case the lowest number (1), will experience the lowest amount of noise (0%) while the 'worst performing' value (500) will experience the most noise (120%). Noise is distributed linearly across the values in-between. The maximum noise value of 120% represents the worst value of 500 varying by possibly 600. It is clear that, any noise above 100% is no more effective than 100%, and yet, to achieve a fair Noise Value between noise patterns, we increase the value to 120%, effectively maximizing the noise applied to values below 500 and increasing the overall Noise Value.

Analysis

The results of this experiment are displayed in Figure 3.8.

BBL handles this noise pattern relatively well compared to other noise patterns, performing similarly to the control experiment with a good accuracy but high number of comparisons.

RBN also achieves a high accuracy, but again is the worst algorithm with the greatest number of comparisons.

INS, GSH and GSL once again perform similarly, more accurately than experiment 2 — 29.208%, 29.579%, 29.955% respectively — but with significantly more comparisons — 149.207%, 137.479%, 141.228% respectively.

SWS continues to perform well in terms of accuracy but is outperformed on number of comparisons by both QCK and GSR. QCK, performing exceptionally well here with a very strong accuracy and very low number of comparisons.

GSR achieves the lowest number of comparisons of all algorithms but with the worst accuracy.

3.6 Discussion

Once again, the main question driving these experiments is whether Graphsort is robust to noise and whether it provides a suitable approach to ranking noisy data in comparison to other common approaches. Graphsort was designed specifically with a 2AFC study in mind, gathering data from human beings participating in a study where subject fatigue may affect the final result far more than any other factor. In addition to fatigue, the practical aspects of conducting a study of this nature necessitates an approach that can be implemented and executed quickly to minimize the time required of a participant in the study and of the researchers conducting the study. In these respects, analysis of algorithm performance is heavily weighted in favour of lowering the number of comparisons.

The accuracy of any algorithm is also less important in the context of a 2AFC study. This is predominantly due to the slight variance in personal preference between subjects, contextual issues arising from the order in which comparisons are made, and possible environmental issues such as ambient lighting, the time of day, the mood of the subject, and many more. The solution here is simply to accept a certain level of uncertainty about the correctness of any one ranking, increase the sample size of subjects and make use of statistical inference to estimate a population average.

3.6.1 Algorithm performance

We now discuss the performance of each algorithm under various noise patterns. In the interest of clarity and to avoid repetition, some similarly performing algorithms have been grouped together.

GSR. GSR performs well across all noise patterns with a low number of comparisons. Indeed GSR actually achieves a lower number of comparisons in the presence of noise than in the control experiment. With the exception of the control, GSR achieves the lowest number of comparisons of any algorithm for each noise pattern. GSR also achieves higher accuracy with some noise patterns, notably handling `Increasing noise` quite well. Considering the intended purpose of this noise pattern is to mimic human fatigue and the intended purpose of Graphsort is to handle this particular circumstance, this is a positive result.

GSH, GSL and INS. GSH and GSL perform similarly to each other across all experiments, which suggests that the impact of breaking higher ties versus breaking lower ties is negligible. GSH and GSL also perform strikingly similarly to INS, which may suggest that these algorithms may be slight variations of a generalized sorting pattern. GSH and GSL also perform worse than GSR in every experiment, both in terms of the number of comparisons and accuracy, with the exception of `Tail noise`, where GSH, GSL and INS outperform GSR in accuracy.

BBL and RBN. BBL and RBN consistently display the largest number of comparisons. This is expected due to the average case complexity of both algorithms. While the performance of these algorithms was perhaps predictable and unnecessary to verify, they do demonstrate a general inverse correlation between the number of comparisons and accuracy achieved. It is plain to see that with the exception of `Increasing noise`, a higher accuracy can be obtained using a less efficient algorithm. This finding may have particular utility in specific circumstances; where a noisy ranking environment is present, making a comparison does not increase the noise, a large number of comparisons is not a drawback and higher accuracy is necessary. A computer controlled biological sampling system may have similar requirements.

SWS. SWS performs incredibly consistently across all noise patterns including the control experiment. This consistency is certainly a major advantage, together with the accuracy it achieves with a low, and predictable number of comparisons. However, in contrast to other algorithms which produce a more homogeneous accuracy distribution, SWS ranks elements with less accuracy towards the center of the array than to either extreme. We suspect the elements ranked closer to the center of the array may be ranked with far less accuracy than our results show. In this respect, SWS cannot be recommended for any application where a homogeneous accuracy is required.

In addition, the particular function used to calculate accuracy may have a strong impact on these results. Further work is needed to investigate the homogeneity of accuracy which may involve comparing accuracy using alternative similarity measures or testing the accuracy of different segments of the arrays.

QCK. QCK would perhaps be the best choice for experiments 1, 2, 4, and 5 if there was less weight on the number of comparisons made. This algorithm achieves very good accuracy across all noise patterns, with a good number of comparisons for all noise patterns. QCK performs worst in the presence of `Flip noise` with a relatively high number of comparisons combined with a lower accuracy than in other experiments.

3.7 Chapter Conclusion

The results presented above show that GSR performs well in the presence of noise, achieving a reasonable accuracy with a low number of comparisons relative to other common ranking and sorting approaches. Further, a random tie breaking strategy is preferred over high or low tie breaking strategies to achieve the lowest number of comparisons. In terms of accuracy, a high or low tie breaking strategy is preferable only in the presence of `Tail noise`.

With a small increase in the number of comparisons, QCK can achieve a higher accuracy than GSR. In most circumstances, this would perhaps be a positive trade-off as long as the required number of comparisons are guaranteed to be made. Alternatively, GSR was designed to obtain a ranking based on human input where comparisons may cease at any point due to boredom or fatigue. GSR has a distinct advantage in this respect as a partial ranking is possible at any point.

As mentioned above, SWS performs consistently across all noise patterns. While SWS produces a symmetric heterogeneous accuracy distribution, GSH and GSL sort the highest and lowest elements of an array first, respectively. This produces an asymmetric heterogeneous accuracy distribution wherever a full ranking is not possible. In particular circumstances these heterogeneous distributions may be a preferable choice — and indeed a partial ranking is also possible at any point with these approaches — but building a model of human perception requires homogeneous accuracy across the entire ranking. Graphsort (GSR) has a distinct advantage in that no part of the array is favoured in terms of accuracy.

With respect to **H1**, the primary hypothesis being tested in this chapter, we aim to disprove the assertion that aesthetic data can be gathered in a way that minimises participant effort. The results presented show that a number of algorithms — QCK, SWS and Graphsort in particular — are capable of ranking subjective responses and therefore capable of gathering aesthetic data. These algorithms are all capable of ranking pairwise comparisons with less comparisons than the theoretical maximum which is often required in a 2AFC study.

While a number of algorithms may be somewhat useful in this respect, other design considerations have been highlighted. When contextual issues, cognitive and memory load and fatigue or boredom are brought into consideration, the emphasis on minimising participant effort is clear. In this respect, out of all the algorithms tested, Graphsort most effectively minimises the number of comparisons required between objects, and therefore the effort required of a participant in a study. We must therefore conclude that **H1** is true.

This conclusion provides a foundation for the verification of **H2** which will require the application of Graphsort to gather data outside of a simulated environment. Aesthetic attributes in the chosen domains of music and visuals may now be measured using the Graphsort algorithm.

Chapter 4

Gathering Aesthetic Data with GraphSort

This chapter is based upon research that has been published in part in the following conference proceedings:

- Aidan Breen and Colm O'Riordan. "Capturing and Ranking Perspectives on the Consonance and Dissonance of Dyads". In: *Sound and Music Computing Conference*. Maynooth, 2015, pp. 125–132

4.1 Introduction

In the previous chapter we demonstrated how the *Graphsort* algorithm can be effectively used to gather data in the presence of noise. In this chapter, the algorithm is used to gather aesthetic data in the domains of music and visuals. This aesthetic data will form the core knowledge-base upon which an Artificial Intelligence or Machine Learning system will be built. We begin by presenting our approach and methodology for gathering data, first in relation musical data, followed by visual data.

In both domains, there are many aspects and potential dimensions, or attributes, of aesthetic data which may be gathered. As we will discuss in later chapters, it is useful to chose an aspect that is common to both domains. We have therefore chosen to measure aesthetic data relating to musical and visual harmony. Both music and visuals can be described in terms of harmony, though the nature of harmony in each domain is varied. The relevant details of music and visual harmony are described in their respective sections 4.2 and 4.3 below.

Chapter Layout

The layout of this chapter is as follows. Subsection 4.1.1 describes the relevant research questions and hypotheses that are tackled in this chapter in detail.

The gathering of musical and visual aesthetic data is described separately in Sections 4.2 and 4.3 respectively. Each of these sections are structured identically. We begin each section with an overview of previous studies in Subsections 4.2.1 and 4.3.1. We follow this with the details of experiments carried out in Subsections 4.2.2 and 4.2.2, in which we describe our methodology and results.

We present a discussion of the two sets of experiments together in Section 4.4 including a comparison of results for both sets of experiments and a comparison of our results to previous studies.

The chapter is concluded in Section 4.5.

4.1.1 Research Questions and Hypotheses

The reader is once again reminded of the first research question of interest in this work, described in Section 1.2.2 as follows: *Can a computationally intelligent system be built to create aesthetic analogies?* Of particular interest in this chapter are the second and third hypotheses associated with this question:

H2 Aesthetic attributes in the chosen domains of music and visuals can be objectively measured.

H3 Aesthetic data can be gathered in the chosen domains.

The second hypothesis, *H2*, has been described in detail in Section 3.1.1 and in Chapter 3 we conclude that the Graphsort algorithm provides a foundation to objectively measure aesthetic attributes in the chosen domains of music and visuals. While the work in in Chapter 3 demonstrates that this is possible in a general sense, the work in this chapter serves to demonstrate the practical implementation of the approach in a specific domain and in doing so, aims to verify *H2* and *H3*.

4.2 Musical Data

We begin with the domain of music and musical harmony as an aesthetic data attribute to be measured. The concept of musical harmony has been introduced previously in Section 2.3.2, and is now described in further detail.

The concept of harmony is important in any artistic domain, but its effects are never more obvious than in music. In music, with the absence of any tangible art piece, the harmony of a piece of music is the primary aspect that influences our aesthetic experience. It should be no surprise then, that musical harmony has been the subject of much research and philosophical discussion throughout musical history.

In this work we discuss musical harmony in a non-temporal context. By this we mean that harmony is measured instantaneously and the duration of notes or the harmony of any preceding or succeeding music will not influence the harmony at any instant.

In most discussion, musical harmony is used synonymously with consonance or dissonance. A set of musical notes in harmony, are said to be consonant, and conversely those that are not in harmony are dissonant. Consonant music is often described as pleasant sounding or agreeable but that is not to say that all dissonant music is unpleasant. On the contrary, dissonant musical phrases are quite desirable in some genres like jazz. In some older texts, dissonance in music is also described as "roughness"[179], possibly referring to the physical perception of lower frequency oscillations which can be felt as the higher frequencies of notes fall in and out of alignment. This phenomenon is also known is "beating".

While music theorists are in general consensus about the relationship between harmony and consonance, at an individual level the perception of harmony is entirely subjective. While certain notes can produce distinctively consonant or dissonant sounds, the perception of harmony, and what is perceived as pleasant, can vary from person to person. A commonly held theory is that older people tend to have less aversion to dissonant sounds. While there is some evidence to suggest that infants prefer consonant over dissonant music [169], there are no strong indications that the psychoacoustic mechanisms at play change over the course of a human lifetime.

In the following section we present an introduction to previous studies on the perception of consonance and dissonance of dyads[1]. Later, we compare results obtained using the Graphsort Algorithm with the results of these previous studies.

4.2.1 Previous Studies

There have been numerous approaches to modelling the consonance of musical intervals, both mathematically and experimentally based on trials involving human subjects.

Helmholtz wrote at length about the topic of roughness [179]. His observations are often cited and reflect more recent empirical studies but were based on his own musings and the musical theory of the 19th and early 20th centuries. However his work has lead to much discussion and further research in the area.

---

[1] A dyad is a musical term for two notes that are played at the same time, similar to the more common term of triad, for three notes played simultaneously.

Malmberg published a study which tested 1045 subjects in three distinct groups between 1913 and 1915 for their preferences of consonance [111]. Subjects were played two pairs of tones – dyads, "or two-clangs" in Malmberg's words – on a piano or violin and asked to select which they preferred in terms of consonance. A strong effort was made to ensure all subjects understood the concept of consonance as Malmberg noted "the fundamental reason for the great divergence in the ranking by experts and the consequent disparagement of the ranking of consonance and dissonance has been due to the failure to take common ground in the definition of these terms".

Malmberg's study was conducted on a scale that has not since been reproduced. However, he did not build a ranking solely based on the responses of his subjects. Malmberg's tests aimed to compare the responses of his 1045 subjects – the "empirical ranking" – to a "norm", a ranking which had been developed based on the responses of "eight observers [who] were carefully selected on the basis of their training and fitness for the work". The eight observers were originally required to produce independent rankings, of which the averages would become the "norm", however Malmberg notes that after an initial conference, the "discussion and mutual criticism was so stimulating and interesting that all the observers agreed to sit again and continue by the same method until all should agree and a unanimous verdict could be handed in as in the case of a jury". The results of this study appear unprincipled and highlight potential difficulties in obtaining agreement in this domain.

Though the results obtained from Malmberg's observers did not directly affect the responses of his empirical study, the order of presentation and the particular dyads compared by subjects were based on his initial flawed study which may have affected his final results.

Plomp and Levelt described their approach to investigating the effect of critical bandwidth — the sensory limitations of hearing — on consonance by having subjects rate pairs of tones (generated by sine-wave oscillators) on a seven point scale, consonant to dissonant [139] with 1 corresponding to most dissonant, 7 corresponding to most consonant. They mention that some subjects had to ask for a definition of "consonant" which they provided as "beautiful and euphonious" as it had be ascertained that "consonant, beautiful and euphonious are highly correlated for naive subjects." Obviously Plomp and Levelt conducted these experiments with a relationship between consonance and aesthetics in mind, which serves to highlight a point made by Malmberg regarding a failure to adopt a common ground on the definition of consonance and dissonance leading to disagreement in the domain.

Though the experimental set up that Plomp and Levelt used was rather inelegant — equipment had to be readjusted by hand between each test — their tests have a number of aspects in common with future studies. Firstly, as mentioned above the subjects in their tests were not all musically trained. Secondly, the tone pairs, generated for their tests were composed of simple sine waves which reduced the tonal complexity of the resulting sound presented to the subject. Finally, their subjects rated tone pairs on a graduated scale.

Plomp and Levelt generated tone pairs about a set of mean frequencies (125, 250, 500, 1000 and 2000 Hz) and used a separate sample group for each. The sample groups were reduced in size from 19, 22, 18, 11 and 18 to 11, 10, 11, 10 and 8 respectively when they removed subjects who displayed incoherent responses.

Kameoka and Kuriyagawa conducted an independent study on the absolute and relative consonance of dyads [83]. Like Plomp and Levelt, their sample groups included two groups and their sample tones were generated using sine-wave oscillators, similar to [139]. In contrast however, subjects were not asked to rank consonance and dissonance absolutely (as in[139]) but rather in a relative manner. Subjects were presented with two dyads (A and B) and asked to provide an answer on a five point scale, -2, -1, 0, 1, or 2 "according to the subjective distance in consonance between A and B. If B is more consonant than A, a plus sign and if more dissonant, a minus sign was given" [83].

The groups chosen by Kameoka and Kuriyagawa for their experiments are rather puzzling. The first, "audio engineers, who are regarded as ordinary people" and the other, "mixers of ... the Japan Broadcasting Corporation", later referred to as "specialists" [83]. It should be obvious to any reader that a person working in the field of audio engineering is far from an ordinary person in the context of their perception of musical consonance and dissonance. The authors proceed to mention that their "experiments were carried out with audio engineers, and the results in this paper should be interpreted as for ordinary people", a contradiction in itself.

It should be noted that Kameoka and Kuriyagawa used exclusively Japanese speaking subjects. This resulted in the concepts of consonance and dissonance being defined in an entirely different language, adding a further layer of complexity to Malmberg's notion of the definitions of these concepts and how that may affect a subject's response. Additionally, there may also have been a cultural influence which, as stated previously, may have a strong effect on a subject.

Hutchinson and Knopoff later produced a "formalism" for calculating the consonance of a pair of notes based on the work of Helmholtz, Malmberg and Plomp and Levelt.[79]. In their paper they discuss a number of formulas which serve to produce absolute values of consonance for different intervals. The comparison of their values to the results produced by Malmberg seem to show a large degree of correlation, however they fail to provide a direct assessment of their values by means of experiment.

It has been noted that there are issues with the reproducibility of both studies carried out by Kameoka and Kuriyakawa, and Hutchinson and Knopoff [113]. This would suggest that the methods they used were simply not informative enough to accurately describe the behaviour of their subjects.

Plomp and Levelt had their subjects rank note pairs in an absolute seven point graduated scale [139]. Malmberg based his ranking on the unanimous decisions of hand picked experts and then tested on large groups of subjects all at once, with a common test pattern [111]. The work of Plomp and Levelt, and Malmberg has been the basis of other papers attempting to produce a formulation of consonance, such as Hutchinson and Vassilakis [79, 176]. None of these studies have taken into account the contextual effects of juxtaposing particular note pairs or ranking intervals absolutely and the effects these approaches may have on results as described in Section 3.1.2.

It could be argued that successive dyads may be presented to subjects with a large enough interval to ensure short term memory does not effect the response. Indeed Plomp and Levelt adopted this strategy with a 4 second interval [139], however other studies do not follow this paradigm, for example Kameoka and Kuriyakawa who mention an interval of only 0.5 seconds [83].

A second issue that faced earlier studies in particular was the reproducibility of musical notes and tones. Without modern audio recording equipment, playing musical notes to subjects was only possible by the live performance of musical instruments. Malmberg went to great lengths to find reproducible tones, testing pianos, organs, violins and bottles filled with wax among other instruments. If the subjects had been tested individually, it would have been arduous on the musicians and unreliable in terms of producing identical note pairs due to the fluctuations in timbre and pitch over time due to human fatigue and temperature or humidity changes. With modern computer systems, it is possible to produce identical tones over and over again with almost no variation.

Another problem we have seen is how researchers may simply disregard incoherent, or contradictory responses from subjects [139]. We believe there is information to be gathered from contradictory responses and that rather than being disregarded, contradictions may actually provide a greater insight into the subjective preferences of test subjects, whether it be consonance, dissonance, aesthetics or any other subjective – or noisy – domain. Once again, it may have been the technological constraints of the time that forced researches to reduce the complexity of their analytical processes, but with modern technology it should be achievable to handle these cases.

4.2.2 Experiment

We now describe an experiment conducted to show the effective implementation of the algorithm described in chapter 3 for gathering aesthetic data relating to musical harmony. A sample group of 25 subjects were presented with the experiment using recordings of piano notes to dynamically produce dyads. Dyads were produced using notes across two octaves, from A4 (440 Hz) to A6 (1760 Hz). Only intervals from 1 to 11 semitones were tested — unison and octave dyads were excluded to avoid confusion. The single sample group included both musicians and non-musicians, males and females, aged 21 to 57. All subjects were brought up within the western music culture.

Method

Each subject was tested separately. Subjects were first presented with a guideline document which described the testing software, the format of the experiment and the definition of consonance as: "Agreement or compatibility between opinions or actions", and further: "When two tones tend to blend or fuse and produce a relatively smooth and pure resulting sound, they are said to be consonant". This second definition is based on the description by Malmberg [111].

Each subject was then presented with the software as shown in Figure 4.1. Dyads — described as "pairs" to subjects — were played by clicking sound icons at the top of the screen, and a preference was defined by clicking the appropriate buttons below. After each preference selection, a screen with the words "Click to Continue" was displayed in order to prevent accidental preference selection.

Each preference was presented in order to break ties as describe in 3.2. Testing finished when all ties had been broken and a full ranking was achieved.

Click icon to play

Which pair is more consonant?

FIGURE 4.1: A screen shot of the testing software. Users were asked to click on the red sound icons and select the appropriate pair using the buttons below.

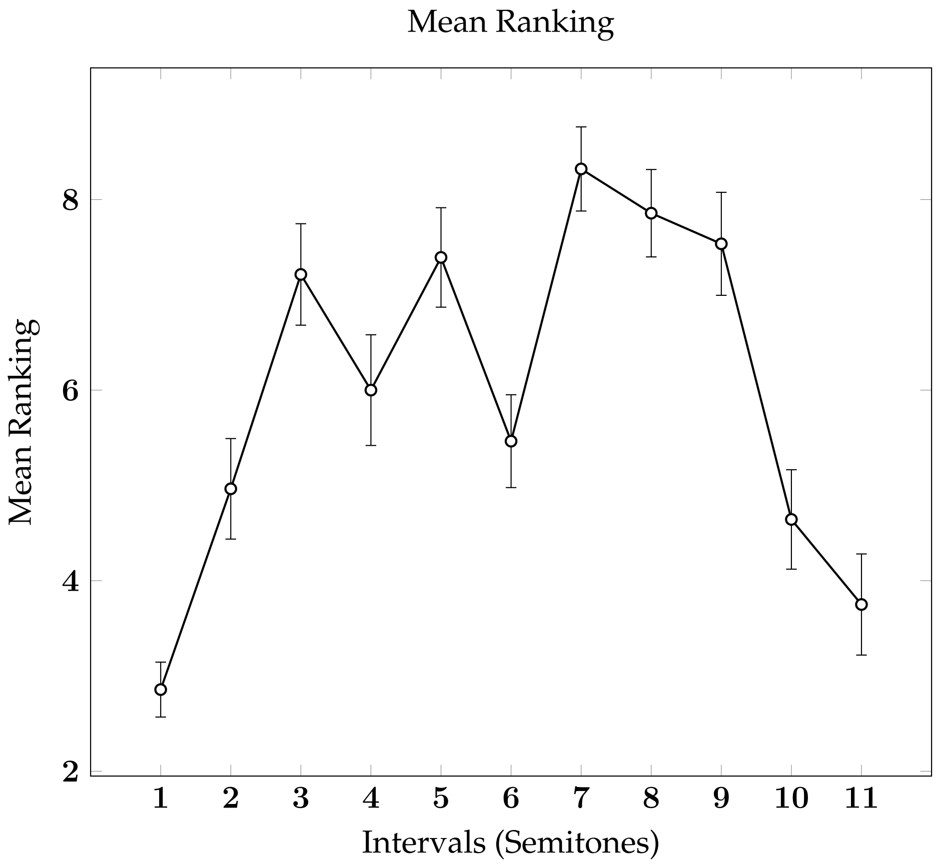


FIGURE 4.2: The mean ranking of intervals across a population of size 25 using an un-weighted digraph. Error bars show standard error of the mean.

Results

Subjects provided an average of 24.8 preferences before a full ranking could be calculated and completed the test in an average of approximately 6 minutes. This is in contrast to the minimum of 55 potential preferences required with a complete graph with 11 nodes ($\boldsymbol{V}$), given as: $\frac{\boldsymbol{V}^2-\boldsymbol{V}}{2}$ The sample mean ranking of each interval is shown in Figure 4.2. This ranking is compared to previous data — we use Hutchinson and Knopoff [79], and Malmberg as examples [111] — as shown in Figure 4.3 and Table 4.1, also see [83, 176] for other comparable results. It can be seen that our ranking follows a similar pattern to previous data where the intervals of 1 and 11 semitones tend to be less consonant (more dissonant, rough) and the interval of 6 semitones is conspicuously lower than its highly ranked neighbours of 5 and 7 semitones.

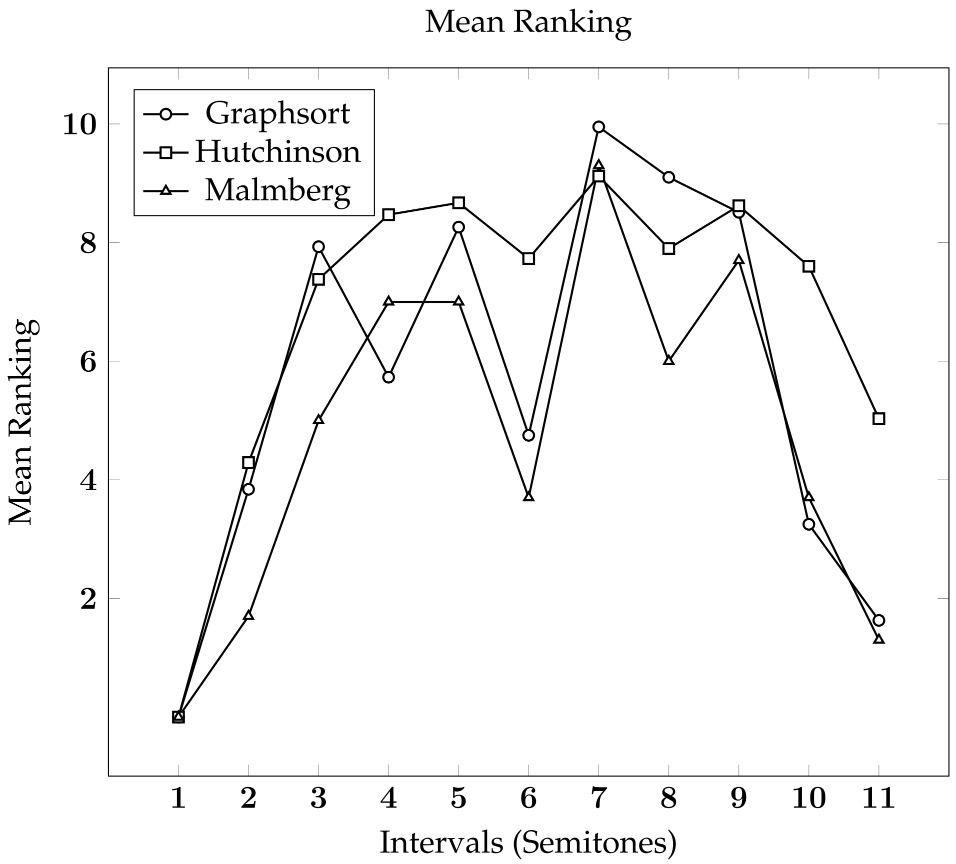


FIGURE 4.3: Comparison of results to previously published rankings [79, 111].

	Correlation Coefficient	P-value
Hutchinson & Knopoff	0.788	0.004
Malmberg	0.852	0.001

TABLE 4.1: Correlation of digraph rankings with previous studies.

4.3 Visual Data

We now introduce the domain of visuals and colour harmony as an aesthetic measure. Colour harmony has been previously introduced in Section 2.3.3 and is now described in further detail.

Visual art is a highly complex domain where minuscule details can have large impacts on the overall aesthetic quality of an image. Line, shape, scale, space and texture all contribute strongly and while general artistic guidelines can help, oftentimes it is the exception to the rule that impacts most on the overall artistic and aesthetic merit of an image. It is therefore difficult to measure the aesthetic value of an image by any one attribute, and even with many attributes in consideration, the task has not been reliably achieved. We must, however, begin with individual attributes if we are to successfully combine them later, and one of the strongest attributes available to us is harmony. In contrast to music, the aesthetics of a visual art piece may be less reliant on harmony. However, the effects of colour harmony cannot be disregarded.

Again, in this work we discuss colour harmony in a non-temporal context. The objective harmony of a visual displayed to a viewer will not be affected by the harmony of proceeding or succeeding visuals. The problem domain could be further simplified to remove the challenges faced by intermixing colours in various patterns in a two dimensional representation, however, in this work we have chosen to explore this aspect of the problem by comparing a number of visual configurations.

In contrast to musical harmony, colour harmony is a concept that most of us are quite familiar with. Most people believe they understand what colours go well together and are able to recognise which colours clash and why, but while colour harmony may seem to be an innate perception, the psychological mechanisms behind this phenomenon are not well understood.

As we have previously introduced, Burchette's 2001 review of 12 separate books concerning colour harmony "failed to demonstrate a general acceptance of any ranked list of terms" used to describe or define colour harmony[28]. This indicates that even as recently as 2001, no singular approach has been found to adequately describe the phenomenon. In a more recent review it is noted that "from a current theoretical perspective, many early definitions of colour harmony appear simplistic or overly generalized in nature; or seem to be based on unfounded opinion or unsubstantiated claims" [126]. This would suggest that our current theoretical models are becoming more accurate, tangible and indeed, implementable, though much more work is required.

In the following section we present some of the related work related to the measurement of visual harmony. Later, we will compare results obtained using the Graphsort algorithm to this work, and to the results obtained for musical harmony.

4.3.1 Previous Studies

In the proceeding section we discussed the previous studies relating to musical harmony. These studies are strongly based on theoretical models and empirical work is not common. In visuals however, with millions of images freely available in digital formats that are easy to manipulate, empirical approaches are far more common.

Indeed, an empirical model may prove to be more useful than purely theoretical approaches. There are two main approaches in this regard. Firstly, employing machine learning techniques to find common colour combinations used in aesthetically pleasing images. Secondly, conducting a study of human subjects to compare various colour combinations. We begin by discussing machine learning based approaches.

Nishiyama et al. utilise a support vector machine to categorise images into aesthetic groups. They make use of a public database of images and rankings (DPChallenge [32]) with crowd-sourced metadata [124].

Lu et al. make use of a large colour harmony evaluation dataset (CHE-dataset), consisting of 29,844 images with associated aesthetic meta-data. They use a latent Dirichlet allocation and Gaussian mixture model to discover a selection of categorised colour palettes. These palettes are then used to identify aesthetic images [108].

The work of Nishiyama et al. and Lu et al. seems to suggest that aesthetic images tend to make use of common colour palettes, although they do not provide any insight into the specific relationships between those colours.

One potential downside to machine learning techniques is a reliance on the underlying dataset. For example, the aesthetic visual analysis dataset (AVA [122]), of which the CHE-dataset is a subset, provides 250,000 images along with rich meta-data including multiple aesthetic scores for each image. The information available in this dataset is incredibly useful in training AI systems, however the authors provide no information on the source of the meta-data, when it was obtained, and from what cultural context. The DPChallenge dataset proves more problematic as it is continuously changing based on public use of the the website. Making the comparison of different algorithms difficult.

Studies of human subjects are typically more difficult to conduct — suffering from all of the same challenges musical studies would face — but lead to results less prone to issues related to over-fitting, changing datasets and misclassified data points. These studies also tackle the problem from a different viewpoint which may serve to provide better insights into the problem domain.

Chuang et al. investigated the relationship between 46 pairs of points within a CIELAB uniform colour space and their effect on perceived colour harmony. The study found subjects to be extremely unreliable in their responses. However, they still report that "the color interval of lightness may be the dominating factor with respect to the influence of (perceived) color difference" [36].

Schloss et al. conducted a number of experiments testing the aesthetic responses to colour combinations with particular focus on the distinction between a subject's preference for a colour combination (whether they liked the colours or not) versus the colour harmony of a combination (whether they felt the colours went well together) [154].

In their first experiment, focusing on preference, Schloss et al. find that "pair preferences are highest when the figure and ground have the same hue (but differ in saturation and/or lightness levels)". This finding is repeated in their second experiment focusing on colour harmony, with the relation between hue offset and harmony being more pronounced. The variance between experiments is attributed to the influence of a subject's preferred colours which may be present in a colour pair. With subjects instructed to rate harmony rather than preference, the effect is reduced.

Szabó et al. developed a series of mathematical models of colour harmony for both two-colour and three-colour combinations [165]. Of particular interest here is the influence of colour hue for two-colour pairs where "hue difference was ... found to be a relevant factor of colour harmony impression".

4.3.2 Experiments

We now describe experiments conducted to show the effective implementation of the algorithm described in chapter 3 for gathering aesthetic data relating to colour harmony and to test the impact of hue distance variance on colour harmony. A sample group of 30 subjects were presented with five separate experiments in various configurations using a computer screen to produce a visual display. The single sample group was made up of both males and females, ranging in age from 18 to 67. No subjects had any colour vision deficiency.

Method

Each subject was tested separately. Subjects were first presented with a guideline document which described the testing software, the format of the experiment and the definition of colour harmony as "how appealing two colours are in combination or how well colours go together". In contrast to our previous experiment, all subjects understood the concept of colour harmony and did not require any further explanation or clarification.

Each subject was then presented with the software as shown in Figure 4.5. The method by which colour pairs were displayed varied between experiment configurations. These methods are described in detail below. A preference for either colour pair was defined by clicking the appropriate buttons below the displayed colour pair. After each preference selection, a screen with the words "Click to Continue" was displayed. This prevented accidental responses. When the experiment was over, a screen with the text "End" was displayed. No further action could be taken by the subject.

The colours shown to the user were varied only in hue using the HSL colour model. The lightness and saturation were kept constant at 75 percent. This study aimed to test only the impact of hue distance on colour harmony, so each 'object' ranked by subjects represents the hue distance, or angle between two hues on a continuous colour wheel. We reduced the granularity of this experiment to include only hue distances in multiples of 30 degrees, from 30 to 330 degrees. Hues of 0, or 360 degree offsets were not tested as this would represent a single hue.

We now describe the specific configurations of each experiment and the motivation for that configuration. Visual examples are shown in Figure 4.4.

Experiment 1 Colours were displayed in configuration 1, a square, halved vertically with a colour in each side, ordered randomly. This configuration was selected to represent the most basic arrangement of two colours, similar to two coloured objects side-by-side.

Experiment 2 Colours were displayed in configuration 2, a square grid of 10 by 10 small coloured squares arranged randomly, with vertical symmetry. This configuration was selected to investigate the effects of mixing colours. The random pattern of squares was used to prevent any possible shape associations. Vertical symmetry was used to ensure a reasonably pleasant display without any distracting asymmetries.

Experiment 3 Colours were displayed in configuration 3, a split figure-ground configuration, similar to [154], with a large square of one colour and a smaller, inset square of the other. However, both squares were split vertically similar to *Experiment 1*, and in opposite orientations. This configuration ensured a uniform crossover of colours without an over-complex or distracting pattern.

Experiment 4 Colours were displayed in similar configuration to *Experiment 3*, however, the left colour of each figure was the same and remained constant for the entire experiment. This configuration was selected to investigate if the position of a colour had any large impact on harmony in comparison to *Experiment 3*.

Experiment 5 This experiment was conducted to test the effect of memory load on subjects. Identical in configuration to *Experiment 1*, except that subjects had to click a button to view each colour combination, and so could only view one combination at a time. Subjects could view each figure as often as they pleased.

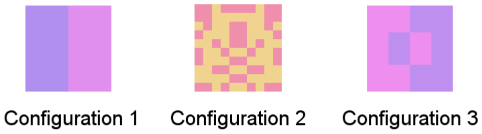
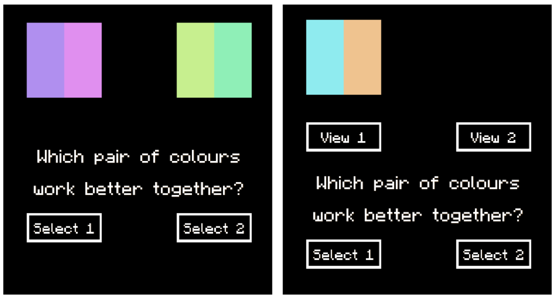


FIGURE 4.4: Various colour combination configurations.



Which pair of colours work better together?

Which pair of colours work better together?

Experiment 1

Experiment 5

FIGURE 4.5: Comparison of user interfaces used for experiments 1 and 5

Results

Subjects ranked 11 hue distances, in 5 separate experiments, with an average of 23.572 comparisons per experiment, producing a full ranking (no tied nodes) in an average time of 82.26 seconds. In comparison to the upper limit of $(11^2 - 11) \times 5 = 550$ comparisons for a standard 2AFC study with 5 configurations, this is a significant reduction. No subject required a break, or reported fatigue or boredom.

For each of the experiments, we record subject preferences for each pair of objects and generate a ranking of the objects based on the subjects' given preference. We then plot the ranking versus the offset in hue in order to determine any existing correlation.

Rankings calculated as a result of data gathered in each experiment are shown in Figures 4.6 and 4.7. We now present and discuss the results of the individual experiments as well as results obtained by combining rankings from all experiments.

Experiment 1 The graph of average rankings shown in Figure 4.7 shows no obvious trend. We believe this is an effect of the configuration used. Because colours were displayed side by side, subjects may have found it more difficult to truly perceive the colours in harmony. Indeed, Chuang et al. used a spiral configuration in order to avoid similar issues.

Experiment 2 We see a relatively strong trend here in favour of colours of increasing hue contrast. This experiment also showed the highest variation between highest and lowest average rank, indicating the population was in agreement using this configuration more than any other. We did not find this result surprising, as subjects often mentioned that they found this experiment the easiest, or most interesting to complete.

Experiment 3 In this experiment we also see a strong trend, far stronger than the similarly configured *Experiment 1*. The only difference being the intermixing of colours. In comparison to *Experiment 2*, we see less variation between maximum and minimum average ranks, indicating that the population was in less agreement about rankings. With these comparisons in mind, we believe the intensity of perceived colour harmony (at least in terms of hue) may be affected by the degree to which those colours are intermixed.

Experiment 4 This experiment was conducted in order to test if varying a single colour had any effect on subjects responses, rather than varying both colours. There seems to be no evidence that this has any effect on subject responses. The data seems to correspond roughly with that of *Experiment 2* and *Experiment 3*.

Experiment 5 This experiment tested the effect of memory load on subject ranking. By displaying only one figure at a time, we simulated the memory load of a 'Likert scale' type study. The data shows no discernible trend, with minimal variation between maximum and minimum average ranks. Perhaps a larger study would allow a stronger population trend to appear, however at this scale, the benefit of reduced memory load is clear.

Combined Experiments The results displayed in Figure 4.6 show the combined average ranking of hue distances for all 5 experiments. This, in effect, represents a combined population of 150 rankings under varying configurations. We see a very strong trend towards colours of greater hue distance, much clearer than any individual experiment shows. This indicates that there is indeed a correlation between perceived colour harmony and hue distance for colours of similar lightness and saturation.

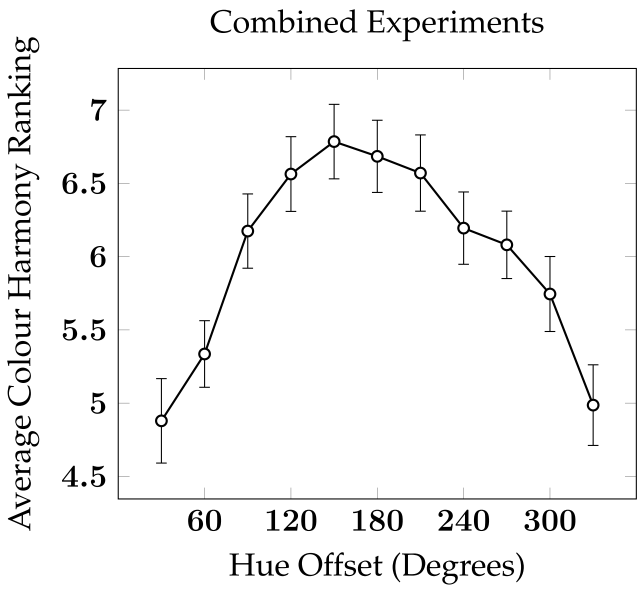


FIGURE 4.6: Combined rankings of all 5 experiments. Error bars represent the standard error of the mean.

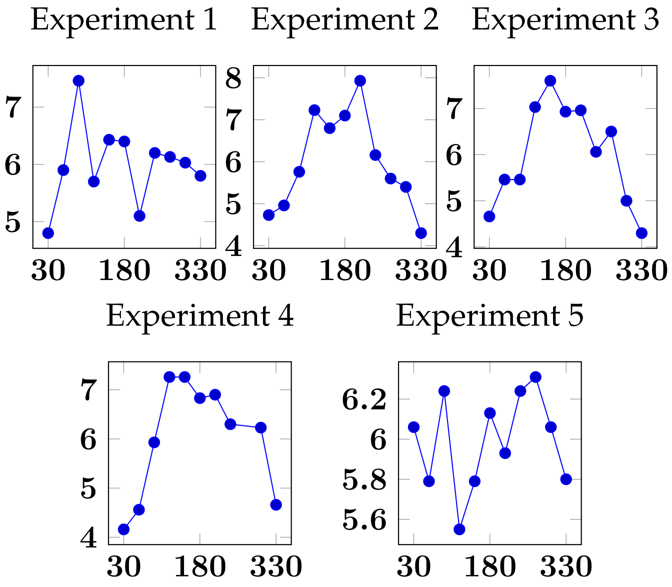


FIGURE 4.7: Average rankings from individual experiments.

4.4 Discussion

The sample populations for both the musical and visual experiments were relatively small. However, it is particularly clear in the musical experiment that the results obtained show a strong resemblance and correlation to previous studies.

Comparable studies in the visual domain have not been found. Colour hue certainly has an impact on colour harmony [165, 36, 154], however no studies have looked at hue in isolation of lightness and saturation. Studies in this area tend to test an entire colour model, with hue, saturation and lightness, in the HSL model for example, all representing possible distinct dimensions.

While an increase in hue distance seems to have a stronger effect on colour harmony, it is generally shown in other studies to have a negative correlation. Our study also found that an increase in hue distance had an effect on colour harmony, however in isolation of any other factor, a greater hue distance seems to increase colour harmony.

One possible explanation for this result might be that while in general, increased colour hue distance has an adverse affect on colour harmony, this is not true for all values of saturation and lightness. Further, the specific saturation and lightness values used in this work may represent values that invert the effect of colour hue on harmony. Further research is required to determine if this is correct.

It should also be noted that the colour model used may not be perceptually uniform and that some colours may be perceived as lighter or more saturated even with the same lightness and saturation values. We aimed to reduce this effect as much as possible by choosing lightness and saturation values of 75% which seemed to minimise perceptual changes. Furthermore, while the hue offsets are fixed, the compared hues were randomly generated which will further reduce this effect.

A similar observation can be made in relation to our musical results. Previous work has used live musical performance, recorded musical tones and pure sine waves generated by a tone generating device. Each of these methods produces a sound that varies in tonal complexity. This tonal complexity may have an impact on the overall consonance or harmony of dyads. Further work is required to demonstrate potential effects.

In both musical and visual experiments, the notion of harmony was explicitly defined for subjects before they provided any response. This step was taken because each subject may have a different idea of what constitutes harmony, an issue that has been mentioned in previous work. In our work we have noted that no participant required clarification of colour harmony. This is in contrast to musical harmony which had to be clarified to almost every participant. This might suggest that we have a more innate understanding of colour harmony, but it may also simply reflect the level of exposure a typical individual has to colour or musical harmony on a daily basis. We think about the colour of the clothes we wear, the cars we drive, the furniture in our houses and the knick-knacks we buy for our living rooms. Rarely do we consider the musical harmony of the noises we hear on a daily basis.

The results shown in both experiments represent harmony on a circular scale. Clearly a hue offset of 30 degrees is equal to a hue offset of 330 degrees on a continuous colour wheel. Similarly, a note offset of 1 semitone is similar to a note offset of 11. This concept is common in music theory and is the basis of a chord inversion. With this in mind, we would expect the graph of musical or visual harmony to be somewhat symmetrical. Indeed, this can be observed in the results presented above.

4.4.1 Experimental Improvements

Along with a larger sample size, a number of other improvements could be made to the experiments presented in this chapter.

Over-sampling

Our current approach of calculating ties and asking questions to break those ties does not allow subjects to specify cycles. Subjects may express cycles if given the opportunity and this information may be an important factor in the ability of any graph to accurately reflect the preferences of an individual. For example, some objects might be equally ranked — such as two musical dyads of equal harmony. This is currently not represented by Graphsort.

To improve our current approach and allow cycles to form we propose **over-sampling**: continuing to ask preferences even after all ties have been broken. A different method of ranking may be used to implement this and calculate new ties, or simple random sampling could be used to continue gathering data until the subject has had enough. We did not attempt to take this approach due to the possible impact of user fatigue over time.

Weighted Graph

By using a weighted graph it may be possible to both allow cycles to form, and fully rank objects, even when cycles are present in the graph. To achieve this we propose the following potential solutions.

In an un-weighted graph, cycles of more than 2 nodes may form, leading to many nodes with an explicit tie. In the case of a weighted graph, where a weight W $\in \boldsymbol{N}$ and 0<W<1, we make the assumption that in a tie, the edge with the lowest W can be safely excluded in order to break the tie and fully rank all nodes on a graph. In a cycle where there is no singular lowest W, we cannot break the tie and proceed with all nodes in the cycle tied, and of equal value.

This solution relies on the proper ranking of nodes within a sparse weighted digraph.

The weighted graph may provide a deeper expressiveness at the expense of the simplicity of implementation. A number of ranking methods exist for weighted directed acyclic graphs [7, 175] however the vast majority of research in the area assumes semi-complete or complete graphs and will not handle cycles. We therefore outline our own algorithm as one possible approach to this challenge.

Given a digraph $\boldsymbol{D} = (\boldsymbol{N}, \boldsymbol{E})$, a subset $\boldsymbol{N}_{\boldsymbol{x}}^{\boldsymbol{f}}$ (forward neighbours of $\boldsymbol{x}$, nodes that $\boldsymbol{x}$ points to), and a subset $\boldsymbol{N}_{\boldsymbol{x}}^{\boldsymbol{b}}$ (backward neighbours of $\boldsymbol{x}$, nodes that point to $\boldsymbol{x}$) we make use of a two pass approach as described below. A two pass approach is necessary to handle ranking imbalances which occur when nodes form disconnected branches, of differing weights, sharing a common root.

First pass

Populate the working set $\boldsymbol{W}$ with all end nodes (nodes with zero outgoing edges). For each node $\boldsymbol{n} \in \boldsymbol{W}$, move it to the visited set $\boldsymbol{V}$, give it a ranking $\boldsymbol{R}(\boldsymbol{n}) = 1$ and add all nodes pointing to $\boldsymbol{n}$, $\boldsymbol{N}^{\boldsymbol{b}}_{\boldsymbol{n}}$ to $\boldsymbol{W}$. To iterate through the entire graph, for each node $\boldsymbol{n} \in \boldsymbol{W}$ with forward neighbours ($\boldsymbol{N}^{\boldsymbol{f}}$) such that $\boldsymbol{N}^{\boldsymbol{f}}_{\boldsymbol{n}} \subseteq \boldsymbol{V}$, calculate the rank of $\boldsymbol{n}$ as

$$\boldsymbol{R}(\boldsymbol{n}) = \sum_{\boldsymbol{i} \in \boldsymbol{N}^{\boldsymbol{f}}_{\boldsymbol{n}}} \boldsymbol{E}_{\boldsymbol{i},\boldsymbol{n}} \boldsymbol{i} \tag{4.1}$$

where $\boldsymbol{E}_{\boldsymbol{x},\boldsymbol{y}}$ is the weight of the edge between nodes x and y. Continue until $|\boldsymbol{W}| = \boldsymbol{0}$.

Second pass

$\boldsymbol{V} = \boldsymbol{\emptyset}$. Populate $\boldsymbol{W}$ with all start nodes (nodes with zero incoming edges). For each node $\boldsymbol{n} \in \boldsymbol{W}$, move it to $\boldsymbol{V}$, keep its ranking the same as the first pass and add $\boldsymbol{N}^{\boldsymbol{f}}_{\boldsymbol{n}}$ to $\boldsymbol{W}$. Iterate through to the entire graph, for each node $\boldsymbol{n} \in \boldsymbol{W}$ where $\boldsymbol{N}^{\boldsymbol{b}}_{\boldsymbol{n}} \subseteq \boldsymbol{V}$, calculate the rank of $\boldsymbol{n}$ as

$$\boldsymbol{R}(\boldsymbol{n}) = \sum_{\boldsymbol{i} \in \boldsymbol{N}^{\boldsymbol{b}}_{\boldsymbol{n}}} \boldsymbol{E}_{\boldsymbol{i},\boldsymbol{n}} \boldsymbol{i} \tag{4.2}$$

Continue until $|\boldsymbol{W}| = \boldsymbol{0}$.

4.5 Conclusion

The results presented above demonstrate the use of Graphsort to gather data in the domains of music and visuals. A series of experiments have been outlined that demonstrate how the musical harmony of dyads and the visual harmony of colour pairs in various configurations can be measured using this approach.

The results of the musical experiment demonstrates that the variation of note intervals has an impact on perceived musical harmony, or consonance. The similarity between the results obtained in the music experiment and those of comparable previous studies demonstrates that the measurement made using Graphsort is objective, and reproducible rather than a subjective experience of any one person.

The results of the visual experiments demonstrates that the variation of colour hue has an impact on perceived colour harmony. This observation is clearly observed when the results of each experiment are combined, but analysis of individual experiment results shows that certain configurations of visuals can greatly impact the results.

The most dramatic effect is seen when memory load is increased by displaying only one colour pair at a time. This increased memory load results in completely unpredictable results.

Unlike the results of the musical experiment, the results of the visual experiment do not show any correlation to previous results. This is partly due to a lack of similar studies in this particular area. However, the results of each configuration of the experiment, where memory load is not introduced, are reproduced, showing a similar change in perceived colour harmony with a change in colour hue.

With respect to **H2**, we aim to demonstrate that *aesthetic attributes in the chosen domains of music and visuals can be objectively measured*. The results presented in this chapter show that the aesthetic attribute of harmony can be measured in each domain and that these results are reproducible and similar to previous studies, indicating that they are indeed objective. We therefore conclude that **H2** is true.

Further, with respect to **H3**, we aim to demonstrate that *aesthetic data can be gathered in the chosen domains*. The results obtained in all experiments are the product of aesthetic data gathered using the Graphsort algorithm. We therefore conclude that **H3** is true.

Chapter 5

Evolving Analogies

This chapter is based upon research that has been published in part in the following conference proceedings:

- Aidan Breen and Colm O Riordan. “Evolving Art Using Aesthetic Analogies: Evolutionary Supervised Learning to Generate Art with Grammatical Evolution”. In: *8th International Conference on Evolutionary Computation, Theory and Applications*. Porto, Portugal, 2016, pp. 59–68
- Aidan Breen, Colm O Riordan, and Jerome Sheahan. “Evolved Aesthetic Analogies to Improve Artistic Experience”. In: *6th International Conference on Computational Intelligence in Music, Sound, Art and Design*. Amsterdam, 2017

5.1 Introduction

In the previous chapter we demonstrated how aesthetic data can be effectively gathered in the domains of music and visuals. In this chapter, a system is introduced which uses this aesthetic data to create analogies between music and visuals. The system uses an evolutionary approach to map the aesthetic data from one domain to another.

Analogy is the comparison of separate domains. The process of analogy has strong applications in communication, logical reasoning, and creativity. A human artist will often take some source material as inspiration and create an equivalent, or related art piece in their chosen artistic domain. This process of metaphor is the equivalent of making an artistic analogy and has been used successfully in a literal form by artists like Klee [92], Kandinsky [84] and more recently Snibbe [162]. Similar approaches are often taken in a less direct form by stage lighting designers or film soundtrack composers.

One of the major challenges of computational art is to understand what makes an art piece *good*. Indeed the cultural and contextual influences of an art piece may define what makes it emotive, such as Duchamp's Fountain [30] or René Magritte's The Treachery of Images [110], but beyond that we rely on the aesthetics of an object to decide if it is pleasurable to perceive. The study of aesthetics provides an objective description of this perception. We use this objective description as a tool upon which to later build our analogies.

Our aim is to make computational analogies between the domains of music and visuals by making use of aesthetic models, computational analogy, and grammatical evolution.

We begin by presenting our theoretical approach to computational analogy using an evolutionary computation technique called Grammatical Evolution to evolve what we have termed *Mapping Expressions* which allow mapping aesthetic data between domains. These expressions have been designed to flexibly map attributes between domains while allowing for evolution to occur, gradually increasing the accuracy and performance of the mapping over time.

Evolution requires many generations or iterations of the mapping algorithms to be evaluated which can be time consuming if evaluation is computationally intensive. The format of mapping expressions, however, has been designed to allow for fast evaluation which both reduces the time taken to run a full evolution simulation, and allows expressions to be used in real-time systems without significant delays.

The evaluation of mapping expressions is only one small aspect of a real-time analogy making system, therefore we also present the details of our implementation of the system which enables the use of live audio input to produce a visual display in real-time.

Chapter Layout

The layout of this chapter is as follows.

Beginning with Subsection 5.1.1, the relevant research questions and hypotheses that are tackled in this chapter are presented in detail.

The analogical system is then presented in two separate sections. A Preliminary Implementation is presented in Section 5.2 focusing on the structure of the analogy, introducing the evolutionary approach, *mapping expressions*, and the fitness metrics used. A primary implementation is then presented in Section 5.3 which focuses on the practical aspects of the system, dealing with real-time musical input, evaluation of *mapping expressions*, and visual display generation.

Next, in Section 5.4 details of a study conducted to investigate the effects of the analogical system on aesthetic experience are presented. This is followed by a discussion in Section 5.5.

The chapter is concluded in Section 5.6.

5.1.1 Research Questions and Hypotheses

The reader is once again reminded of the first research question of interest in this work, described in Section 1.2.2 as follows: *Can a computationally intelligent system be built to create aesthetic analogies?* Of particular interest in this chapter are the fourth and fifth hypotheses associated with this question:

H4 An *analogical structure* can be created using the gathered aesthetic data.

H5 The analogical structure can be *evaluated* in *real-time.*

Analogical Structure refers to a formal definition of a structure between a set of attributes (or functions of attributes) within one domain to a set of attributes (or functions of attributes) in another domain. For example, we often use analogies to map the functions of two objects to better explain their operation — charging a battery is like filling a bucket with water — and in this analogical structure we are interested in the similarities of function between the two domains: the bucket, holding water, and the battery, holding electrical charge.

In this example we expect the observed attributes of each domain to function similarly, producing similar outputs: water will flow from a hose connected to the bucket until it is empty at which point it must be refilled. Electrical current will discharge through a wire connected to a battery until it is completely discharged at which point it must be recharged. There is an inherent structure to this analogy connecting the water bucket domain to the electrical battery domain. The function of flowing water is connected to the function of electrical discharge, and the outcome of an empty bucket to a discharged battery.

In audio and visuals, we observe a music piece in the audio domain and a 2-dimensional display of colours in the visual domain. The function of musical harmony affects the aesthetics of the music piece. The function of visual harmony affects the aesthetics of the visual display. These functions are analogically connected: if harmony increases, the aesthetic experience increases.

Evaluation of an analogical structure, in the context of *H5*, refers to the process to taking the value of an attribute within one domain and calculating the expected value of any connected attributes in another domain. In our water-bucket to battery analogy, this would be the equivalent of taking the flow rate of water, which should lead to a decrease in the amount of water in the bucket, and mapping this to a discharge rate of electrical charge, and calculating a decrease in the electrical charge of the battery.

Real-time evaluation of this analogical structure implies that the calculation can be completed without introducing a noticeable delay in the system. The system presented in this chapter could be described as a “soft real-time system”, which implies the time delay has rather large margins and is bounded by human perception. As opposed to a "hard real-time system" which must maintain calculation times below some physically determined limit often for safety purposes.

In addition, the reader is also reminded of the second research question of interest in this work, described in Section 1.2.3 as follows: *Does the analogy generated affect aesthetic response?* Each of the hypotheses associated with this question are explored later in this chapter:

H6 The use of aesthetic analogy has an impact on the *enjoyment* of an observer in artistic material.

H7 The use of aesthetic analogy has an impact on the *interest* of an observer in artistic material.

H8 The use of aesthetic analogy has an impact on the *fatigue* of an observer while viewing artistic material.

Enjoyment refers to an individual's overall satisfaction with an art piece. This can be seen as a combination of the aesthetics of that art piece, the individual's mood and any other external factors.

Interest refers to how interesting an art piece is to an individual. This is distinct from enjoyment as an individual might identify an art piece as interesting, but not have the inclination to give it the attention they might believe it deserves due to some other external factor such as mood or fatigue.

Fatigue refers simply to the tiredness or boredom of the individual. It is quite possible that an art piece that would be interesting and enjoyable at first, would be disregarded and disliked if the individual has been bombarded by similar pieces or has not rested.

The work presented in this chapter aims to verify these hypotheses, with Section 5.2 focusing on *H4*, Section 5.3 focusing on *H5* and Section 5.4 focusing on *H6, H7* and *H8* specifically.

5.2 Preliminary Implementation: Demonstrating Evolution

In this section we present a preliminary implementation of an evolutionary system for real-time aesthetic analogies. This implementation was carried out in order to demonstrate a proof of concept, that an analogical structure could be created using evolutionary computation and that such a structure could be used to take musical input to produce analogical visual output.

In particular, we explore the practical implementation of the evolution of the analogy structure and the challenges faced therein. We also present results showing the efficacy of the evolution system, together with brief observations on the visual output which is detailed in later sections.

5.2.1 Analogical Structure for Evolution

We now introduce the design of an analogical structure suitable for evolution, used to bridge the attributes of music to the attributes of visuals. This structure is the basis of a system to make use of analogies to computationally create art. The structure makes use of *mapping expressions* to connect aesthetic attributes from one domain to another. Figure 5.1 demonstrates a sample analogical structure with *mapping expressions* across two domains. The details of this structure are now introduced.

Every artistic domain has its own aesthetic measures — musical harmony, visual symmetry, rhythm and combinations thereof are simply functions of attributes in an artistic domain. In some cases, these measures can be used to describe objects in more than one domain. Symmetry, for example, can describe both a visual image, and a phrase of music. Though the specific way symmetry works in each domain may vary, we understand that symmetry describes a similar attribute, a sort of mirroring across some axis. In visuals, this axis is a line but in music, the axis may be a point in time.

In this chapter we build upon our previous work, ranking subjective preferences, and gathering aesthetic data, but with a focus on harmony. It has been demonstrated that musical harmony may be measured by the consonance or dissonance of musical notes and that visual harmony can be measured directly as the harmony of colour hues. Harmony in the domain of music is a function of the musical notes being played at any one instance. Harmony in the domain of visuals is a function of the hues of visible colours.

To bridge the gap between the two domains in a way that is well defined and suitable for computational evolution, we introduce the concept of *mapping expressions*, which are used to connect the attributes of musical and visual harmony. The operation of a *mapping expression* between musical and visual harmony is formally defined as follows:

*Given some musical input with harmony value $\boldsymbol{x}$, a **mapping expression** can be created to generate a visual output with a harmony value $\boldsymbol{y}$ such that $\boldsymbol{x} \simeq \boldsymbol{y}$.*

For this simple example, it is clear that a suitable expression could be created by hand with some knowledge of music and colour theory. However, if we extend the system to include more aesthetic measures, such as symmetry, intensity, contrast or granularity, defining an analogy by hand becomes impossibly complex, with the interactions of aesthetic attributes representing a strongly connected high dimensional complexity. The use of *mapping expressions* — which may be based upon a single attribute in each domain or multiple attributes in each domain — allows this analogical structure to be extended, and the use of Computational Evolution allows the inherent complexities of the system to be explored efficiently.

While implementing a system to capture more complex mappings is beyond the scope of this work, we aim to design a system that is a candidate for future extension that will allow applications to include complex, multi-attribute structures.

The analogical structure is shown in Figure 5.1 with measurable aesthetic attributes in separate domains which are connected by a set of *mapping expressions*. The implementation in this chapter uses a single attribute in each domain, however, the structure is not restricted to such a bijective mapping.

Many possible aesthetic attributes have been mentioned, but some aesthetic attributes may be more suitable than others for use in a structure as described in Figure 5.1. Harmony is selected for use in this thesis as it has a strong impact on the overall aesthetic

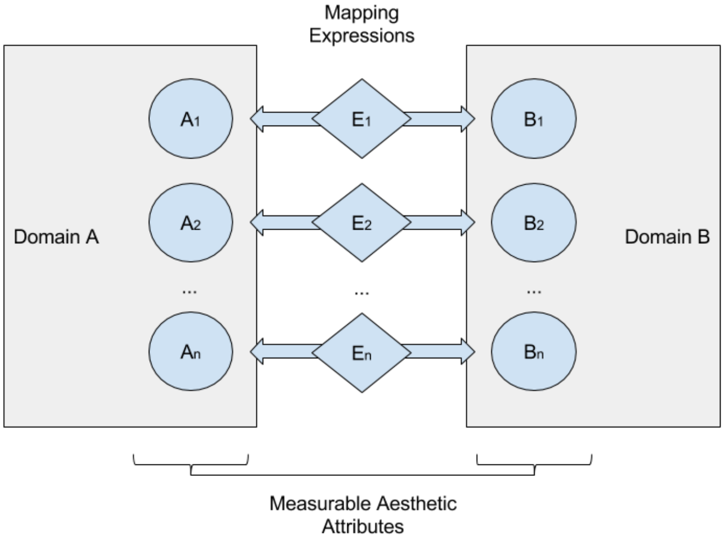


FIGURE 5.1: Analogy structure overview. The *Mapping Expressions*, $\boldsymbol{E_1}$ to $\boldsymbol{E_n}$, are encoded as chromosomes and evolved using the genetic algorithm.

quality of an art piece, and can be measured quite easily in separate domains, as demonstrated in Chapter 4.

We have discussed previously that in music, the harmony of notes being played is often referred to as the consonance — or conversely, dissonance — of those notes. While the timbre of notes has an impact on consonance, an estimate can be obtained from pitch alone. In the visual domain, colour harmony, or how pleasing a set of colours are in combination, can be measured as a function of the positions of those colours in some colour space. The conceptual similarities between musical consonance and visual harmony provide a convenient and understandable starting point.

Consonance values for any two notes have been measured as part of this thesis, and in previous studies [111, 83, 20]. However, when dealing with typical musical input, it is clearly possible to encounter larger sets of notes, or chords. For a system to be capable of handling real musical input, consonance values must be calculated for these larger sets of notes. Indeed there is a finite number of note combinations for which consonance values may be obtained, and numerous methods have been proposed that suggest measurement of consonance for larger note sets is possible [179, 139, 79, 176]. However, with a larger number of note combinations, the difficulty in gathering reliable data increases.

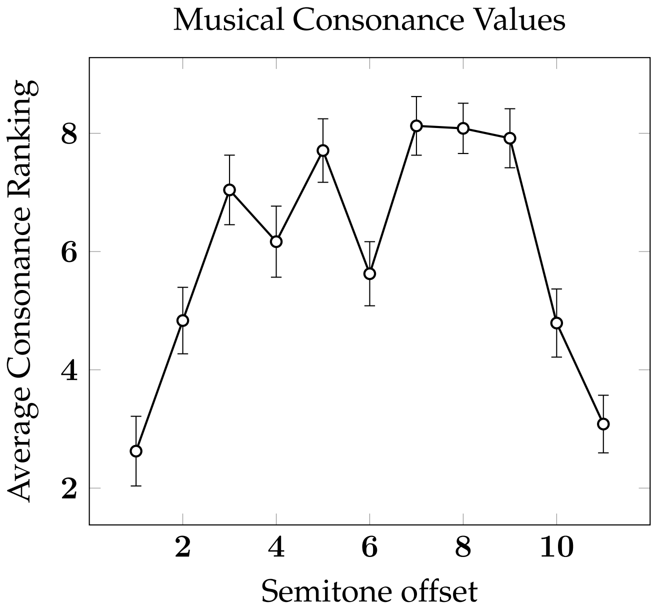


FIGURE 5.2: Consonance Values for musical intervals used to calculate musical Harmony Values.

Alternative approaches do exist which allow us to make use of existing consonance values for two-note pairs. The simplest general approach is to average the consonances for all pairs of notes in a set. This provides a good estimation for chords with different number of notes. We adopt this approach due to its simplicity, ease of implementation and use of two-note consonance values.

For our preliminary implementation, we enforce a number of restrictions. Firstly, we restrict the number of inputs to two musical notes at any one time. This removes issues with the consonance of larger note sets, simplifies the grammar used for evolution and allows us to more easily analyse the output *mapping expressions*. Secondly, musical harmony is calculated using just 12 note classes within a single octave. This helps to avoid consonance variations for lower frequencies.

Figure 5.2 shows the consonance values for note pairs used in this work, as reported in Chapter 4.

Similarly, colour harmony values can be measured and modelled [36, 165, 154]. While the harmony of more than two colours may be obtained with a similar approach to music chords, the pattern in which colours are displayed adds an extra level of complexity. To combat this, we assume our visual display is not a strict two dimensional image, but rather a pair of lights emitting coloured light into

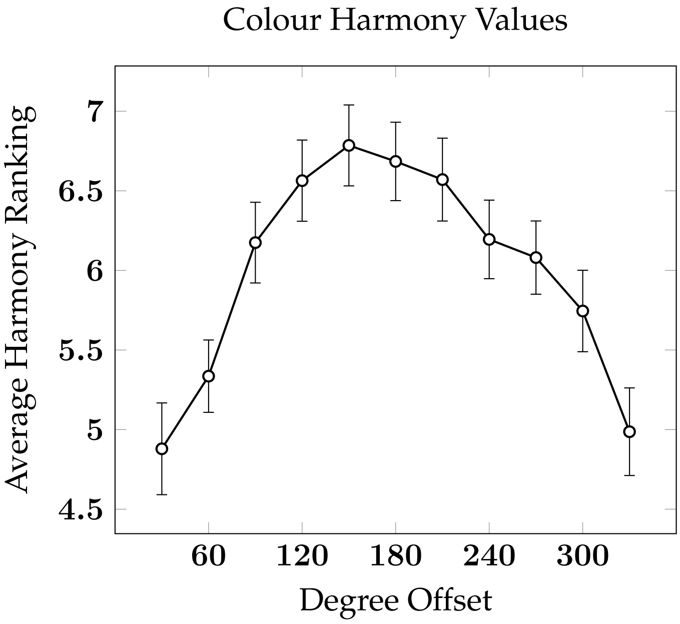


FIGURE 5.3: Average Colour Harmony values.

some space. For example, a pair of LED stage lights for a small musical performance.

The harmony values for colour pairs were obtained using the same methodology used to obtain musical data. A *two alternative forced choice* style study using the same ranking algorithm was employed to prevent subjectivity and contextual pitfalls and to reduce subject fatigue. The results were compared, once again, to previous studies and historical observations. Figure 5.3 shows the harmony values obtained in this study and used in the work presented here.

5.2.2 Evolving Analogies

Our evolutionary approach is based upon Grammatical Evolution (GE) as proposed by O'Neill and Ryan [128]. GE is a computational evolution approach whereby a grammar is used in a genotype-phenotype mapping. The application of this genotype-phenotype mapping enforces a stronger resemblance to biological systems where DNA codons are used to create proteins of a particular shape. In both systems, a many-to-one relationship occurs where many genotypes may produce one particular phenotype, which introduces a natural robustness while still allowing crossover and mutation to take effect. By defining a grammar, we can take any genotype and guarantee a grammatically correct phenotype. Using GE also allows the robustness of the many-to-one genotype to phenotype mapping to take effect.

GE provides a number of distinct advantages over other evolutionary approaches. Primarily, a GE system suits the creation of executable expressions which can be easily defined by a relatively small grammar. In comparison to other Evolutionary Programming systems, the implementation is relatively straightforward and simple to implement. The output of the GE system, in this case a lisp-like s-expression, can be parsed and executed simply and efficiently. This is of particular value in this work as a real-time execution of the expression is necessary when generating visuals with live music input. In this way, any particular performance is not limited to a strict musical input thereby allowing improvisation, timing and phrasing variation, and handling of human error.

Another advantage of this particular GA approach is the flexibility by which we can incorporate aesthetic data. By using a grammar, we are decoupling the structure of the output of the GA from the actual process of evolution. To extend the current implementation to include intensity as an aesthetic attribute, for example, we simply edit the grammar to accommodate this. We can also extend the grammar to include other operators, such as predefined functions if we feel their inclusion may improve the performance of the system.

Further, the human readability of the output expression is extremely valuable, not only as a sanity check to debug unexpected behaviour, but also to analyse the output of a particular evolutionary run. This is important to understand the process by which expressions are created. The readability of output expressions at each stage of the evolutionary process is therefore a huge advantage over other *black box* AI techniques which produce results without any way of understanding how or why they were produced.

Representing Mapping Expressions

Mapping expressions are created by using an individual chromosome to guide the construction of a **symbolic expression** by use of the given grammar. The following is an example of an expression representing a nested list of operators (addition and multiplication) and parameters (2, 8 and 5) using prefix notation. This is the same structure used by *mapping expressions*.

TABLE 5.1: Example Grammar

Non-terminals	**Possible replacements**
Parameter	Integer, (Operator Parameter Parameter)
Operator	$+$,$*$
Integer	Any integer

Terminals
$+$
$*$
Any Integer

```
(+ 2 (* 8 5))
```

The symbolic expression is evaluated from the inside out, by evaluating nested expressions and replacing them with their result. In this example, the nested expression multiplying 8 by 5 will be evaluated producing the following intermediate expression.

```
(+ 2 40)
```

Once all nested expressions are evaluated, the final expression can return a result, in this case, **42**.

Grammar

The grammar defines the structure of an expression using terminal and non-terminal lexical operators. Terminals are literal symbols that may appear within the expression, such as $+$, $*$ and the integers in the example expression above. Non-terminals are symbols that can be replaced. Non-terminals represent a class of symbols such as operators of a specific cardinality, other non-terminals, or specific terminals. In the example expression above, the $+$ and $*$ symbols might be represented by a single *operator* non-terminal that accepts two parameters. Parameters may be an integer, or a sub expression. Finally, the integers in the expression above could be replaced by an integer non-terminal. This leaves us with the simple grammar shown in Table 5.1 which is capable of representing not only the example expression, but also any other expression, of any size, that adds and multiplies integers.

Given a defined grammar, such as the example grammar in table 5.1, an expression can be built from a chromosome using the following approach.

Beginning with a starting non-terminal, each value in the chromosome is used in series as the index of the next legal terminal or non-terminal. This mapping continues until either the expression requires no more arguments, or a size limit is reached. If the chromosome is not long enough to complete the expression, we simply begin reading from the start of the chromosome again.

The values of a codon — a specific value within a chromosome — must have a fixed, uniform range so that evolution can occur. In contrast, the number of valid values that any codon may be mapped to will be varied. The solution is to distribute the valid terms across the codon range. There are a number of possible approaches here.

Firstly, a modulo based approach could take the remainder of the codon value, divided by the number of valid terms. This approach will ensure the terms are well distributed even if the codon range is small. However, this approach means that a slight change in a codon value, an increment of one for example, will result in a completely different phenotype.

Secondly, a roulette wheel style approach could define a range of codon values for each term. This approach makes for a more robust genotype to phenotype mapping, where codon values must change significantly before a phenotypic change happens. This approach is slightly more computationally complex, but because the grammar and codon value ranges will not change throughout the course of evolution, lookup tables can be used. This approach is adopted throughout this work.

Evolutionary Process

The evolutionary process is straightforward thanks to the simple structure of chromosomes in GE: an array of integers which map to valid terms in the given grammar. Each *mapping expression* becomes an individual possessing a single chromosome. Each individual is part of a larger population of individuals with varying chromosomes. Beginning with a population of individuals with randomly generated chromosomes, known as Generation 0, or $\boldsymbol{G_0}$, successive generations are produced using the Selection, Crossover and Mutation genetic operators.

Tournament selection with elitism is used in this implementation. At generation G_n, the fitness of each individual is calculated. A small number of the highest fitness individuals are moved to generation G_{n+1}. This elitism is used to prevent the maximum fitness of the population from declining. The selection phase now begins to populate the rest of generation G_{n+1}. Two individuals are selected at random. The individual with a more favourable fitness value is placed into generation G_{n+1}. The process continues until the desired size of generation G_{n+1} is reached.

When selection is complete, crossover begins using both single point and double point crossover. Two individuals, $P1$ and $P2$, are selected at random from the newly selected generation G_{n+1}, minus elite individuals. If single point crossover is to be used, a random crossover point is selected and two children, $C1$ and $C2$, are produced. These children take the place of their parents in the new generation.

$C1$ will receive a chromosome made up of the $P1$ codons to the left of the crossover point, and the $P2$ codons to the right of the crossover point. $C2$ will receive the alternative, that is, a chromosome made up of the $P1$ codons to the right of the crossover point, and the $P2$ codons to the left of the crossover point. If double point crossover is to be used, the same steps are taken, but twice. This results in one child receiving outer codons from one parent and inner codons from the other.

The effect of double point crossover is quite strong when using the described implementation of Grammatical Evolution. Due to the tree-like, recursive way in which an expression is built, single point crossover invariably has large effects on the final expression. Large portions of the expression are likely to be completely different, even to the parent expression, as previous codons determine the meaning of following codons. Double point crossover alternatively, can operate similar to subtree substitution, affecting only a small portion of the final expression and retaining the overall structure of the expression tree.

Finally, once generation G_{n+1} is complete, a mutation operator is applied. Mutation consists of setting a very small number of codons

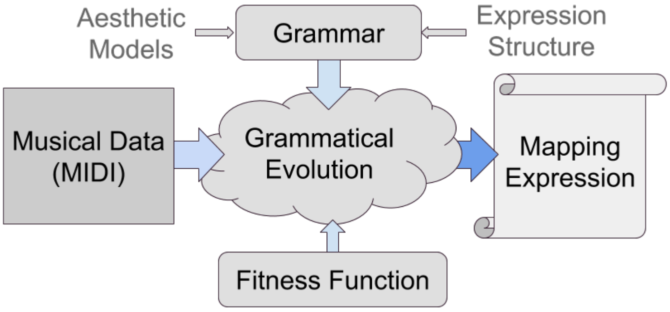


FIGURE 5.4: Evolution Phase overview.

TABLE 5.2: Genetic Algorithm Parameters.

Parameter	**Value**
Population Size	50
Chromosome Length	60
Crossover Rate	0.8
Standard Mutation Rate ($\boldsymbol{Mut_1}$)	See Equation 5.2
Hyper-Mutation Rate ($\boldsymbol{Mut_2}$)	1.0
Mutation Threshold	100
Halting Threshold	200

to new values. Both the mutated codon and the new value are randomly selected.

This process relies heavily on the fitness function. Calculating the fitness of any *mapping expression* without some guidelines would be extremely subjective. In the following section we describe our implementation, which takes a heuristic approach that rewards solutions that aims to maximise the similarity between input and output harmony.

Implementation of Evolutionary Process

Figure 5.4 shows the high-level structure of the system implemented for the evolution phase of experimentation. This phase is centred about the GE evolutionary approach. In this implementation of GE we use a population of 50 chromosomes. Chromosomes are stored as 8 bit integer arrays, representing values between 0 and 255. A chromosome length of 60 integer values was used.

Musical input is taken in the form of *Musical Instrument Digital Interface* (MIDI) data. The MIDI protocol represents digital music

TABLE 5.3: Grammar terminal operators.

Expression	Arguments
Plus 90 degrees	1
Plus 180 degrees	1
Sin	1
Cos	1
Log	1
Addition	2
Subtraction	2
Multiplication	2
Division	2
Music Harmony Constant	2
Visual Harmony Constant	2
Ternary Conditional Operator	3

TABLE 5.4: Grammar terminal values.

Expression	Range
Constant integer value	0-255
Musical input 1	0-255
Musical input 2	0-255

signals, originally designed as a transmission protocol to allow musical signals to be sent between instruments and synthesizers. Musical notes are sent as packet pairs (*note on* and *note off*) containing the note pitch and the 'velocity', or strength of the note which is often translated to volume. The MIDI protocol also allows data to be stored as a file with each packet containing a timing value. In this experiment, we use a file to store a sample musical input using this format and determine which notes are being played using the timing value.

The implemented grammar contains a list of operators, and values (variables and constants) which are presented in Tables 5.3 and 5.4. Of note here are the aesthetic values for music and visuals which can be inserted directly into an expression as constants, or read at run-time as variables in the case of musical input. These aesthetic values are normalised and based on the musical harmony, and visual harmony values shown in Figures 5.2 and 5.3 respectively. Aesthetic constant expressions such as *Musical Harmony Constant* and *Visual Harmony Constant*, accept two arguments representing two music notes or two colour hues. The expression returns the aesthetic value of those two notes or colours.

The fitness of an individual with any $\boldsymbol{n}$ pairs of input musical notes is calculated as follows, where $\boldsymbol{M}$ is a function representing the musical harmony of a pair of notes, and $\boldsymbol{V}$ is a function representing the visual harmony of a pair of colour hues.

$$\boldsymbol{fitness = \frac{1}{n}\sum_{i=1}^{n} 255 - |M(input) - V(output)|} \tag{5.1}$$

Both $\boldsymbol{M}$ and $\boldsymbol{V}$ are normalised between 0 and 255, which produces a fitness range of 0 to 255.

In this implementation, the fitness of an individual is calculated with 11 pairs of input musical notes, representing the 11 possible musical intervals in a single octave excluding the octave and unison intervals.

Tournament selection is carried out to select individuals for evolution. A combination of single point and double point crossover is used to build a succeeding generation. Elitism is used to maintain the maximum fitness of the population by promoting the best performing individuals to the next generation without crossover or mutation.

Mutation is applied at the gene level. A gene is mutated by randomly resetting its value. The mutation rate is the probability with which a gene will be mutated. The mutation rate is varied based on the number of generations since a new peak fitness has been reached. This allows us to optimise locally for a period, and introduce hypermutation after an appropriate number of generations without any increase in peak fitness. We call this the Mutation Threshold. The standard mutation rate ($\boldsymbol{Mut_1}$) is calculated as:

$$\boldsymbol{Mut_1 = \Big(\frac{0.02}{70}\alpha\Big) + 0.01} \tag{5.2}$$

where $\boldsymbol{\alpha}$ represents the number of generations since a new peak fitness was reached.

After the Mutation Threshold is reached, indicating a local optima, hyper-mutation ($\boldsymbol{Mut_2}$) is introduced to encourage further exploration of the fitness landscape.

$$\boldsymbol{Mut_2 = 1.0} \tag{5.3}$$

If a fitter solution is discovered, mutation is again reduced to $\boldsymbol{Mut_1}$ to allow smaller variations to occur.

At each generation, the *mapping expressions* represented by chromosomes are stored. This allows us to monitor the structure and size of the expressions as they are created. We can also use the stored expressions to compare earlier generation to later generations.

Evolution is halted after a Halting Threshold has been reached. The Halting Threshold is measured as the number of generations without an increase of peak fitness. Details of the parameters used can be found in Table 5.2.

The output of this process is a set of *mapping expressions* representing the final generation. From this set, the fittest expression is selected for evaluation and comparison to the fittest expression from previous generations.

5.2.3 Experiment: Fitness of Mapping Expressions

Method

The core function of the evolutionary system is to produce solutions that increase in fitness over time. This experiment demonstrates that solutions both increase in fitness with successive generations, and also that the solutions generated by the system are distinctly fitter than initial or randomly generated solutions.

After an initial phase of parameter tuning, 100 evolutionary runs were completed using the parameters shown in Table 5.2. The individuals in each generation were stored, with the chromosome values, generated expression, and individual fitness. Further data was stored about each generation including the maximum fitness of any individual, the average fitness of all individuals, and the size of fit expressions in the generation.

Finally, the fitness of both Generation 0 (randomly generated) and Final Generation (highly fit) individuals was stored.

Results

Using the approach outlined in Section 5.2.2 above we successfully evolved *mapping expressions* capable of mapping musical input to visual output.

Many of the random Generation 0 expressions such as the following example simply produced constant values:

```
['plus180',['plus90',['sin', 215]]]
```

In later generations however, we see more complex expressions producing better fitting results:

```
['add',56, ['musicalHarmony',94,['cos','mus2']]]
```

Here we see the expression makes use of the input variable `mus2` and the musical harmony constant `musicalHarmony` which produces a dynamic output. The example chosen here is one of the smallest expressions created.

Further visual analysis of the smallest evolved expressions shows regular use of dynamic input variables. A script was used to search through the entire set of final expressions from all evolutionary runs. The smallest expressions were selected for manual evaluation and visual inspection. The sampled expressions achieved high fitness scores, above 200 in all cases, while using less than 10 subexpressions. Leaf nodes of these sampled expression trees consist of at least one input variable in all cases. All sampled expressions also made use of musical or visual harmony constants at least once.

Figure 5.5 shows a comparison of 100 Generation 0, or randomly generated *Mapping Expressions* to 100 highly fit *Mapping Expressions* from the final generation of an evolutionary run.

We see the distribution for random expressions is heavily skewed towards the minimum value of 100. This is due to the number of expressions which produce a constant output. Evolved expressions however show a much tighter distribution with significantly higher fitness values.

Figure 5.6 shows the fitness of a single typical population across a number of generations in one evolutionary run. In blue we see the maximum fitness of each generation. This value never decreases due to the use of elitism. Highly fit individuals are brought from one generation to the next to prevent the population from decaying and losing a possible fit solution. In red we see the average fitness of the population.

During early generations we see dramatic improvements in maximum fitness (blue) followed by a series of incremental increases

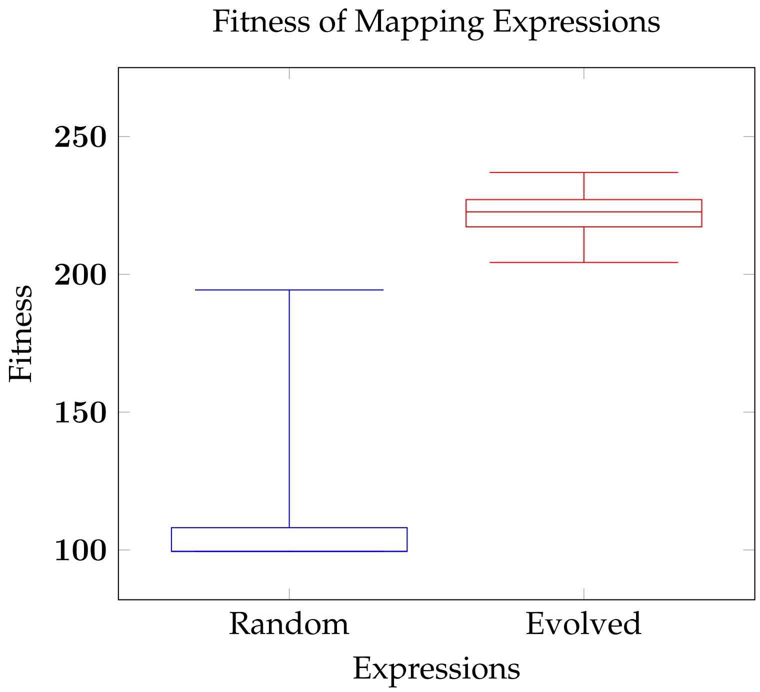


FIGURE 5.5: Fitness of 100 randomly generated *Mapping Expressions* vs Evolved expressions.

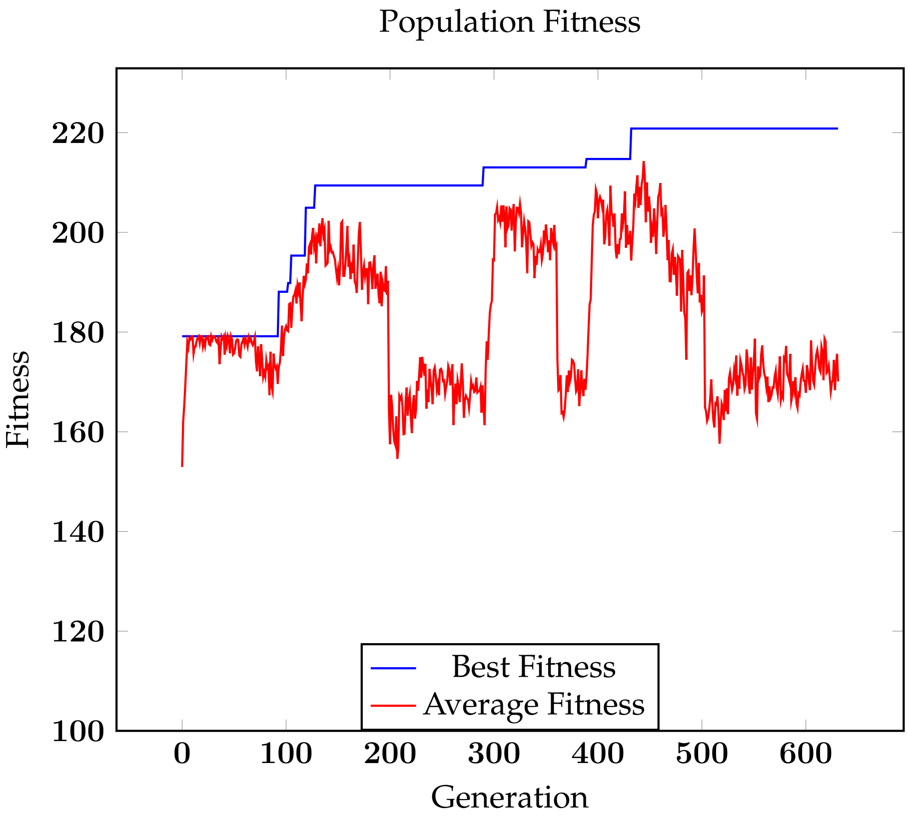


FIGURE 5.6: Population fitness for 631 generations of a typical run.

punctuated by dramatic decreases in average fitness (red). These incremental increases in maximum fitness are an indication of better local optima being discovered due to low mutation. Hyper-mutation is then introduced causing the dramatic reduction in average fitness as more individuals mutate in more extreme ways. While this seems to have a negative effect on the population as a whole, it allows us to find fitter solutions and prevents premature population convergence at a local optima.

5.2.4 Experiment: Effectiveness of Fitness Function

Method

An evolutionary system capable of producing solutions that increase in fitness over time is only as effective as the fitness function it utilises. This experiment demonstrates the effectiveness of the fitness function used.

A maximally fit *mapping expression*, when given a pair of musical notes, should produce a hue offset with a colour harmony equal to the musical harmony value of the input notes. That is to say: if we take the musical harmony at each interval $\mathbf{1 - 11}$, and evaluate a maximally fit *mapping expression* with these values, the output should be equal to the colour harmony values for the intervals $\mathbf{1 - 11}$.

We can use this as a guideline to visualise the effectiveness of the fitness function. As the fitness of a sample *mapping expression* increases, we would expect the output to more closely match the colour harmony curve shown in Figure 5.3.

Results

Previous generations are shown as dotted lines with the final fittest individual in solid black. The target output, the output that would produce an optimum fitness value, is shown in red. We see as the generations pass, the output matches the target more closely. Of note here are the horizontal dotted lines indicating older generations producing constant outputs which have been superseded by generations producing closer matching dynamic outputs.

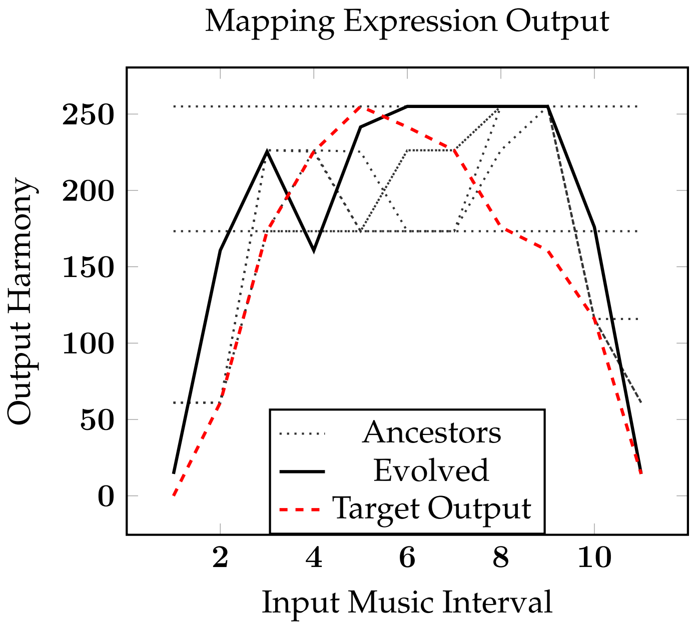


FIGURE 5.7: Output of one evolved expression and its ancestors compared to target visual harmony.

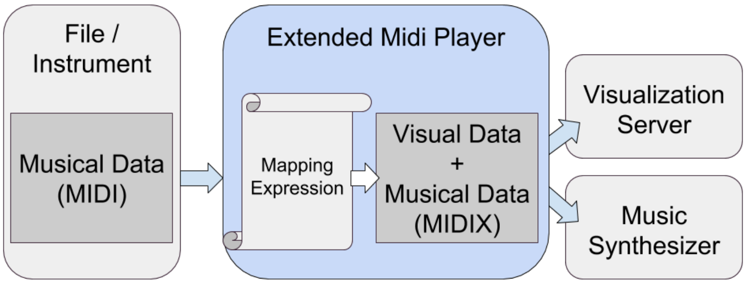


FIGURE 5.8: Overview of system design to produce visuals.

5.2.5 Experiment: Producing Visuals

Method

In order to evaluate the performance of an evolved *mapping expression*, we must play both music and visuals together. To this end, we have built the evaluation system as outlined in Figure 5.8. While evolved expressions are capable of generating visual output in real-time, the evaluation for this work is conducted offline, using a pre-calculated file containing the original musical data, and the generated visual data.

In order to perform music in synchrony with generated visuals, a system is required to send musical and visual signals to subsystems which output the musical and visual results. This synchronisation system is called the *extended MIDI player*.

Musical data (MIDI) and visual data are combined in an *extended MIDI file* (MIDIX). The *extended MIDI player* then parses this file and uses an internal time synchronisation process to send MIDI signals to two separate systems: a music synthesizer and a visualization server. Both systems are capable of reading MIDI signals and producing output in real time.

The music synthesizer is a common tool in music creation and performance. Historically, synthesizers existed as hardware devices connected to some digital instrument producing MIDI signals. The synthesizer would listen for input from a physical MIDI cable, and produce a signal which could be sent to an amplifier and speakers to create audio. Modern synthesizers are typically digital systems that listen to digital MIDI ports for message packets and produce audio using modern audio interfaces. While hardware based synthesizers are highly sought after by electronic musicians, software based synthesizers are indistinguishable in most circumstances.

We make use of digital synthesizers in this work due to their flexibility and cost effectiveness. We use a standard MIDI port to send and receive musical data to an open source software synthesizer, part of the Reaper digital audio workstation [37]. The signals received by the synthesizer are used to produce realistic sounding music.

The visualization server is a software system created specifically for this implementation. The server works in a similar fashion to a music synthesizer; however, rather than using a MIDI port, the server uses HTTP and websockets. With this approach, a webserver accepts HTTP messages sent from the *extended MIDI player* and uses websockets to update a javascript application running in a web browser. This technology was chosen due to the cross platform support for javascript applications and also due to the ease of rendering a visual display using standard web technologies. When the javascript application receives an update through the websocket, the display is updated accordingly. This ensures the visuals remain synchronized with the audio being played.

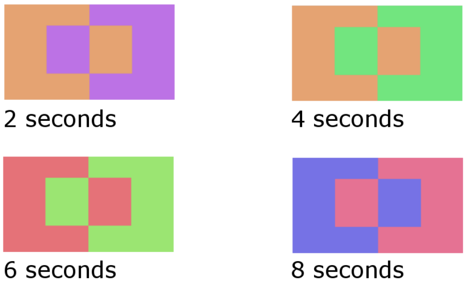

2 seconds
4 seconds
6 seconds
8 seconds

FIGURE 5.9: Generated visual display

Results

Preliminary results have been obtained based on the initial implementation. These results demonstrate that a visual display, however rudimentary, can be produced based on musical data in real time. The *extended MIDI player* was used to play a file containing a 10 second music piece with musical intervals of varying harmony. Visuals generated by a *mapping expression* were displayed on a computer screen. Visuals were observed to be in time with the synthesized music. An example of the visual display with screenshots taken at 2 second intervals are shown in Figure 5.9. The colour pattern used in the visual display was similar to that used to collect colour harmony data. This ensures that there is no variation between observed colour harmony and generated colour harmony.

Our initial observations indicate the the evolved expressions produce an analogy that results in more enjoyable visuals than randomly generated colours. The colours tend to follow a general pattern where similar note pairs, with similar intervals, produce similar colours.

These results serve as a proof of concept and are expanded upon in the following section.

5.3 Primary Implementation: Creating Analogies

In this section we examine the implementation and performance of an evolutionary system for real-time aesthetic analogies. We explore the practical implementation of real-time evaluation of evolved mapping expressions, possible musical input and visual output approaches, and the challenges faced therein.

5.3.1 Real-Time Musical Input

In Section 5.2.5 we introduced the design of a preliminary system capable of evolving aesthetic analogical structures using mapping expressions. A visual overview of the system is shown in Figure 5.8. This preliminary system reads musical input data from a MIDI file, passes it to the Extended Midi Player which evaluates a mapping expression to produce visual data. The musical and visual information is then sent in synchrony to a visualization system and music synthesizer system. The result is music and visuals being played and displayed in synchrony.

This preliminary system is capable of operating without any observable delays, however, it does require the input music file to be read into memory before any visuals can be generated.

In this Primary Implementation, we build upon the preliminary system by incorporating live musical input and produce real-time analogical visuals. There are no significant differences in terms of system structure, but differences in implementation can be found within the Extended Midi Player. We implement an input buffer, enabling musical input to be read by the system continuously.

We discuss in later sections the practical challenges surrounding musical input approaches, and these challenges impact the system design. The preliminary system was designed to accept Midi data, however, if we wish to accept live musical input, it would be useful to handle analogue audio signals, enabling the use of acoustic instruments.

The solution is a separate system which uses signal processing techniques to identify notes being played in analogue audio signals. This system may then send the identified notes in Midi format to the analogy system.

5.3.2 Experiment: Musical Input Approaches

Method

In order to test the effectiveness of an input system to read analogue audio and convert it to a format suitable for the analogy system, the following experiment was conducted, comparing input from sampled analogue audio signals and digital midi signals.

In the first approach using sampled analogue audio signals, a Fast Fourier Transform (FFT) was implemented to identify the frequencies being played in a live audio sample. In testing this approach, we identified the strongest frequencies to be the fundamental frequencies of notes being played on an instrument. The detected fundamental frequencies were then used to produce Midi signals, from which the harmony value was calculated.

In the second approach, using digital midi signals, we took advantage of the digital audio workstation (DAW) used to produce the original track. Because the music was composed within a DAW, Midi data was available for each instrument. The DAW could then be used to play the song in Midi format, passing the output Midi data directly to the analogy system where the harmony was calculated.

In both cases, the harmony value was calculated based on musical harmony values shown in Figure 5.2, using an average of note pair harmonies to get reasonable approximations of chord harmonies as described in Section 5.2.

The same musical input was used in both cases. The input was composed by the authors to be enjoyable without focusing on any particularly polarising genre and also to avoid any intellectual property issues. The input also displays a varied level of harmony and distinct parts of the song including a verse, bridge and chorus.

Results

Figure 5.10 shows the harmony values, frequency spectrum and the manually identified parts of the the song for the first approach using sampled analogue audio signals.

There are some obvious drawbacks to this approach. Chord recognition is a complex and actively researched field. Simply taking the strongest frequencies of a FFT is an unreliable, naive approach. The fundamental frequency of a note being played is not guaranteed

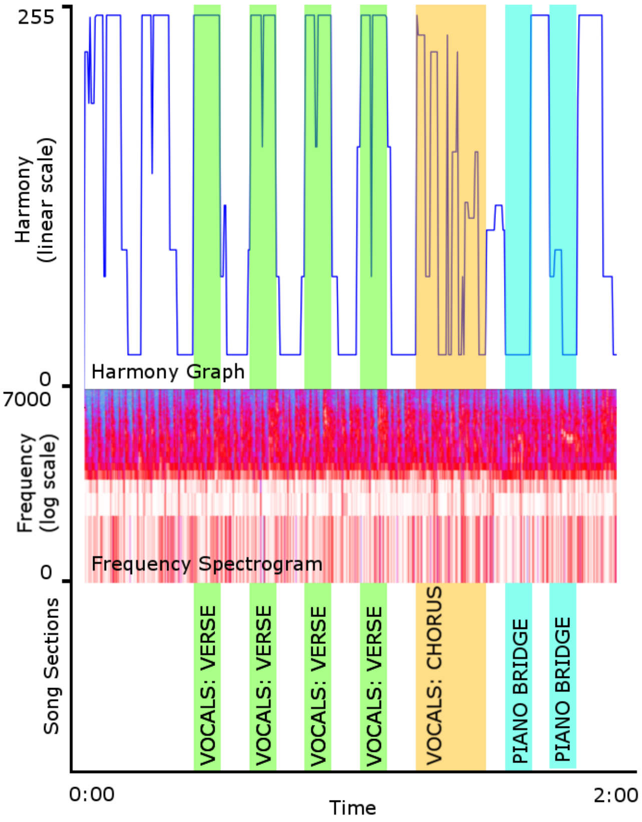


FIGURE 5.10: Harmony calculated using a naive Fast Fourier Transform approach compared to the frequency spectrum and manually identified parts of a sample song.

to be the strongest frequency and the usefulness of this approach is diminished further as more instrumentation is included and the frequency spectrum becomes more crowded with overtones and noise. Another issue with this approach is the speed at which the calculated harmony changes. In testing, we found the harmony to fluctuate wildly as the frequencies identified changed. To combat this, a smoothing window was used to find the mode harmony value. Using a short window of ten samples was quite effective, improving the signal to noise ratio greatly without adding any significant delay.

Another major drawback of using a FFT with live audio is the timing of visual updates. Without a smoothing window, colour changes are rapid and distinctly unenjoyable. With a smoothing window, colours change at a more enjoyable pace, but do not seem to change in synchrony with any musical cadence.

Regardless of these drawbacks, using the FFT approach produced some interesting data. Even using the naive FFT approach, by comparing the calculated harmony, and the manually identified sections of a sample song, it is clear that the harmony value is indicative of the part of the song, showing obvious differences between the verses, chorus and bridge.

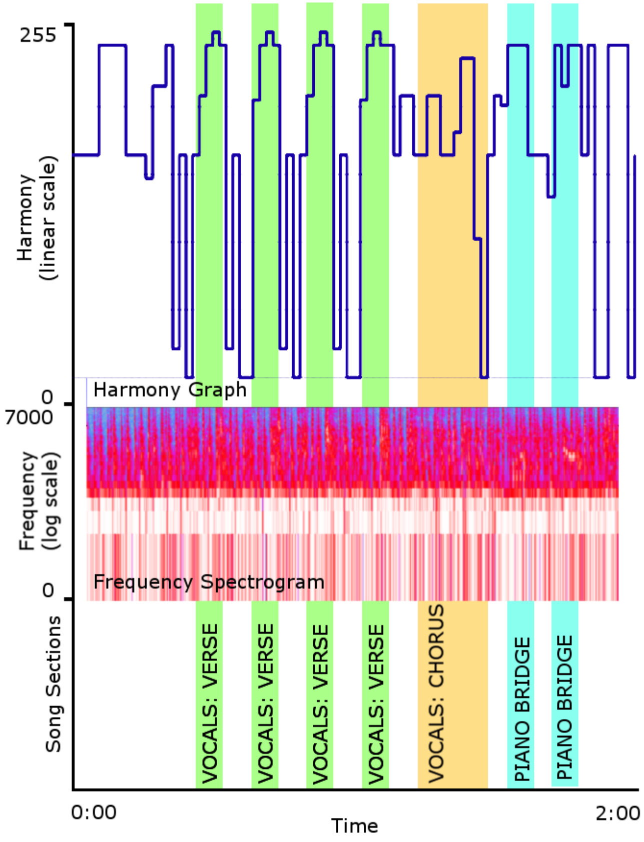


FIGURE 5.11: Harmony calculated using a MIDI score compared to the frequency spectrum and manually identified parts of a sample song.

Figure 5.11 shows the harmony values, frequency spectrum and the manually identified parts of the the song for the second approach using digital Midi signals.

By using the MIDI protocol, we can ensure that we know exactly what notes are being played at any time without any noise or interference. This clearly restricts us to monitoring digital instruments in a live setting. However, for the purposes of the Primary Implementation of the analogy system, we can take a sample recording and score each instrument in MIDI manually, as we have done for this experiment. This approach proved to be the most successful, producing a true representation of the music being played without any noise interference, and harmony values updating in synchrony with the music.

Interestingly, as illustrated in Figure 5.11, the harmony value calculated using the MIDI approach is remarkably similar to the value calculated using the FFT approach, with separate parts of the music piece similarly distinguishable. This suggests that while using an FFT to detect the timing, notes and chords of live audio is unreliable, it may still be a useful approach to calculate harmony.

5.3.3 Experiment: Evaluation of Mapping Expressions

Method

With a reliable musical input available, the next step in generating a live visual output is evaluating a *mapping expression* with musical input. This experiment investigates the time taken to evaluate a set of evolved Mapping Expressions.

Sample Mapping Expressions were evolved using the approach described in Section 5.2. A sample size of 1000 expressions was used, from all stages in the evolutionary process, across 100 separate populations. Evolved expressions varied in size between less than 10 sub-expressions, to over 7000, however, expression sizes were not evenly distributed. Expression size tended towards smaller expressions with a large portion of expressions containing less than 1000 sub-expressions.

Typically, in a full analogical system, a *mapping expression* is evaluated as follows. First, the *mapping expression* is read from a file and parsed into memory as a set of nested sub-expressions. Evolution of *mapping expressions* will be conducted offline, based on a sample performance or display, before any real time output is required. Second, when real-time evaluation commences, the parsed expression is passed to an evaluation script, alongside musical input values. Third, the evaluation script recursively evaluates each sub-expression, replacing terminals representing musical input with the current real-time values.

During evolution, the fitness of a *mapping expression* is calculated by using a sample input of 11 note intervals. This requires the *mapping expression* to be evaluated 11 times, once for each interval. For this experiment, we measured the time taken for the fitness of each expression to be calculated. This ensures that all intervals are evaluated, removing any short-cuts that an expression may take for specific inputs. As a consequence of this, the actual evaluation time of an individual expression can only be estimated by dividing the measured time by 11.

The measurement was taken by noting the system time immediately before and after the fitness of an expression was calculated. The difference between these times is the time taken to calculate the fitness.

Results

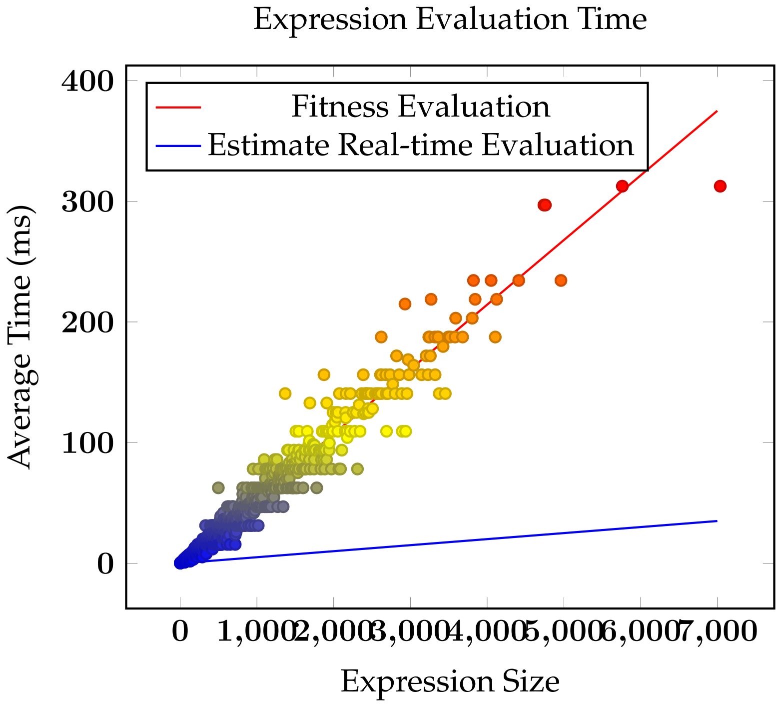


FIGURE 5.12: Average execution time to evaluate the fitness of mapping expressions with an estimation of real-time evaluation. Mapping expression size is measured by number of sub-expressions.

The results of this experiment are illustrated in Figure 5.12, where we plot the time taken to calculate the fitness of *mapping expressions*, by size, measured in number of sub-expressions. We also display the linear trend of evaluation time in red, and the estimated real-time evaluation time in blue. This real-time evaluation time was obtained by dividing the fitness evaluation time by 11.

It is clear from the figure that evaluation time is strongly correlated with the size of an expression in terms of sub-expressions.

Some apparent striation can be observed parallel to the x-axis in the figure indicating that similar time delays were measured for multiple expressions of varying sizes. This may be the result of low granularity of time measurements available to the experimental program, or the effects of other background process running alongside the experimental program.

5.3.4 Experiment: Visual Display Generation

Method

Presenting and updating a visual display with minimal time delay is also a challenge. This experiment investigates the time taken to read some input values, evaluate a mapping expression with those input values, send the result to a visualisation system, and render the visualisation.

To achieve this, both experimentally, and in a full analogical system, we implemented a *visualisation server* subsystem. This subsystem was made up of a web server with API endpoints accepting colour update requests, and web pages served to render the display. Websockets were used to update the display as new colour update requests were received.

The display could then be shown on a data projector, external monitors or any other display device. This approach was not ideal, but presented a cost effective and robust solution without the need for expensive stage lighting equipment and proprietary hardware for data transfer.

Data was gathered by selecting an evolved *mapping expression* with a high fitness value of reasonably large size, and using this expression to produce a visual display using the *visualisation server* and a sample MIDI input file. As the file was played, each time a new note was detected, the system time was stored, alongside the time stamp of the music note from within the MIDI file. Similarly, when the *visualisation server* both received a new input, and completed rendering a new visual, the system time was noted, alongside the time stamp of the music note.

This time series data was collected for the duration of the full song at normal speed, while synthesizing the associated audio.

Results

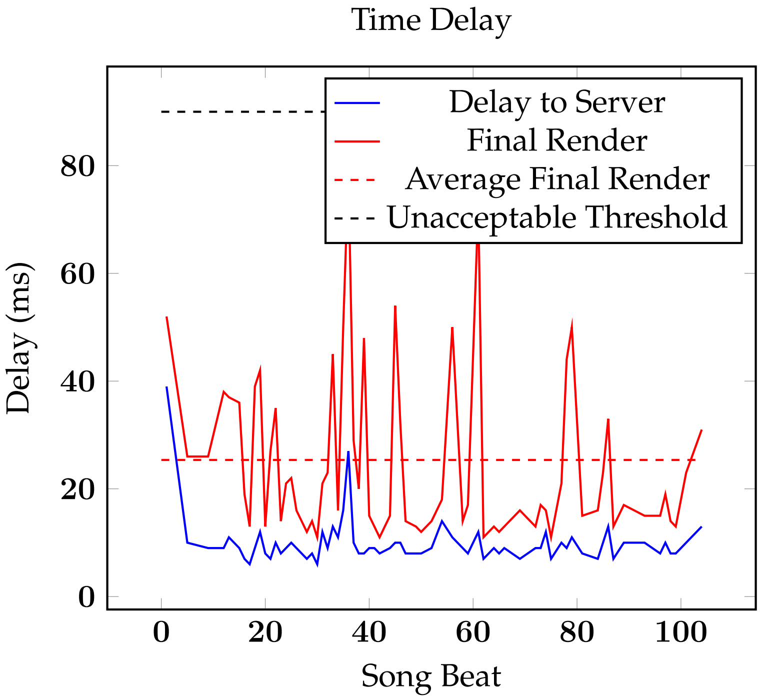


FIGURE 5.13: Time delay in milliseconds from a MIDI message being sent to the Visualisation Server, and the visualisation being rendered for a single typical run.

The overhead of evaluating a larger *mapping expression* as well as communicating with a visualisation tool could result in unacceptable delays, preventing suitable time synchronisation between audio and visuals. According to the ITU-R BT.1359-1 (1998) standard, a sound/vision timing difference of more than 90 milliseconds is unacceptable, while anything less than 45 milliseconds is undetectable. By plotting the time delay between a MIDI signal being received and a visual being rendered, we can estimate the acceptability of the time delay of the system.

In Figure 5.13 we plot the time delay between a MIDI signal being received by the analogical system and a visual being rendered in red, and the time delay between the analogical system and the visualisation server in blue.

We also plot the unacceptability threshold according to the ITU-R BT.1359-1 (1998) standard in dotted black, with the average full time delay in dotted red.

The figure displays a typical run with maximum delays of 79ms, minimum delays of 11ms, and average delay of 24.357ms indicating an acceptable delay.

5.4 Observer Survey: Effect of Analogy on Aesthetic Response

In this section we examine the effects that an an evolutionary system for real-time aesthetic analogies may have on observers listening to input music and watching a generated visual display. We present the results of an exploratory study testing whether evolved mapping expression between the measurable aesthetic attributes of musical and visual harmony will produce an improved aesthetic experience compared to a random mapping expression. A sample set of expressions with various fitness values were used and the participants were surveyed on their enjoyment, interest, and fatigue.

5.4.1 Survey Design Considerations

In the previous Sections 5.2 and 5.3 we describe the design and implementation of an evolutionary system for real-time aesthetic analogies. We test the effectiveness of the individual parts of this system to ensure that they perform as expected and within tolerable ranges. A visual overview of the system is shown in Figure 5.8.

While each of the various sub-systems appear to perform well in isolation, and indeed the entire system is shown to operate — generating visuals based on a musical input, both from a recorded Midi file and live instruments — the effect of this system on the aesthetic response of an observer remains untested.

Surveying aesthetic response is a task we have discussed in great detail previously in this thesis, and the challenges of this task are only compounded when a more complex system is considered with detailed music and temporally changing visuals. Previously, we were able to directly compare short pieces of information, a pair of musical notes or colours, in terms of a single response: preference. However, in evaluating the effect of the analogy system, we are interested in a more complex response, an experience sustained over a period of time and a combination of response axes[1].

This sustained experience means that comparing like with like is impossible. In the time it takes for each sample display to be performed, the survey participant is likely to have forgotten most of the preceding display. We opt instead for a more typical survey methodology where each subject is simply asked a series of questions immediately after watching a display.

Our hypotheses focuses on the response dimensions of enjoyment, interest and fatigue. These three dimensions, we believe, should adequately capture the state of an individuals aesthetic response to an audio-visual display, such as the display the analogy system produces. In previous chapters we have discussed the effects of fatigue on survey participants and with this discussion in mind we aim to reduce subject fatigue. Using just these three response dimensions reduces the number of questions required of a participant, and therefore reduces the overall fatigue experienced by each participant.

5.4.2 Experiment: Observer Survey

Method

Using the approach described in Section 5.3 above, a live music performance can be used to create real-time visual displays with an aesthetic analogy system and mapping expressions.

Using the evolutionary approach described in Section 5.2, over 500 generations of *mapping expressions* were generated and stored. Three sample generations were selected and the lowest, median and highest fitness individual expressions were chosen for display. For each *mapping expression* (9 in total), a visual display was generated using a sample music piece. The visual display was recorded to ensure the output was identical for all survey participants and no issues

---

[1] We make the assumption that a response is multidimensional and can be broken into several aspects, axes, or dimensions. This allows us to better understand the overall response in terms of more concrete dimensions such as enjoyment, interest and fatigue.

with audio-visual synchrony were encountered. These recordings are henceforth referred to as 'videos'.

In the interest of displaying the same colours to all participants, and to avoid colour grading or exposure related loss of colour strength, videos were displayed on a large screen in a darkened conference room, using the same screen for all participants. The display consisted of a constant red background colour, and a varying foreground colour. The colours were arranged on screen as three vertical stripes, with the foreground colour in the center. This allowed the colours to illuminate the room distinctly.

The music piece, an arrangement of guitar, vocal harmony, piano, and light drums was used as it demonstrated a relatively strong periodic variation in musical harmony. This is the same musical piece shown in Figures 5.10 and 5.11. The piece was composed by the authors to be enjoyable without focusing on any particularly polarising genre and also to avoid any intellectual property issues. An external speaker was used to ensure high audio quality throughout the study.

Each of the 9 videos were of 2 minutes in length and were presented in a random order for each viewing. After each video, participants were asked to provide one answer to each of 3 multiple choice questions as shown in Table 5.5. In total, 32 participants were surveyed. A summary of the age, gender and musical background distributions are shown in Table 5.6.

TABLE 5.5: Survey Questions.

Category	Question	Value
Enjoyment	I enjoyed neither the music nor the colours.	0
	I enjoyed the music but not the colours.	1
	I enjoyed the colours but not the music.	2
	I enjoyed both the music and the colours.	3
Interest	I found neither the music nor the colours interesting.	0
	I found the music but not the colours interesting.	1
	I found the colours but not the music interesting.	2
	I found both the music and the colours interesting.	3
Fatigue	I feel tired or bored.	
	I do not feel tired or bored.	

TABLE 5.6: Study Participants Summary.

Attribute	Value	Percent (frequency)
Age	15-24	37.5% (12)
	25-44	43.75% (14)
	45-64	18.75% (6)
Gender	Male	56.25% (18)
	Female	43.75% (14)
Musical Background	No experience	43.75% (14)
	Beginner or Intermediate	43.75% (14)
	Expert	12.5% (4)

Results

Each participant was presented with nine visualisations, divided into three fitness levels: low, medium and high. The format of the survey allows us to plot responses for interest and enjoyment at varying levels of fitness as shown in Figure 5.14. Responses to music and colour can be separated, as shown in the figure with grouped bars. The labels of "Music-Low " through to "Colour-High" refer to the response to the music at low fitness level, and the response to colour at high fitness level respectively.

In the plot of "Enjoyment at 3 Fitness Levels", we see very little change in response for music at each fitness level, with a high rate of enjoyment of music reported at all levels. Similarly unchanging results are shown for enjoyment of colour, with the results leaning towards low enjoyment of colour.

Similarly, in the plot of "Interest at 3 Fitness Levels", we see practically no change in responses for music. Again we see a high rate of interest reported at all fitness levels. For colour, however, we do see a very slight increase in interest as fitness increases.

A statistical analysis of reponses was also carried out with each of the four mutually exclusive responses to Enjoyment and Interest given a value, 0 to 3 as shown in Table 5.5.

The input variables include 3 categorical variables (or factors), namely Fitness, Gender, and Musical Background and 1 quantitative input (or covariate), Age. The response values within each fitness level were summed, obtaining an ordered categorical response variable, taking values 0 to 9 representing the aggregate enjoyment or interest of a participant at each fitness level.

One complication in the analysis is that certain observations are correlated, namely those on the same participant at the different fitness levels.

To model either of the responses Enjoyment and Interest as a function of Fitness, Age, Gender and Background, care must be taken to account for any correlation between observations taken on the same subject. The factor Fitness (at 3 levels) is a within-subject variable and observations (of Enjoyment and of Interest) will be correlated due to the same individual being measured at each of the three levels of this factor.

An appropriate method of analysis is Generalised Estimating Equations since this allows us to accommodate the above-mentioned correlation and permits ordered categorical responses, such as those encountered here.

The software SPSS [80], provides suitable analysis through Generalised Estimating Equations. The results of the analysis are shown in Table 5.7. This analysis provides no evidence of an effect of any of the three factors (Fitness, Gender, Musical Background), nor of the covariate (Age) on either of the responses (Enjoyment and Interest).

5.5 Discussion

The results presented in the sections above show a mostly positive outcome. We were successful in defining an analogical structure which was codified such that Grammatical Evolution could be utilised to produce an output that represents an analogy. That analogy — or its representation — can then be used to create a visual output where before there was only sound — or the digital representation of sound.

Realistically, of course, the visual output of the system is lacking in a number of ways. It might be a stretch to call the display shown in Figure 5.9 a piece of art, but it *is* undeniably a visual display which has been created by a computer system. This system knows only the given parameters of aesthetics, a grammar and a way to compare results in the form of a fitness function. In this respect, the results might be compared to the exploratory play of a toddler as she learns. While the results might not be on par with more advanced or mature

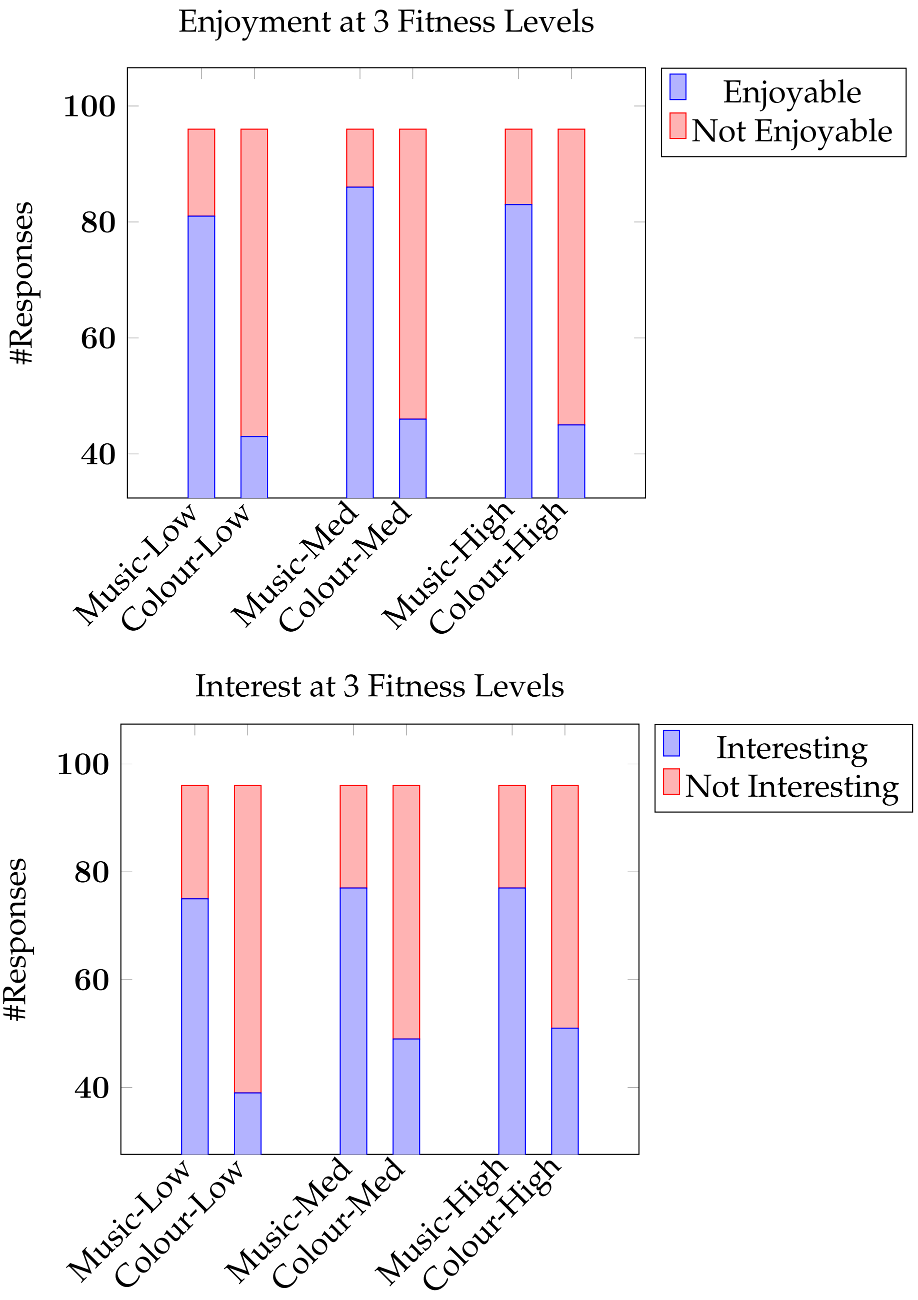


FIGURE 5.14: Responses at Varying Fitness Levels.

TABLE 5.7: Observer Survey Results: Enjoyment

Parameter		B	Std. Err.	95% Wald Confidence Interval		Hypothesis Test		
				Lower	Upper	Wald Chi-Square	df	P-value
Threshold	Enjoyment = 0	**7.170**	**82.3374**	**−154.209**	**168.548**	**0.008**	**1**	**0.931**
	Enjoyment = 2	**7.387**	**82.4698**	**−154.251**	**169.025**	**0.008**	**1**	**0.929**
	Enjoyment = 3	**8.256**	**82.8005**	**−154.030**	**170.542**	**0.010**	**1**	**0.921**
	Enjoyment = 4	**8.556**	**82.9006**	**−153.926**	**171.038**	**0.011**	**1**	**0.918**
	Enjoyment = 5	**11.058**	**83.5490**	**−152.695**	**174.811**	**0.018**	**1**	**0.895**
	Enjoyment = 6	**9.668**[a]	.	.	.	.	.	.
	Enjoyment = 7	**1.657**[a]	.	.	.	.	.	.
Fitness=High		**−0.872**	**0.6498**	**−2.146**	**0.402**	**1.801**	**1**	**0.180**
Fitness = Low		**0.238**	**0.4366**	**−0.617**	**1.094**	**0.298**	**1**	**0.585**
Fitness = Medium		**0**[b]	.	.	.	.	.	.
Gender = Female		**1.497**	**1.3819**	**−1.212**	**4.205**	**1.173**	**1**	**0.279**
Gender = Male		**0**[b]	.	.	.	.	.	.
Background = 1		**8.240**	**82.5327**	**153.521**	**170.001**	**0.010**	**1**	**0.920**
Background = 2		**7.3337**	**82.3433**	**154.053**	**168.727**	**0.008**	**1**	**0.929**
Background = 3		**0**[b]	.	.	.	.	.	.
Age		**0.044**	**0.0367**		**0.116**	**1.435**	**1**	**0.231**

Dependent Variable: Interest
Model: (Threshold), Fitness, Gender
a: Hessian Matrix singularity is caused by this parameter. Estimate for last iteration is shown.
b: Set to zero. Parameter is redundant.
a and b indicate that observed values in the dataset are not numerous enough for meaningful analysis.

TABLE 5.8: Observer Survey Results: Interest

Parameter		B	Std. Err.	95% Wald Confidence Interval		Hypothesis Test		
				Lower	Upper	Wald Chi-Square	df	P-value
Threshold	Interest = 0	−2.812	28.57772	−58.823	53.198	0.010	1	0.922
	Interest = 1	−2.571	31.7508	−64.802	59.659	0.007	1	0.935
	Interest = 2	−2.205	34.9766	−70.758	66.348	0.004	1	0.950
	Interest = 3	−0.168	31.0808	−61.086	60.749	0.000	1	0.996
	Interest = 4	0.010	33.0692	−64.804	64.825	0.000	1	1.000
	Interest = 5	1.043	34.9397	−67.438	69.524	0.001	1	0.976
	Interest = 6	1.774	32.2771	−61.488	65.036	0.003	.	0.956
	Interest = 7	11.426	1940.5091	3791.902	3814.754	0.000	.	0.995
	Interest = 8	578.996^a	.	.	.	.	.	.
Fitness=High		−0.482	3.6284	−7.593	6.630	0.018	1	0.894
Fitness = Low		−0.169	14.4551	−28.501	28.162	0.000	1	0.991
Fitness = Medium		0^b	.	.	.	.	.	.
Gender = Female		0.319	72.9867	−142.732	143.370	0.000	1	0.997
Gender = Male		0^b	.	.	.	.	.	.

Dependent Variable: Enjoyment

Model: (Threshold), Fitness, Gender, Background, Age

a: Hessian Matrix singularity is caused by this parameter. Estimate for last iteration is shown.

b: Set to zero. Parameter is redundant.

a and b indicate that observed values in the dataset are not numerous enough for meaningful analysis.

systems, what we are shown is really the beginning of what might be possible.

5.5.1 Analogical Structure

One of the clear drawbacks of the work presented here is the restriction in the structure of the analogy used. Currently, the system is designed to create an analogy using a single aesthetic attribute — harmony. This attribute was chosen specifically because it was clearly present and easily measured in both domains.

However, harmony is only one of many potential attributes within the context of the aesthetics of music and visuals. The resulting analogy could therefore be described as a fractional aesthetic analogy, or an analogy of harmony alone which may contribute to the overall aesthetic quality of the output. We hope to achieve a full analogy of aesthetics by making use of more than one attribute. To this end, we have highlighted a number of potential candidates for future attributes to be included.

Symmetry, intensity, contrast and complexity have specifically been highlighted as candidates for our future work. All four attributes can be observed in each domain and we believe they all may be measured, to some degree, without a great deal of subjectiveness or computational expense.

Contrast is perhaps the easiest attribute to measure in both domains with contrast in visuals being measurable on a per-pixel basis and in music on a per-note basis.

Symmetry is similarly directly measurable in visuals, and potentially measurable in music using heuristics such as matching beat patterns or relative pitch changes.

Intensity in visuals may be measured as a function of colour contrast and fractal index while in music it may be measured by volume and tempo.

Finally, **complexity** is commonly measured in visuals using a fractal index and may be measured in music simply by measuring the number of notes being played in a period of time.

These attributes may not represent the full aesthetic gamut but we believe that they will provide a strong toolset, beyond harmony alone.

5.5.2 Effect on Aesthetic Response

In our preliminary implementation, described in Section 5.2, we report a positive result with evolved *mapping expressions* producing visuals that are more pleasant than randomly generated *mapping expressions*. Our primary implementation, described in Section 5.3 also reports a positive result with visuals generated based on MIDI input changing in synchrony with the music.

While these results seem positive, they are based solely on the observations of a single observer. These observations did seem positive however, so a more in-depth survey was conducted.

Section 5.4 presents the results of this survey, conducted to investigate the effect of the analogy system on the aesthetic response of human observers.

A study of human subjects requires the comparison of randomly generated visual displays, to displays created with *mapping expressions* of various fitness. Due to the subjective nature of human aesthetic response, a reasonably large study is required. A similar problem was faced when gathering the data used to build the aesthetic models used in this work. Even when tasked with ranking simple musical note pairs — or colour pairs — the participants in the study suffered from fatigue and boredom extremely quickly which may have had strong effects on the data gathered.

In the case of musical note pairs, the study could be conducted at a reasonably quick speed as each note pair would play for only a few seconds. To evaluate the output of the system presented in this chapter, a longer segment of music must be played, significantly increasing the time it takes to gather information from any individual subject. This has a knock-on effect of increasing fatigue and boredom, reducing the quality of the data gathered.

Nevertheless, results were obtained from the survey and analysis shows mixed results. A statistical analysis of the variables of the study shows p-values that indicate very little strength of correlation between any two factors. Initially, this would indicate that system did not manage to effect either the interest or enjoyment of subjects as the fitness of mapping expressions increased.

However, the results of this statistical analysis would suggest that while there is no strong correlation between fitness and enjoyment or fitness and interest, there is also no strong correlation between subject gender, age or musical experience and enjoyment or interest.

One of the major challenges faced in conducting a study of this kind is avoiding biases like these. Using music or colour that may appeal to a specific age group or gender would skew results. In particular, it could be suggested that participants with more musical experience may actively look for patterns or features in the visual display that less experienced subjects would not be aware of. Finding no correlation here suggests that the responses were no influenced by predefined notions of pattern, complexity or structure.

With this is mind, the raw responses can be viewed with a less critical eye. Figure 5.14 displays these responses plainly. It is clear to see that there is little change in enjoyment of music or colour as the fitness of mapping expressions increases. However, there is a slight, but observable increase in responses of positive interest in relation to colour as fitness increases.

It is possible that this increase is due to a change in fitness, and that statistical analysis did not show a strong correlation due to the small sample size or some other numerical factor. Further, it might also be possible that this increase is not particularly strong due to underlying methodological challenges, such as the environment in which the generated visual display was shown to participants, the design of the visual display and other factors.

It is also interesting that this increase is visible in terms of interest and not enjoyment. This indicates that the two terms of distinct and that a participant may find a display interesting, even if it is not particularly enjoyable. The dependence of this result on the terminology presented to the subject echoes the challenges faced when comparing previous studies of culturally distant populations.

Finally, given the slight increase in interest without a corresponding increase in enjoyment, could this suggest that a threshold must be reached before an increase in interest effects enjoyment? Following this, a question might also be asked: if it possible for art to be enjoyable without being interesting, is possible for art to be interesting without being enjoyable?

In practice of course, no visual display that is interesting, enjoyable or not, can be created in a fully autonomous manner as suggested in this work. A display for a live music performance would typically be created at the discretion of the lighting designer or artist in order to enhance the original performance in a very specific and meaningful way.

With this observation in mind, the ultimate goal of an analogy system should be to act as an artistic tool, enabling and enhancing the creativity of the artist, rather than replacing it.

5.5.3 Future Work

The results presented were obtained using only one *mapping expression* between musical consonance and colour harmony. We have not explored the possibilities of using multiple mapping expressions incorporating many attributes. We believe this will improve the quality of generated visuals dramatically.

As shown in Section 5.2.3, the fitness of a population has certain limitations. We hope to improve the speed at which fitness increases and also increase the maximum fitness achievable by any individual by tuning the parameters of the genetic operators.

The *extended MIDI* format has a number of useful applications beyond its use in this implementation. The format may also be useful for predefined visual displays and synchronised performances. With this in mind, we would like to fully define our version of the protocol and make it available to the public.

In a similar vein, the visualization server, which uses the *extended MIDI* format may also be improved. Most immediately, it should be able to handle all of the attributes used by *mapping expressions* to generate varied and immersing visual displays. Also, the server is currently restricted to displaying visual displays on a computer screen, which is not suitable for a live performance. We hope to develop functionality to allow the visualization server to accept an *extended MIDI* signal and control stage lighting hardware using industry standard protocols and open source lighting controllers.

The results of the preliminary and primary implementations demonstrate that we are moving in the right direction however the study described in Section 5.4 indicates a shortcoming in the existing implementation. The overall goal of this research is to create a system that can create art, and perform it. To this end, the success of the system should be evaluated with a live performance.

The existing system uses an offline evolution process that must be conducted before any performance is possible. Using a supervised fitness technique and real-time evolution, it may be possible for the system to dynamically update the structure of mapping expressions during a single performance, or throughout multiple rehearsal performances.

Using the existing system, we can interactively evaluate the performance of a particular *mapping expression*. We can either use a static MIDI file to compare individual *expressions* or we can use a live MIDI instrument to send live MIDI signals to evaluate how it performs with improvised and varying input.

Expressions that are deemed fit by human supervision may then be reintroduced to the evolution phase to continue the process. This step is independent of the fitness function in order to capture aesthetic results beyond its capabilities.

5.6 Chapter Conclusion

It has been demonstrated that computational evolution can be utilised in the creation of aesthetic analogies between two artistic domains by the use of mapping expressions. When given an artistic input these mapping expressions can be used to guide the generation of content in a separate domain.A number of experiments have been outlined that demonstrate the various aspects of this process.

The results of experiments conducted with a preliminary implementation of the system demonstrate that Grammatical Evolution can be successfully used to evolve mapping expressions representing an analogy. The fitness of these mapping expressions has been shown to increase over a number of generations before reaching a plateau of fitness. Evolved expressions are shown to have a fitness distinctly higher than randomly generated expressions, with highly fit expressions closely mapping musical input to the ideal visual output.

Further results are presented reporting the ability of the preliminary system to produce visual outputs based on musical inputs and play them in synchrony using MIDI input, an *extended MIDI player* sub-system and a *visualisation server* sub-system.

The results of experiments conducted with a primary implementation of the system demonstrate that the system can be used to produce visuals based on a dynamic musical input. The first experiment in the section demonstrates two potential methods of gathering musical input, using a live, analog music signal with a Fast Fourier Transform based approach and a digital signal with a MIDI based approach. It is also demonstrated that mapping expressions can be evaluated in real-time without significant time delays and that these expressions can be used to render a visual display in synchrony with a music with time delays within a suitable threshold.

Finally, the results of a survey conducted to investigate the effect of analogy on aesthetic response are presented. These results do not indicate that the primary implementation is sufficient to significantly affect aesthetic response.

With respect to **H4**, we aim to demonstrate that *An analogical structure can be created using the gathered aesthetic data.* The results presented in this chapter show that using mapping expressions can represent the functions of an analogical structure and can be evolved to map between gathered aesthetic data. We therefore conclude that **H4** is true.

With respect to **H5**, we aim to demonstrate that *the analogical structure can be evaluated in real-time.* The results of our primary implementation demonstrate that mapping expressions, representing the analogical structure, can be evaluated and used to render a visual display with sufficiently small delays. We therefore conclude that **H5** is true.

With respect to *H6, H7* and *H8*, we aim to demonstrate that *The use of aesthetic analogy has an impact on the enjoyment, interest and fatigue respectively of an observer in artistic material.* The results of our survey do not demonstrate any effect of aesthetic analogy on these three responses. We therefore conclude that there is insufficient evidence to support *H6, H7* or *H8*.

Chapter 6

Summary and Conclusions

6.1 Introduction

Machine learning and artificial intelligence techniques have improved by leaps and bounds in recent years with systems now capable of doing work that was previously regarded as impossible for a computer to do. In all this advancement however, one major challenge has remained unmet: creativity.

Of course, computers are very capable of generative creation - producing billions of variations of specific structures - but this is not what most would call creativity. For a human being, simply producing is not enough to be considered creative, but that output must hold some aesthetic value. In this respect, computational creativity remains a topic of science fiction. Primarily, the ethereal nature of human creativity and aesthetic experience is difficult to define and measure, leaving computer systems little solid data to work with.

Nevertheless, as human beings we have a fascination with the topic of creativity and the aesthetics of creative endeavours. Much has been written and many open questions have been identified such as: If the generation of an art piece can be controlled, can an aesthetic response be affected in a predictable way? Are there some factors that have stronger effect on aesthetic response than others? When we consider the application of computational systems in this area, we may also ask if it is possible to use a computational system to test the affects of aesthetic factors?

Continuing this line of inquisition, we take the concept of analogy as a paradigm of creativity - a method by which a computer system can produce an output which may or may not be considered to have aesthetic merit. Can a computationally intelligent system be used to create useful analogies? From a human point of view, the concept of analogy often means the translation of a concept from one area of thought to another. If we consider the translation between the senses, sight and sound for example, to be an aesthetic analogy, we see an effect similar to synesthesia. Further, if the combination of senses is useful and positive in the case of synesthesia, can a computational system be designed to create a positive experience by building analogies between artistic domains?

Computer systems have recently excelled in tasks that have previously been considered creative challenges using specialised implementations of artificial intelligence and machine learning systems. While current computational systems may be capable of performing well in what we consider a creative challenge, we gain no insight into their processes. Is it possible to use a creative computational system that provides insight into the creative process?

Finally, these systems operate in very niche areas where data can be gathered and hundreds of millions of simulated scenarios can be tested offline - without time or memory constraints. However, many real life applications of creativity and aesthetics do not provide this luxury. In this regard, supervised and interactive machine learning techniques are successful but the cost of gathering training data and of interacting with the system is a major pitfall. Can a system be designed that minimises human interactions and thus human fatigue while remaining robust to unreliable input and training data?

This thesis is inspired by these open questions and the research conducted in this respect is described in more detail in the following summary.

6.2 Summary

The work in this thesis aims to make use of machine learning and artificial intelligence techniques with an analogy based approach to explore creativity. It is hoped that in tackling this challenge, we may document and learn about the intricacies of creativity and aesthetics from a computational point of view, contributing to the knowledge base and providing a platform for further work in the area of computational creativity.

Based on the open questions identified in Chapter 1 and Section 6.1, an extensive literature review has been carried out and eight hypotheses have been identified. This literature review is detailed in Chapter 2 and the hypotheses are detailed in Section 1.2 and below.

The first two hypotheses are investigated in detail in Chapter 3. These are:

H1 Aesthetic data can be gathered in a way that minimises participant effort.

H2 Aesthetic attributes in the chosen domains of music and visuals can be objectively measured.

The Graphsort algorithm and its implementation has been designed to test these hypotheses. In Chapter 3 we demonstrate that the Graphsort algorithm can be used to gather data in a way that minimises participant effort by reducing the number of questions required of a subject in a 2-alternative forced choice style study. Aesthetic data can be gathered in a study of this type thus demonstrating Hypothesis 1 to be true.

Analysis of the Graphsort algorithm in comparison to other ranking and sorting algorithms has shown that Graphsort can be used rank items objectively, or without major biases when input is subject to noise. While this result in isolation does not fully demonstrate Hypothesis 2 to be true for the specific case of aesthetic attributes, it demonstrates the effectiveness of the algorithm in a general sense.

Chapter 4 concerns Hypothesis 2 in relation to aesthetic attributes specifically, as well as Hypothesis 3:

H3 Aesthetic data can be gathered in the chosen domains.

Here the Graphsort algorithm has been applied in two studies. In these studies we have measured responses to the aesthetic attributes of visual and musical harmony, demonstrating the objective measurement of aesthetic attributes and verifying hypothesis 2. Aesthetic data in the chosen domains of music and visuals has been gathered showing Hypothesis 3 to be true.

Hypotheses 4 and 5 have been investigated in Chapter 5:

H4 An analogical structure can be created using the gathered aesthetic data.

H5 The analogical structure can be evaluated in real-time.

In this chapter, an analogical system has been outlined that uses Grammatical Evolution to evolve Mapping Expressions to map aesthetic values from the musical domain to the visual domain. The aesthetic data used in this mapping has been gathered in Chapter 4, using the methodology described in Chapter 3.

The first section of Chapter 5 describes the initial, preliminary implementation of the system - designed to test Hypothesis 4 specifically. The designed system uses an analogical structure together with the gathered aesthetic data, successfully verifying Hypothesis 4.

The second section of Chapter 5 described the subsequent, primary implementation of the system. This implementation builds upon the preliminary system with the aim of verifying Hypothesis 5 specifically. The results of this section demonstrate that the system can operate in real-time, mapping values from one aesthetic domain to another, without any noticeable or significant delay.

In addition, the second research question and the associated hypotheses have been explored in this chapter:

H6 The use of aesthetic analogy has an impact on the enjoyment of an observer in artistic material

H7 The use of aesthetic analogy has an impact on the interest of an observer in artistic material.

H8 The use of aesthetic analogy has an impact on the fatigue of an observer while viewing artistic material.

In this regard, a single study has been conducted using the primary implementation of the system as described above to test all three hypotheses. The results of this study are not conclusive and do not demonstrate with any certainty that the use of aesthetic analogy has an impact on the enjoyment, interest of fatigue of an observer while viewing artistic material.

6.3 Contributions

We now present a summary of contributions. These contributions and associated publications are listed in Section 1.4.

- The Graphsort algorithm has been proposed to gather subjective feedback and to form an objective model based on unreliable input. The design and implementation of the Graphsort algorithm is described in detail in Chapter 3. This graph based ranking algorithm reduces the number of questions asked of an individual in the course of a 2-alternative forced-choice study. The reduction of questions is imperative to ensure a study requires no more time or effort of the subject than is absolutely necessary. The use of Graphsort reduces fatigue, boredom and the effects of context and juxtaposition thus improving the reliability of the subject's responses.

- A number of commonly used ranking and sorting algorithms have been tested alongside the Graphsort algorithm demonstrating the robustness of each algorithm to unreliable inputs. This work demonstrates that the defacto standard ranking and sorting algorithms may not be suitable in specific circumstances and defines archetypical *noise patterns* which can be used to classify and identify these circumstances.

- Data has been gathered demonstrating a common ranking of musical dyads (two notes played together) within one octave. Note pair offsets — the space between two notes in a dyad — are ranked in terms of consonance and dissonance, which is synonymous to musical harmony in this context.

- Data has been gathered demonstrating a common ranking of two-colour pairs varied by hue. Colour pairs are ranked in terms of colour harmony.

- The concept of *Mapping Expressions* has been introduced. In an analogical system, Mapping Expressions are a high level construct that allow attributes from one domain to be mapped to analogous attributes in another domain.

- An implementation of an analogical system has been presented which makes use of Grammatical Evolution to produce working Mapping Expressions capable of mapping a musical input into an analogous visual output in real time.

6.4 Conclusion

A primary research question has been defined as:
Can a computationally intelligent system be created which is capable of making useful aesthetic analogies between separate domains such as music and visuals?

This primary question has been broken down into the following two sub-questions:

1. Can a computationally intelligent system be built to create aesthetic analogies?
2. Does the analogy generated affect aesthetic response?

The work in thesis has found evidence supporting each of the hypotheses associated with the first sub-question. The use of the Graphsort algorithm demonstrates that aesthetic data can be gathered in a way that minimises participant effort (H1). Using Graphsort to measure harmony, aesthetic attributes in the chosen domains of music (musical consonance) and visuals (colour harmony) can be objectively measured (H2). The results of the studies conducted have demonstrated that aesthetic data can be gathered in the chosen domains. The use of mapping expressions evolved using Grammatical Evolution has shown that an analogical structure can be created using the gathered aesthetic data. Finally, the evolved mapping expressions, in the form of symbolic expressions demonstrate that an analogical structure can be evaluated in real-time. We therefore conclude that a computationally intelligent system can be built to create aesthetic analogies.

We have not found supporting evidence of any of the hypotheses associated with the second sub-question. No statistically significant results have been gathered that show the use of aesthetic analogy has an impact on the enjoyment of an observer in artistic material (H6). Similarly, no statistically significant results have been gathered that show the use of aesthetic analogy has an impact on the interest of an observer in artistic material (H7). Finally, no statistically significant results have been gathered that show the use of aesthetic analogy has an impact on the fatigue of an observer while viewing artistic material (H8). We therefore conclude that the analogy generated by the system proposed does not affect aesthetic response.

Based on the conclusions of the two sub-questions, we can partially support the primary research question. That is, a computationally intelligent system can be created which is capable of making aesthetic analogies between separate domains such as music and visuals. However, the specific analogies created by the system presented in this thesis has not been useful.

6.5 Future Work

Attempting to create a computationally intelligent system capable of manipulating aesthetic response was a far reaching goal. With this in mind, it should not be surprising that the goals of the hypothetical system have not been satisfied in every way. While our conclusions are not entirely positive, they do provide a very strong foundation upon which further research may build.

Most importantly, it has been found that it is possible to take two separate artistic domains and build a mapping between them that closely matches measured aesthetic responses. In this regard, there are many ways in which this system can be optimised and improved.

Weighted Graphsort

The Graphsort algorithm has been designed and tested in its current form using an unweighted directed graph and has been shown to work effectively. However, there are a number of potential limitations to the current implementation of the algorithm that future work may address.

Primarily, the algorithm assumes the subject has an outright preference for one object over another. This approach disregards the potential for objects that are equally preferred, potentially producing an incorrect, or misleading ranking.

In a similar vein, the algorithm also does not take into account any difference in preference magnitude. That is, it may be true that the subject prefers item $\boldsymbol{A}$ over item $\boldsymbol{B}$, but the subject may prefer item $\boldsymbol{A}$ over item $\boldsymbol{C}$ much more strongly. At present, this is not represented in the ranking graph and items $\boldsymbol{B}$ and $\boldsymbol{C}$ will be given an equal rank.

It is expected that incorporating weights on the directed edges of the graph will improve the performance of the algorithm, particularly when the magnitude of preferences, or equal preferences are encountered. We believe the core tie-breaking logic of the algorithm will not require any change. However, the presented approach to finding an intermediate ranking does not incorporate weighted edges and will require significant changes.

Once these changes have been made, further research will be required to evaluate the performance of the algorithm against the unweighted implementation, as well as the other sorting and ranking methods that were used in the original benchmarks.

Mapping Expressions and Analogy Structure

The concept of mapping expressions has been introduced as a practical method of computational analogy making. In this work, mapping expressions have been implemented using symbolic expressions evolved with a Grammatical Evolution approach.

In this work however, we have only used one-to-one, bijective mappings. This is simply a result of time restrictions and leaves the opportunity for many-to-one mappings to be investigated. In it's current form, the symbolic expression approach to mapping expressions will enable inputs from many domains in the form of variable terminal expressions. However, the output of the expression is restricted to a single value, limiting the application to a many-to-one structure.

A many-to-many structure may be possible if the mapping expression is implemented using some other method. Indeed, the implementation of mapping expressions in another medium may lead to interesting research opportunities. For example, mapping expressions may be implemented using a neural network which would allow continuous evaluation and optimisation while also allowing a many-to-many structure to be implemented.

One of the major drawbacks of the current implementation that we have mentioned previously is the limitation just two aesthetic attributes. The design of the system is such that any number of aesthetic attributes may be used, particularly in a many-to-one mapping structure. The attributes of musical and visual harmony were selected from a shortlist of other potential attributes due to the availability of data and the ease of implementation. Other potential aesthetic attributes that may be useful in future work are:

1. Intensity
2. Complexity / Granularity
3. Predictability
4. Symmetry

Further Research on the Effectiveness of Analogy

Three studies have been carried out where groups of human subjects were shown some stimulus and surveyed for their responses. Every effort was made to ensure the subjects were surveyed in identical environments and in such a way as to produce fair and honest results. However, due to the limitations of time and funding, it was not possible to survey large groups. Subjects could not be offered any compensation for their time and extra personnel could not be employed to help conduct the study. These restrictions necessarily limited the sample sizes to their current values.

The effects of these small sample sizes are not particularly problematic for the first two studies—investigating the ranking of musical and visual harmonies—due to the largely similar feedback from subjects. However, in the final study investigating Hypotheses 6, 7 and 8, the sample size was far too small to identify any behaviours with statistical significance. As a result, these hypotheses remain conspicuously inconclusive. Future work should include a reproduction of the studies in this thesis on a larger scale.

Beyond this, further shortfalls may also be addressed. In repeated studies, the previously mentioned aesthetic attributes should be tested in isolation and in combination, in as many permutations as possible. For hypotheses 6, 7, and 8, the visual displayed, including the shape, brightness, saturation, blending methods, and animations should all be considered. Further, the work presented in this thesis only considers a single musical piece which may have led to excessive boredom or fatigue experienced by subjects which may have effected their responses.

Application of Analogy

The implemented analogical system described in Chapter 5 was capable of mapping aesthetic attributes from one domain (music) to another (visuals) in real time. In this respect, the system is a success. However, there are many limitations in the implementation of the system that prevent this system from being used in a real-world scenario.

The concept of computational creativity and analogy is perhaps best suited to the application of a creative tool: a system that can aid in the creation of art, rather than create art without any human interaction whatsoever. One goal in this regard would be to package the existing evolution and analogy system for use by working musicians, artists or stage performers.

Currently, many configuration options must be hard-coded into the application code in order to operate with musical input and send visual output. This severely limits the usefulness of the system as a flexible creative tool. Beyond this, a great deal of effort would be required to ensure compatibility with standard music and lighting systems as well as cross-platform software standards. This would certainly make for satisfying future work, if not worthwhile academic endeavour.

Beyond polishing the existing system for commercial use, the evolutionary stage of the system can also be improved. Evolution has not been optimised and must be conducted offline. Genetic algorithms, such as the Grammatical Evolution system used, benefit from the ability to be run in a distributed environment which can drastically improve evaluation time. Currently however, the system is limited to a single instance and as a result, evolution takes a long period of time.

The restriction of the evolutionary stage of the system which must occur offline, before mapping expressions can be used for real-time analogy. This is a sever restriction on the utility of the system as a multimedia tool, but there are also academic opportunities for future work here. The use of semi-supervised machine learning techniques may be employed here to enable real-time analogy without offline evolution.

The current system is also limited to the creation of analogies between just two aesthetic attributes. These attributes represent the absolute minimum set of a vast range of potential aesthetic attributes that might be used. Again, time constraints have severely limited the investigation into other potential aesthetic attributes, but further research should be carried out to investigate the potential applications of this larger set of aesthetic attributes.

Finally, beyond the artistic and creative applications of analogy which have been the focus of this work so far, computational analogy may have industrial applications that provide ample opportunity for further academic research. The inputs to a mapping expression, or the source domain of a computational analogy is not restricted to artistic or aesthetic attributes. Any continuously varying temporal variable may be used. Real-time analytics of physical or digital systems, environmental variables or any other streaming data point may be fed into the system which can then be used to provide an analogical summary of the system state.

Bibliography

[1] Martín Abadi et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (2016).

[2] Deepak Ajwani and Tobias Friedrich. "Average-case analysis of incremental topological ordering". In: *Discrete Applied Mathematics* 158.4 (2010), pp. 240–250. ISSN: 0166218X. DOI: `10.1016/j.dam.2009.07.006`. URL: `http://www.sciencedirect.com/science/article/pii/S0166218X09003047`.

[3] Hamzah Al Najada and Imad Mahgoub. "Autonomous vehicles safe-optimal trajectory selection based on big data analysis and predefined user preferences". In: *Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE Annual*. IEEE, 2016, pp. 1–6. ISBN: 1509014969.

[4] Amazon Inc. *Amazon Machine Learning*. 2017. URL: `https://aws.amazon.com/machine-learning/` (visited on 07/20/2017).

[5] John R Anderson. *Rules of the Mind*. Psychology Press, 1993.

[6] Artistotle. *Politics: A Treatise on Government VIII*.

[7] Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008. ISBN: 1848009984.

[8] Simon Baron-Cohen et al. "Synaesthesia: prevalence and familiality". In: *Perception* 25.9 (1996), pp. 1073–1079. ISSN: 0301-0066.

[9] Joseph D Becker. "The Modeling of Simple Analogic and Inductive Processes in a Semantic Memory System". In: *Proceedings of IJCAI-69*. Washington, 1969, pp. 655–668.

[10] Ramona Behravan. "Automatic Mapping of Emotion in Music to Abstract Visual Arts". PhD thesis. University College London, 2007.

[11] Hanan Alsouly Bennaceur and Hachemi. "Enhanced Genetic Algorithm for Mobile Robot Path Planning in Static and Dynamic Environment". In: *Proceedings of the 8th International Joint Conference on Computational Intelligence*. Porto, Portugal, 2016, pp. 121–131.

[12] James Bergstra et al. "Theano: Deep learning on gpus with python". In: (2011).

[13] Michael Betancourt. *Thomas Wilfred's Clavilux*. Wildside Press LLC, 2006. ISBN: 0809556715.

[14] Hans-Georg Beyer and Hans-Paul Schwefel. "Evolution strategies–A comprehensive introduction". In: *Natural computing* 1.1 (2002), pp. 3–52. ISSN: 1567-7818.

[15] G.D. Birkhoff. *Aesthetic Measure*. Cambridge University Press, 1933.

[16] G.D. Birkhoff. "Atti Congressi Bologna". 1928.

[17] Anne J Blood and Robert J Zatorre. "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion". In: *Proceedings of the National Academy of Sciences* 98.20 (2001), pp. 11818–11823. ISSN: 0027-8424.

[18] James K Bowmaker and HJk Dartnall. "Visual pigments of rods and cones in a human retina." In: *The Journal of physiology* 298.1 (1980), pp. 501–511. ISSN: 1469-7793.

[19] Mark Braverman and Elchanan Mossel. "Sorting from noisy information". In: *arXiv preprint arXiv:0910.1191* (2009).

[20] Aidan Breen and Colm O'Riordan. "Capturing and Ranking Perspectives on the Consonance and Dissonance of Dyads". In: *Sound and Music Computing Conference*. Maynooth, 2015, pp. 125–132.

[21] Aidan Breen and Colm O'Riordan. "Capturing Data in the Presence of Noise for Artificial Intelligence Systems". In: *Irish Conference on Artificial Intelligence and Cognitive Science*. Dublin, 2016, pp. 204–216.

[22] Aidan Breen and Colm O Riordan. "Evolving Art Using Aesthetic Analogies: Evolutionary Supervised Learning to Generate Art with Grammatical Evolution". In: *8th International Conference on Evolutionary Computation, Theory and Applications*. Porto, Portugal, 2016, pp. 59–68.

[23] Aidan Breen, Colm O Riordan, and Jerome Sheahan. "Evolved Aesthetic Analogies to Improve Artistic Experience". In: *6th International Conference on Computational Intelligence in Music, Sound, Art and Design.* Amsterdam, 2017.

[24] Ruth V Brittin and Robert A Duke. "Continuous versus summative evaluations of musical intensity: A comparison of two methods for measuring overall effect". In: *Journal of Research in Music Education* 45.2 (1997), pp. 245–258. ISSN: 0022-4294.

[25] Rodney Allen Brooks. "A robust layered control system for a mobile robot". In: *IEEE Journal of Robotics and Automation* 2.1 (1986), pp. 14–23. ISSN: 0882-4967.

[26] Richard Brown. "Use of analogy to achieve new expertise". PhD thesis. MIT, 1977.

[27] Raluca Budiu and John R Anderson. "Integration of Background Knowledge in Language Processing: A Unified Theory of Metaphor Understanding, Moses Illusions and Text Memory". 2008.

[28] Kenneth E. (University of Central Arkansas) Burchette. "Color harmony". In: *Color Research and Application* 27.1 (2002), pp. 28–31. ISSN: 03612317. DOI: `10.1002/col.10004`.

[29] Ludwig M Busse, Morteza Haghir Chehreghani, and Joachim M Buhmann. "The information content in sorting algorithms". In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on* (2012), pp. 2746–2750. ISSN: 2157-8095. DOI: `10.1109/ISIT.2012.6284021`.

[30] William A Camfield. "Marcel Duchamp's fountain: Its history and aesthetics in the context of 1917". In: *Marcel Duchamp: Artist of the century* (1990).

[31] G Carpenter and S Grossberg. “Adaptive resonance theory: Stable self-organization of neural recognition codes in response to arbitrary lists of input patterns”. In: *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Lawrence Eribaum Associates, 1986, pp. 45–62.

[32] Challenging Technologies LLC. *http://www.dpchallenge.com*. 2016. URL: `http://www.dpchallenge.com` (visited on 01/20/2016).

[33] Irène Charon and Olivier Hudry. “An updated survey on the linear ordering problem for weighted or unweighted tournaments”. In: *Annals of Operations Research* 175.1 (2010), pp. 107–158. ISSN: 0254-5330.

[34] Liana De Girolami Cheney. “Edward Burne-Jones’s The Planets: Musical Spheres and Visions of a Benevolent Cosmos”. In: *Journal of Literature and Art Studies* 7.7 (2017), pp. 812–868.

[35] Noam Chomsky. “Three models for the description of language”. In: *Information Theory, IRE Transactions on* 2.3 (1956), pp. 113–124. ISSN: 0096-1000.

[36] Ming-Chuen Chuang and Li-Chen Ou. “Influence of a holistic color interval on color harmony”. In: *COLOR research and application* 26.1 (2001), pp. 29–39. ISSN: 0361-2317.

[37] Cockos. *Reaper*. 2016. URL: `http://www.reaper.fm/`.

[38] Miri Weiss Cohen, Leticia Cherchiglia, and Rachel Costa. “Evolving Mondrian-Style Artworks”. In: *International Conference on Evolutionary and Biologically Inspired Music and Art*. Springer, 2017, pp. 338–353.

[39] Richard E Cytowic and David Eagleman. *Wednesday is indigo blue: Discovering the brain of synesthesia*. MIT Press, 2009. ISBN: 0262012790.

[40] Haisheng Deng et al. “Image Compression Based on Genetic Algorithm and Deep Neural Network”. In: *Bio-Inspired Computing-Theories and Applications*. Springer, 2016, pp. 417–424.

[41] Russell C Eberhart and Yuhui Shi. "Comparison between genetic algorithms and particle swarm optimization". In: *International conference on evolutionary programming*. Springer, 1998, pp. 611–616.

[42] Tuomas Eerola and Adrian C North. "Expectancy-based model of melodic complexity". In: *Proceedings of the Sixth International Conference on Music Perception and Cognition. Keele, Staffordshire, UK: Department of Psychology. CD-ROM*. 2000.

[43] Anikó Ekárt et al. "Modelling the underlying principles of human aesthetic preference in evolutionary art". In: *Journal of Mathematics and the Arts* 6.2-3 (2012), pp. 107–124. ISSN: 1751-3472.

[44] Chris Eliasmith and Paul Thagard. "Integrating structure and meaning: A distributed model of analogical mapping". In: *Cognitive Science* 25.2 (2001), pp. 245–286. ISSN: 0364-0213.

[45] Wilfried Elmenreich, Tobias Ibounig, and István Fehérvári. "Robustness versus performance in sorting and tournament algorithms". In: *Acta Polytechnica Hungarica* 6.5 (2009), pp. 7–18.

[46] Thomas G Evans. *A program for the solution of a class of geometric-analogy intelligence-test questions.* Tech. rep. Air Force Cambridge Research Labs LG Hanscom Field, 1964.

[47] Hugh S Fairman, Michael H Brill, and Henry Hemmendinger. "How the CIE 1931 color-matching functions were derived from Wright-Guild data". In: *Color Research & Application* 22.1 (1997), pp. 11–23. ISSN: 0361-2317.

[48] Brian Falkenhainer. "Analogical interpretation in context". In: *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. 1990, pp. 69–76.

[49] Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. *The structure-mapping engine.* Department of Computer Science, University of Illinois at Urbana-Champaign, 1986.

[50] C Fassnidge, C Cecconi-Marcotti, and E D Freeman. "A deafening flash! Visual interference of auditory signal detection". In: *Consciousness and Cognition* (2016). ISSN: 1053-8100.

[51] Ronald W Ferguson. "MAGI: Analogy-based encoding using regularity and symmetry". In: *Proceedings of the sixteenth annual conference of the Cognitive Science Society*. Citeseer, 1994, pp. 283–288.

[52] David Ferrucci. "Build Watson: an overview of DeepQA for the Jeopardy! challenge". In: *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. ACM, 2010, pp. 1–2. ISBN: 1450301789.

[53] Daniel Filonik and Dominikus Baur. "Measuring aesthetics for information visualization". In: *Information Visualisation, 2009 13th International Conference*. IEEE, 2009, pp. 579–584. ISBN: 0769537332.

[54] Lawrence J Fogel, Alvin J Owens, and Michael J Walsh. "Artificial intelligence through simulated evolution". In: (1966).

[55] Kenneth D Forbus, Ronald W Ferguson, and Dedre Gentner. "Incremental structure-mapping". In: *Proceedings of the sixteenth annual conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Inc Hillsdale, NJ, 1994, pp. 313–318.

[56] Kenneth D Forbus, Dedre Gentner, and Keith Law. "MAC/FAC: A Model of Similarity-Based Retrieval". In: *Cognitive Science* 19.2 (1995), pp. 141–205. ISSN: 1551-6709.

[57] Robert M. French. "The computational modeling of analogy-making." In: *Trends in cognitive sciences* 6.5 (2002), pp. 200–205. ISSN: 1879-307X. URL: `http://www.ncbi.nlm.nih.gov/pubmed/11983582`.

[58] Robert Matthew French. *The subtlety of sameness: A theory and computer model of analogy-making*. MIT press, 1995. ISBN: 0262061805.

[59] Galilée et al. *Dialogues Concerning Two New Sciences*. Dover Publ. Incorporated, 1967.

[60] Dedre Gentner. "Structure-Mapping: A Theoretical Framework for Analogy*". In: *Cognitive science* 7.2 (1983), pp. 155–170. ISSN: 1551-6709.

[61] Dedre Gentner and Kenneth D Forbus. “Computational Models of Analogy”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 2.3 (2011), pp. 266–276.

[62] Dedre Gentner and Kenneth D Forbus. “MAC/FAC: A model of similarity-based retrieval”. In: *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. Vol. 504. 1991, p. 509.

[63] Mark E Glickman. “Example of the Glicko-2 system”. In: *Boston University* (2012).

[64] Mark E Glickman. “The glicko system”. In: *Boston University* (1995).

[65] Joseph A Goguen. “Art and the Brain: Editorial introduction”. In: *Journal of Consciousness Studies* 6.6 (1999), pp. 5–14.

[66] G W Granger. “An experimental study of colour harmony”. In: *The Journal of General Psychology* 52.1 (1955), pp. 21–35. ISSN: 0022-1309.

[67] Bernhard Haeupler et al. “Incremental cycle detection, topological ordering, and strong component maintenance”. In: *ACM Transactions on Algorithms (TALG)* 8.1 (2012), p. 3. ISSN: 1549-6325.

[68] Graeme S Halford et al. “Connectionist implications for processing capacity limitations in analogies”. In: *Advances in connectionist and neural computation theory* 2 (1994), pp. 363–415.

[69] Rogers P Hall. “Computational approaches to analogical reasoning: A comparative analysis”. In: *Artificial Intelligence* 39.1 (1989), pp. 39–120.

[70] Takehiro Hasegawa et al. “Learned audio-visual cross-modal associations in observed piano playing activate the left planum temporale. An fMRI study”. In: *Cognitive Brain Research* 20.3 (2004), pp. 510–518. ISSN: 0926-6410.

[71] James Hendler. “Developing hybrid symbolic/connectionist models”. In: *Advances in Connectionist and Neural Computation Theory* 1 (1991), pp. 165–179.

[72] James A Hendler. “Marker passing over Microfeatures: Towards a Hybrid Symbolic/Connectionist Model”. In: *Cognitive Science* 13.1 (1989), pp. 79–106. ISSN: 1551-6709.

[73] John Henry Holland. *Adaptation in Natural and Artificial Systems*. MIT press, 1975.

[74] Keith J Holyoak and Paul Thagard. "Analogical mapping by constraint satisfaction". In: *Cognitive science* 13.3 (1989), pp. 295–355. ISSN: 1551-6709.

[75] Amy K Hoover, Paul A Szerlip, and Kenneth O Stanley. "Interactively evolving harmonies through functional scaffolding". In: *Proceedings of the 13th annual conference on Genetic and evolutionary computation*. ACM, 2011, pp. 387–394. ISBN: 1450305571.

[76] Mengfei Huang. "The Neuroscience of Art". In: *Stanford Journal of Neuroscience* 2.1 (2009), pp. 24–26. URL: `http://www.stanford.edu/group/co-sign/Huang.pdf`.

[77] John E Hummel and Keith J Holyoak. "Distributed representations of structure: A theory of analogical access and mapping." In: *Psychological Review* 104.3 (1997), p. 427. ISSN: 1939-1471.

[78] Patrick Hutchings and Jon McCormack. "Using Autonomous Agents to Improvise Music Compositions in Real-Time". In: *International Conference on Evolutionary and Biologically Inspired Music and Art*. Springer, 2017, pp. 114–127.

[79] William Hutchinson and Leon Knopoff. "The acoustic component of Western consonance". In: *Journal of New Music Research* 7.1 (1978), pp. 1–29. ISSN: 0303-3902.

[80] IBM Corp. *IBM SPSS Statistics for Windows*. Armonk, NY, 2013.

[81] Nilendu G Jani and Daniel S Levine. "A neural network theory of proportional analogy-making". In: *Neural Networks* 13.2 (2000), pp. 149–183. ISSN: 0893-6080.

[82] Jörg Jewanski. "Colour and music". In: *Grove Music Online* (2002).

[83] Akio Kameoka and Mamoru Kuriyagawa. "Consonance theory part I: Consonance of dyads". In: *The Journal of the Acoustical Society of America* 45.6 (1969), pp. 1451–1459. ISSN: 0001-4966.

[84] Wassily Kandinsky and Hilla Rebay. *Point and line to plane*. Courier Corporation, 1947. ISBN: 0486238083.

[85] Richard M Karp. *Reducibility among combinatorial problems*. Springer, 1972, pp. 85–103. ISBN: 1468420038.

[86] Brett Kashmere. "Len Lye". In: *Senses of Cinema*. 2007.

[87] Irit Katriel and Hans L Bodlaender. "Online Topological Ordering". In: *ACM Trans. Algorithms* 2.3 (2006), pp. 364–379. ISSN: 1549-6325. DOI: `10.1145/1159892.1159896`. URL: `http://doi.acm.org/10.1145/1159892.1159896`.

[88] Mark T Keane and Mike Brayshaw. *The incremental analogy machine: A computational model of analogy*. Human Cognition Research Laboratory, 1988.

[89] Mark T Keane, Tim Ledgeway, and Stuart Duff. "Constraints on analogical mapping: A comparison of three models". In: *Cognitive Science* 18.3 (1994), pp. 387–438. ISSN: 1551-6709.

[90] Claire Kenyon-Mathieu and Warren Schudy. "How to rank with few errors". In: *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 2007, pp. 95–103. ISBN: 1595936319.

[91] Ulrich Kirk et al. "Brain correlates of aesthetic expertise: a parametric fMRI study". In: *Brain and cognition* 69.2 (2009), pp. 306–315. ISSN: 0278-2626.

[92] Paul Klee. *Pedagogical Sketchbook*. Washington: Praeger Publishers, 1925. URL: `http://monoskop.org/images/3/30/Klee{\_}Paul{\_}Pedagogical{\_}Sketchbook{\_}1960.pdf`.

[93] Adrian Bernard Leopold Klein and Adrian Cornwell-Clyne. *Colour-music: The art of light*. London: Lockwood, 1926.

[94] B Kokinov. "Associative memory-based reasoning: Some experimental results". In: *Proceedings of the 12th annual conference of the cognitive science society*. Citeseer, 1990, pp. 741–749.

[95] Boicho Kokinov and Alexander Petrov. "Integrating memory and reasoning in analogy-making: The AMBR model". In: *The analogical mind: Perspectives from cognitive science* (2001), pp. 59–124.

[96] Boicho N Kokinov. “The DUAL Cognitive Architecture: A Hybrid Multi-Agent Approach.” In: *ECAI*. 1994, pp. 203–207.

[97] John R Koza. “Genetic programming as a means for programming computers by natural selection”. In: *Statistics and computing* 4.2 (1994), pp. 87–112. ISSN: 0960-3174.

[98] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[99] Sven Kuehne et al. “SEQL: Category learning as progressive abstraction using structure mapping”. In: *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. 2000, pp. 770–775.

[100] Trent E Lange. “Hybrid connectionist models: Temporary bridges over the gap between the symbolic and the subsymbolic”. In: *The symbolic and connectionist paradigms: Closing the gap* (1992), pp. 237–289.

[101] Trent E Lange and Michael G Dyer. “High level inferencing in a connectionist network”. In: *Connection Science* 1.2 (1989), pp. 181–217. ISSN: 0954-0091.

[102] George E Lewis. “Too many notes: Computers, complexity and culture in voyager”. In: *Leonardo Music Journal* 10 (2000), pp. 33–39.

[103] James W Lewis et al. “Human brain regions involved in recognizing environmental sounds”. In: *Cerebral cortex* 14.9 (2004), pp. 1008–1021. ISSN: 1047-3211.

[104] Marie-Christine Lichtlé. “The effect of an advertisement’s colour on emotions evoked by attitude towards the ad: The moderating role of the optimal stimulation level”. In: *International Journal of Advertising* 26.1 (2007), pp. 37–62. ISSN: 0265-0487.

[105] Aristid Lindenmayer. “Mathematical Models for Cellular Interactions in Development”. In: *Journal of Theoretical Biology* 18.3 (1968), pp. 280–299.

[106] Adam Linson, Chris Dobbyn, and Robin Laney. "Improvisation without representation: artificial intelligence and music". In: *in Proceedings of Music, Mind, and Invention: Creativity at the Intersection of Music and Computation*. Citeseer, 2012.

[107] Róisín Loughran, James McDermott, and Michael O'Neill. "Tonality driven piano compositions with grammatical evolution". In: *Evolutionary Computation (CEC), 2015 IEEE Congress on*. IEEE, 2015, pp. 2168–2175. ISBN: 1479974927.

[108] Peng Lu et al. "Discovering Harmony: A Hierarchical Colour Harmony Model for Aesthetics Assessment". In: *Computer Vision–ACCV 2014*. Springer, 2015, pp. 452–467. ISBN: 331916810X.

[109] Katherine Lubar. "Color Intervals : Applying Concepts of Musical Consonance and Dissonance to Color". In: *Leonardo* 37.2 (2004), pp. 127–132. ISSN: 0024-094X. DOI: `10.1162/0024094041139283`.

[110] René Magritte. "The Treachery of Images". In: *Oil on canvas* 231.2 (1928), pp. 1928–1929.

[111] Constantine Frithiof Malmberg. "The perception of consonance and dissonance." In: *Psychological Monographs* 25.2 (1918), pp. 93–133. ISSN: 0096-9753.

[112] Heather Mann et al. "Time–space synaesthesia–A cognitive advantage?" In: *Consciousness and cognition* 18.3 (2009), pp. 619–627. ISSN: 1053-8100.

[113] Keith Mashinter. "Calculating sensory dissonance: Some discrepancies arising from the models of Kameoka & Kuriyagawa, and Hutchinson & Knopoff". In: *Empirical Musicology Review* 1.2 (2006), pp. 65–84. ISSN: 1559-5749.

[114] Dan Mayer. *What Font Should I Use?: Five Principles for Choosing and Using Typefaces*. 2014. URL: `http://www.smashingmagazine.com/2010/12/14/what-font-should-i-use-five-principles-for-choosing-and-using-typefaces/`.

[115] John McDermott. “Learning to use analogies”. In: *Proceedings of the 6th international joint conference on Artificial intelligence-Volume 1*. Morgan Kaufmann Publishers Inc., 1979, pp. 568–576. ISBN: 0934613478.

[116] Gary E McGraw Jr. *Letter Spirit (part one): Emergent high-level perception of letters using fluid concepts*. 1995.

[117] Ingo G Meister et al. “Playing piano in the mind—an fMRI study on music imagery and performance in pianists”. In: *Cognitive Brain Research* 19.3 (2004), pp. 219–228. ISSN: 0926-6410.

[118] Julian F Miller and Peter Thomson. “Cartesian genetic programming”. In: *European Conference on Genetic Programming*. Springer, 2000, pp. 121–132.

[119] Melanie Mitchell. *Analogy-making as perception: A computer model*. Mit Press Cambridge, MA, 1993. ISBN: 0262132893.

[120] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. “Inceptionism: Going deeper into neural networks”. In: *Google Research Blog. Retrieved June* 20 (2015), p. 14.

[121] Campbell Murray, A Joseph Hoan Jr, and Feng-hsiung Hsu. *Deep Blue*. Tech. rep. 2001.

[122] Naila Murray, Luca Marchesotti, and Florent Perronnin. “AVA: A large-scale database for aesthetic visual analysis”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2408–2415. ISBN: 1467312266.

[123] Jay Neitz and Gerald H Jacobs. “Polymorphism of the long-wavelength cone in normal human colour vision”. In: (1986).

[124] Masashi Nishiyama et al. “Aesthetic quality classification of photographs based on color harmony”. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 33–40. ISBN: 1457703947.

[125] Martin A Nowak. “Five rules for the evolution of cooperation”. In: *science* 314.5805 (2006), pp. 1560–1563. ISSN: 0036-8075.

[126] Zena O’Connor. “Colour harmony revisited”. In: *Color Research and Application* 35.4 (2009), pp. 267–273. ISSN: 03612317. DOI: `10.1002/col.20578`.

[127] Atsuyuki Okabe et al. *Spatial tessellations: concepts and applications of Voronoi diagrams*. Vol. 501. John Wiley & Sons, 2009. ISBN: 047031785X.

[128] Michael O'Neil and Conor Ryan. "Grammatical evolution". In: *Grammatical Evolution*. Springer, 2001, pp. 33–47. ISBN: 1461350816.

[129] Kyle Orland. *Minecraft Draws Over $33 Million In Revenue From 1.8M Paying Customers*. 2011. URL: `http://www.gamasutra.com/view/news/33961`.

[130] Fabian Ostermann, Igor Vatolkin, and Günter Rudolph. "Evaluation Rules for Evolutionary Generation of Drum Patterns in Jazz Solos". In: *International Conference on Evolutionary and Biologically Inspired Music and Art*. Springer, 2017, pp. 246–261.

[131] Francois Pachet. "Representing knowledge used by jazz musicians". In: *Proceedings of the International Computer Music Conference*. INTERNATIONAL COMPUTER MUSIC ACCOCIATION, 1991, p. 285. ISBN: 1026-1087.

[132] Stephen E Palmer, Karen B Schloss, and Jonathan Sammartino. "Visual aesthetics and human preference". In: *Annual review of psychology* 64 (2013), pp. 77–107. ISSN: 0066-4308.

[133] Chaoyi Pang et al. "Topological sorts on DAGs". In: *Information Processing Letters* 115.2 (2015), pp. 298–301. ISSN: 0020-0190.

[134] Maximos A Kaliakatsos Papakostas, Andreas Floros, and Michael N Vrahatis. "evoDrummer : Deriving Rhythmic Patterns through Interactive Genetic Algorithms". In: (2013), pp. 25–36.

[135] Denys Parsons and Bernard Levin. *The directory of tunes and musical themes*. S. Brown Cambridge, England, 1975. ISBN: 090474700X.

[136] Kenneth Peacock. "Instruments to perform color-music: Two centuries of technological experimentation". In: *Leonardo* 21.4 (1988), pp. 397–406.

[137] Alexander Petrov. "Dynamic extension of episode representation in analogy-making in AMBR". In: *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society: August 13-15, 2000, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA*. Vol. 22. Psychology Press, 2000, p. 274. ISBN: 0805838791.

[138] Tony A Plate. "Holographic reduced representations". In: *Neural networks, IEEE transactions on* 6.3 (1995), pp. 623–641. ISSN: 1045-9227.

[139] Reinier Plomp and Willem J M Levelt. "Tonal consonance and critical bandwidth". In: *The journal of the Acoustical Society of America* 38.4 (1965), pp. 548–560. ISSN: 0001-4966.

[140] Riccardo Poli, James Kennedy, and Tim Blackwell. "Particle swarm optimization". In: *Swarm intelligence* 1.1 (2007), pp. 33–57. ISSN: 1935-3812.

[141] Riccardo Poli et al. *A field guide to genetic programming*. Lulu. com, 2008. ISBN: 1409200736.

[142] Jeffrey Pressing. "Cognitive complexity and the structure of musical patterns". In: *Proceedings of the 4th Conference of the Australasian Cognitive Science Society*. 1999.

[143] V S Ramachandran and William Hirstein. "The science of art: a neurological theory of aesthetic experience". In: *Journal of Consciousness Studies* 6.6 (1999), pp. 15–35.

[144] Geber Ramalho and Jean-Gabriel Ganascia. "Simulating creativity in jazz performance". In: *AAAI*. Vol. 94. 1994, pp. 108–113.

[145] Alexander Rimington. "Colour-music: the art of mobile colour". In: (1911).

[146] Nicolas Rothen and Beat Meier. "Higher prevalence of synaesthesia in art students". In: *Perception* 39.5 (2010), pp. 718–720. ISSN: 0301-0066.

[147] Romke Rouw and H Steven Scholte. "Personality and cognitive profiles of a general synesthetic trait". In: *Neuropsychologia* 88 (2016), pp. 35–48. ISSN: 0028-3932.

[148] Ariel Rubinstein. “Ranking the participants in a tournament”. In: *SIAM Journal on Applied Mathematics* 38.1 (1980), pp. 108–111. ISSN: 0036-1399.

[149] Stuart Russell, Peter Norvig, and Artificial Intelligence. “A modern approach”. In: *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25 (1995), p. 27.

[150] Oliver Sacks. *Musicophilia*. New York, NY, USA: Vintage Books, 2008.

[151] Eyal Sagi, Dedre Gentner, and Andrew Lovett. “What difference reveals about similarity.” In: *Cognitive science* 36.6 (2012), pp. 1019–50. ISSN: 1551-6709. DOI: `10.1111/j.1551-6709.2012.01250.x`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/22578040`.

[152] Arthur L Samuel. “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229. ISSN: 0018-8646.

[153] Adam N Sanborn, Thomas L Griffiths, and Richard M Shiffrin. “Uncovering mental representations with Markov chain Monte Carlo”. In: *Cognitive psychology* 60.2 (2010), pp. 63–106. ISSN: 0010-0285.

[154] Karen B Schloss and Stephen E Palmer. “Aesthetic response to color combinations: preference, harmony, and similarity”. In: *Attention, Perception, & Psychophysics* 73.2 (2011), pp. 551–571. ISSN: 1943-3921.

[155] Ilya Shmulevich and Dirk-Jan Povel. “Complexity measures of musical rhythms”. In: *Rhythm perception and production* (2000), pp. 239–244.

[156] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529.7587 (2016), pp. 484–489. ISSN: 0028-0836.

[157] Robert Lee Simpson Jr. *A Computer Model of Case-Based Reasoning in Problem Solving: An Investigation in the Domain of Dispute Mediation.* Tech. rep. 1985.

[158] John A Sloboda. “Music structure and emotional response: Some empirical findings”. In: *Psychology of music* 19.2 (1991), pp. 110–120. ISSN: 0305-7356.

[159] Paul Smolensky. "Tensor product variable binding and the representation of symbolic structures in connectionist systems". In: *Artificial intelligence* 46.1 (1990), pp. 159–216. ISSN: 0004-3702.

[160] Scott Sona Snibbe. *Bubble Harp*. 1999. URL: `http://www.snibbe.com/projects/interactive/bubbleharp`.

[161] Scott Sona Snibbe. *Motion Phone*. 1996. URL: `http://www.snibbe.com/projects/interactive/motionphone`.

[162] Scott Sona Snibbe and Golan Levin. "Interactive dynamic abstraction". In: *Proceedings of the 1st international symposium on Non-photorealistic animation and rendering*. ACM, 2000, pp. 21–29. ISBN: 1581132778.

[163] Jan Corfixen Sørensen et al. "DynaGrow - Multi-Objective Optimization for Energy Cost-Efficient Control of Supplemental Light in Greenhouses". In: *8th International Conference on Evolutionary Computation Theory and Applications* 1.Ijcci (2016), p. 9.

[164] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[165] Ferenc Szabó, Peter Bodrogi, and János Schanda. "Experimental modeling of colour harmony". In: *Color Research & Application* 35.1 (2010), pp. 34–49. ISSN: 1520-6378.

[166] Andranik S Tangiuane. *Artificial perception and music recognition*. Springer-Verlag New York, Inc., 1993. ISBN: 0387573941.

[167] Sebastian Thrun, Michael Montemerlo, and Andrei Aron. "Probabilistic Terrain Analysis For High-Speed Desert Driving." In: *Robotics: Science and Systems*. 2006, pp. 16–19.

[168] Michael Torke. *Color Music*. 1988.

[169] Laurel J Trainor and Becky M Heinmiller. "The development of evaluative responses to music:: infants prefer to listen to consonance over dissonance". In: *Infant Behavior and Development* 21.1 (1998), pp. 77–88. ISSN: 0163-6383.

[170] Hung-Cheng Tsai and Jyh-Rong Chou. "Automatic design support and image evaluation of two-coloured products using colour association and colour harmony scales and genetic algorithm". In: *Computer-Aided Design* 39.9 (2007), pp. 818–828. ISSN: 0010-4485.

[171] Christopher W Tyler. "Is Art Lawful?" In: *Science* 285.5428 (1999), pp. 673–674. DOI: `10.1126/science.285.5428.673`. URL: `http://www.sciencemag.org/content/285/5428/673.short`.

[172] Alexandra Uitdenbogerd and Justin Zobel. "Melodic matching techniques for large music databases". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 57–66. ISBN: 1581131518.

[173] Unity. *Procedurally generated content in Sir, You Are Being Hunted: a Unite Nordic 2013 presentation*. 2013. URL: `https://www.youtube.com/watch?v=GYYuhuarTA0`.

[174] Crétien Van Campen. "Synesthesia and artistic experimentation". In: *Psyche* 3.6 (1997).

[175] René Van den Brink and Robert P Gilles. "The outflow ranking method for weighted directed graphs". In: *European Journal of Operational Research* 193.2 (2009), pp. 484–491. ISSN: 0377-2217.

[176] Pantelis N. Vassilakis. "Auditory roughness as means of musical expression". In: *Selected Reports in Ethnomusicology* 12 (2005), pp. 119–144. URL: `http://www.acousticslab.org/papers/Vassilakis2005SRE.pdf`.

[177] Manuela M Veloso and Jaime G Carbonell. "Derivational Analogy in PRODIGY : Automating Case Acquisition , Storage , and Utilization". In: 278 (1993), pp. 249–278.

[178] Georg Von Békésy and Ernest Glen Wever. *Experiments in hearing*. Vol. 8. McGraw-Hill New York, 1960.

[179] Hermann Von Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green, 1912.

[180] Geoffrey Warner et al. "Modeling tax evasion with genetic algorithms". In: *Economics of Governance* 16.2 (2015), pp. 165–178. ISSN: 1435-6104.

[181] Marcus R Watson et al. "The prevalence of synaesthesia depends on early language learning". In: *Consciousness and Cognition* 48 (2017), pp. 212–231. ISSN: 1053-8100.

[182] Fabian L Wauthier and Michael I Jordan. "Efficient Ranking from Pairwise Comparisons". In: *Proceedings of the 30th International Conference on Machine Learning* 28 (2013), pp. 1–9.

[183] Joseph Weizenbaum. "ELIZA - A computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1996), pp. 36–45.

[184] William H Wilson et al. "The STAR-2 model for mapping hierarchically structured analogs". In: *The analogical mind, ed. D. Gentner, KJ Holyoak & BN Kokinov* (2001), pp. 125–160.

[185] Patrick H Winston. "Learning and reasoning by analogy". In: *Communications of the ACM* 23.12 (1980), pp. 689–703. ISSN: 0001-0782.

[186] Patrick H Winston. "Learning by creatifying transfer frames". In: *Artificial Intelligence* 10.2 (1978), pp. 147–172. ISSN: 0004-3702.

[187] Patrick H Winston. "Learning new principles from precedents and exercises". In: *Artificial Intelligence* 19.3 (1982), pp. 321–350. ISSN: 0004-3702.