Generative Lighting Design with Synesthesia

Master Thesis

Simon Borst Tyroll

Aalborg University
Institute for Architecture and Media Technology

Institute for Architecture and Media Technology
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Generative Lighting Design with Synesthesia

Theme:
Lighting Design

Project Period:
Spring Semester 2020

Project Group:

Participant(s):
Simon Borst Tyroll

Supervisor(s):
Stefania Serafin
George Palamas

Copies: 1

Page Numbers: 75

Date of Completion:
September 28, 2020

Abstract:

A system for audiovisual experiences is developed with inspiration from the human neurophysiological condition of synesthesia. Research suggests that synesthesia may be the result of a neural short circuit between the sensory systems in the brain, which is replicated by short circuiting two artificial sensory systems. It is emulated by the use of artificial neural networks (ANN) to replicate the sensory systems of the human brain. An autoencoder (AE) is used to recognise patterns in music, while a compositional pattern producing network (CPPN) is used to generate visual compositions. The AE extracts features from music and used to push the CPPN, which creates movements in the patterns it produces. In an experiment with 30 participants, the system was compared against a spectrogram and a CPPN producing movements based on Perlin noise. Results show that the developed system differentiates itself even in worst case conditions, and has potential for use as a generative design tool for dynamic lighting design. A lighting design was made using the system to show some different compositions of light and music. Further development is imperative to release the full potential of the system, by using a cybernetic approach to generative design.

Institut for Arkitektur og Medieteknologi
Aalborg Universitet
http://www.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Generativt Lys Design med Synestesi

Tema:
Lys Design

Projektperiode:
Forårssemestret 2020

Projektgruppe:

Deltager(e):
Simon Borst Tyroll

Vejleder(e):
Stefania Serafin
George Palamas

Oplagstal: 1

Sidetal: 75

Afleveringsdato:
28. september 2020

Abstract:

Et system til audiovisuelle oplevelser er udviklet med inspiration fra den menneskelige neurofysiologiske tilstand synestesi. Forskning tyder på, at synestesi kan være resultatet af en neural kortslutning mellem sensoriske systemer i hjernen, hvilket replikeres ved kortslutning af to kunstige sensoriske systemer. Det efterlignes ved brug af kunstige neurale netværk (ANN) til at replikere den menneskelige hjernes sensoriske systemer. En autoencoder (AE) bruges til at genkende strukturer i musik, mens et kompositionsmønsterproducerende netværk (CPPN) bruges til at generere visuelle kompositioner. Den uviklede AE finder mønstre i musikken der bruges til at skubbe til den uviklede CPPN hvilket skaber bevægelser i de mønstre den producerer. I et eksperiment med 30 deltagere blev systemet sammenlignet med et spektrogram og en CPPN, der producerede bevægelser baseret på Perlin-støj. Resultaterne viser, at det udviklede system adskiller sig selv i værste tilfælde og har potentiale til at blive brugt som et generativt designværktøj til dynamisk lysdesign. Et lysdesign blev lavet ved hjælp af systemet til at vise nogle forskellige kompositioner af lys og musik. Yderligere udvikling er oplagt for at frigøre systemets fulde potentiale ved hjælp af en cybernetisk tilgang til generativt design.

Contents

Preface **xiii**

1 Introduction **3**

2 Background **7**
2.1 Cybernetics 7
2.2 Generative Art 8
2.3 Generative Design 9
2.4 Computer Vision 10

3 Theory **15**
3.1 Our Auditory and Visual Perception 15
3.1.1 Gestalt Principles 15
3.1.2 In the Brain 17
3.2 Cross Modality 20
3.2.1 Synesthesia 21
3.3 Artificial Intelligence 23
3.3.1 Neural Networks 23
3.3.2 Genetic Algorithms 27

4 Implementation **35**
4.1 Imitating Human Hearing 35
4.1.1 Feature Extraction 37
4.1.2 Technical Performance Factors 38
4.2 Imitating the Retina 39
4.3 Evolving the Visual Cortex 40
4.4 Merging the Senses 41
4.5 Testing the Technology 44
4.6 Manual CPPN Construction 44

5 Test **47**
5.1 Preparations 47
5.1.1 Cross Modal Interaction 47
5.1.2 Perlin Noise 48
5.1.3 Spectrogram 48
5.2 Experiment Setup 48
5.2.1 Participants 48
5.2.2 Online Survey 49

6 Results **51**
6.1 Wilcoxon 52
6.1.1 Comparing Artificial Synesthesia with Perlin noise 52
6.1.2 Comparing Artificial Synesthesia with Spectrogram 52

7 Analysis **55**

8 Discussion **57**
8.1 The Experiment 57
8.1.1 Choice of Videos 57
8.1.2 Defining the Scale 58
8.1.3 The Aesthetics 58
8.1.4 Iterations of Design 58
8.2 Taking Control of AI 58
8.2.1 Morphology of Compositional Movements 60
8.3 Interaction Parameters 62
8.4 Cybernetic Design 62
8.5 The Future 62
8.5.1 Reference Projects for Development 63

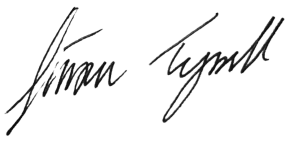
9 Design with Artificial Synesthesia **65**
9.1 The Virtual Design 65

10 Conclusion **69**

Bibliography **71**
03

Preface

Aalborg University, September 28, 2020



Simon Borst Tyroll
<styrol18@student.aau.dk>

Abstract

A system for audiovisual experiences is developed with inspiration from the human neurophysiological condition of synesthesia. Research suggests that synesthesia may be the result of a neural short circuit between the sensory systems in the brain, which is replicated by short circuiting two artificial sensory systems. It is emulated by the use of artificial neural networks (ANN) to replicate the sensory systems of the human brain. An autoencoder (AE) is used to recognise patterns in music, while a compositional pattern producing network (CPPN) is used to generate visual compositions. The AE extracts features from music and used to push the CPPN, which creates movements in the patterns it produces. In an experiment with 30 participants, the system was compared against a spectrogram and a CPPN producing movements based on Perlin noise. Results show that the developed system differentiates itself even in worst case conditions, and has potential for use as a generative design tool for dynamic lighting design. A lighting design was made using the system to show some different compositions of light and music. Further development is imperative to release the full potential of the system, by using a cybernetic approach to generative design.

Chapter 1

Introduction

Audiovisual experiences are deeply rooted in human nature with 90-95% of human perception processing coming from vision and hearing [54]. The fundamental understanding of human language is deeply rooted in the audiovisual perception as demonstrated by the McGurk effect (can be seen on Youtube[7]) which reveal how interconnected the senses are. The importance of this connection can be leveraged to create audiovisual atmospheres through communicating the movement of sound in light. Some research suggests that vision is dominantly responsible for spatial perception while hearing is responsible for the perception of time. This project describes an approach to developing a lighting design tool for the translation of music to light inspired by the phenomenon of synesthesia.

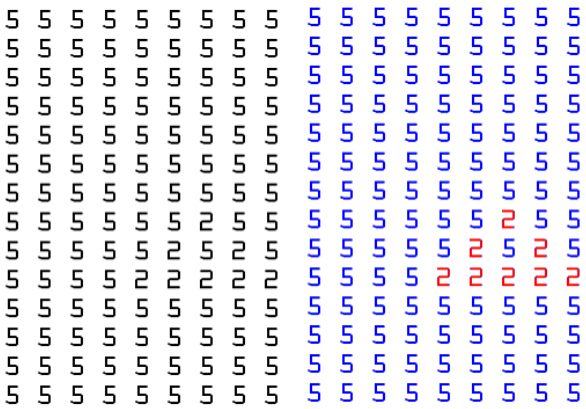


Figure 1.1: Grapheme-colour synesthesia makes it much faster to recognise different numbers. Left: a grid of numbers - 5 and 2. Right: example of grapheme-colour number associations.[41]

Synesthesia means "joint sensation", and is the opposite of anesthesia which means "no sensation". Synesthesia which applies to 4% of the population is a neurological condition in which information meant to stimulate one of your senses stimulates several of your senses. Depending on which kind of synesthesia it is,

it can allow people who have synesthesia (synesthetes) to actually see colours in music, or taste words, among other combinations. Grapheme-colour synesthetes sees every letter and number with colours, allowing them to identify different numbers efficiently as seen in fig. 1.1.

In a neuroscientific perspective, synesthesia is the involuntary response of a secondary sensory system due to a stimuli in a primary sensory system. This involuntary response is believed (at least in some cases) to be caused by additional neural pathways, connecting the sensory cortices in the brain. Some sensory systems are naturally very well connected like taste and smell.

On the topic of taste and smell, synesthesia was used as the inspiration for the Walt Disney movie Ratatouille in 2008, when the animator Michel Gagné (who is a synesthete) visualised synesthesia for the rat "Remy" who eats cheese and strawberry seen in fig. 1.2. Each sensory stimulus is unique, but as Remy eats both at the same time, the experience is new and amplified. Michael Gagné has created several other works, combining shapes and colors to visualise his synesthetic experiences with music seen in fig. 1.3 and fig. 1.4.

Figure 1.2: Picture from the Pixar movie "Ratatouille" where the rat named Remy experiences synesthesia of tastes as he combines the flavors of cheese and strawberry.[40]

Modern technology advancements in usable interfaces for Machine Learning (ML), enable the development of Artificial Neural Networks (ANN) that can be used to emulate functions of the human brain. The technology is even inspired by the structures of the brain which consists of interconnected neurons. This can be seen in some of Ryuichi Kurokawas audiovisual works as he explores the technologies to - in his words - create artworks in two themes: the reuse of nature, and creating synesthetic experiences[49]. Images from his work "ad/ab Atom" can be seen in fig. 1.5.

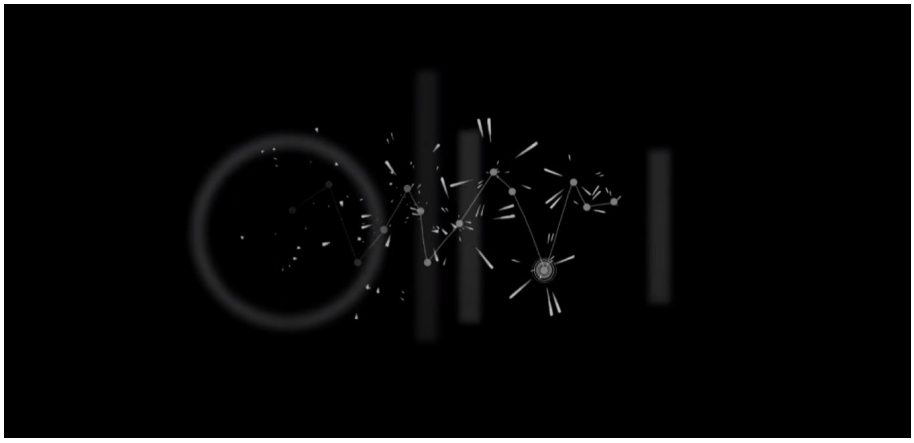


Figure 1.3: Still image from the animation by Michel Gagné "Sensology"[43]

Figure 1.4: Still image from the animation by Michel Gagné "Synesthesia"[48]

Research points toward a hierarchical structure in the human brain with specialised neural networks that handle specific tasks, which then are combined in the higher cognitive functions of the brain. This model can be emulated by creating specialised ANNs that perform specific tasks emulating the sensory systems. In this project, the sensory modalities in focus are the auditory system and the visual system. Artificial neural networks as a cross modal synthesizer can create a dynamic mapping from one sense to another, an artificial synesthesia.

Viewing an artificial neural network as a phenomenological entity that encodes sensory experiences into qualia[1], can offer an inherently subjective point of view on synesthetic experiences. Inspired by synesthesia, this project seeks to explore

[1] *Qualia* are the individual subjective experiences of consciousness, singular: *quale*.

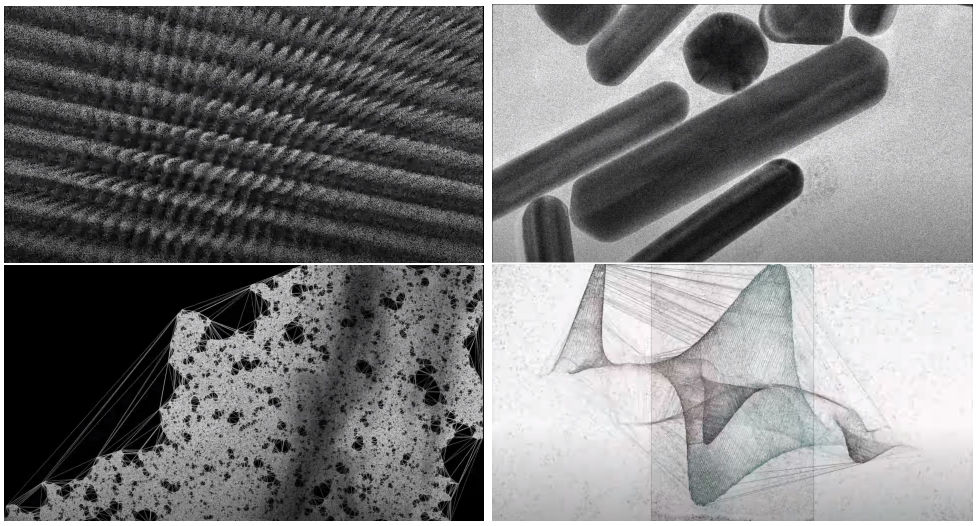


Figure 1.5: Still frames from Ryuichi Kurokawas audiovisual piece "ad/ab Atom" which can be seen on Youtube[42].

the opportunities of developing a dynamic lighting design system that integrates the movement of sound to light.

Lou Michel refers to the focal accents hierarchy of vision as: people, movement, brightness, high contrast, vivid color, strong patterns, meaning and combinations of the former. Especially two of these are highlighted: meaning and movement. Movement is the continuous perception through time. Meaning is the subjective lens that we understand the world through[34]. In letting a neural network derive meaning, it can be used through time to create meaning and movements in time.

What if software could enable the lighting designer to explore dynamic lighting designs through meaningful interactions and exploration? Exploring the behaviours of Artificial Synesthesia may show new perspectives for dynamic lighting design.

Chapter 2

Background

2.1 Cybernetics

> "Cybernetics is a young discipline which, like applied mathematics, cuts across the entrenched departments of natural science; the sky, the earth, the animals and the plants. Its interdisciplinary character emerges when it considers economy not as an economist, biology not as a biologist, engines not as an engineer. In each case its theme remains the same, namely, how systems regulate themselves, reproduce themselves, evolve and learn. Its high spot is the question of how they organize themselves." - Gordon Pask (1961)[38]

Defined in 1948 by Norbert Wiener as "the scientific study of control and communication in the animal and the machine"[61], cybernetics can be described as the exploration of the structures, constraints and possibilities of regulatory systems. In cybernetics a regulatory system has a cybernetic loop, meaning that it acts based on the causal relationship of its actions. Cybernetics comes the Greek word kubernētēs which means "steersman"[37] - we see the destination in the distance, so we correct our course through the cybernetic loop. In this way cybernetics is about steering to get to the goal, which is a fundamental property of intelligent systems. Although an applicable concept to many things in our lives like thermostats and automatically dimming screens, it is especially central to the modern discussions on adaptation and learning in AI systems. In a broader sense, cybernetics concerns itself with any system that has a goal or an objective. A system is defined as a set of things working together as parts of a mechanism or an interconnecting network. Cybernetics then is concerning itself with explaining anything that has a causal relationship with any other thing.

> "Once you see the world in a cybernetic way - through the cybernetic lens, all things are cybernetic, because all systems becomes part of this

> set of languages of: action, and sensing and comparing and understanding and taking a meta-view. This is fundamental to intelligence because if I know what I want, and I act to achieve what I want - successfully - then that is the best definition of intelligence that I know. And of course we all want to be intelligent, it is part of what drives us as humans, and because cybernetics is about physical, technological, biological and social systems - all - it is the most comprehensive language to describe these things." - Paul Pangaro [37]

2.2 Generative Art

> "Much of the innovation today is not achieved within the precious bubble of fine art, but by those who work in the industries of popular culture–computer graphics, film, music videos, games, robotics and the Internet" - Jon McCormack (2003)[16]

With its theoretical roots in cybernetics and general systems theory, the generative arts, along with computer arts and electronic arts, have been technologically enabled especially in the last 20 years as Machine Learning and AI techniques have gained massive mainstream traction. It was defined as "Cybernetic Vision" in 1966/1967 by Roy Ascott who became an influential figure in the cybernetic arts[11]. Another theoretical seed from cybernetic theory is the methodologies behind the digitised concepts of evolution and self-organization, which also takes inspiration from cognitive science and artificial life. An example of a simulation of life is Conway's Game of Life created by mathematician John Horton Conway in 1970. An image of this can be seen in fig. 2.1.

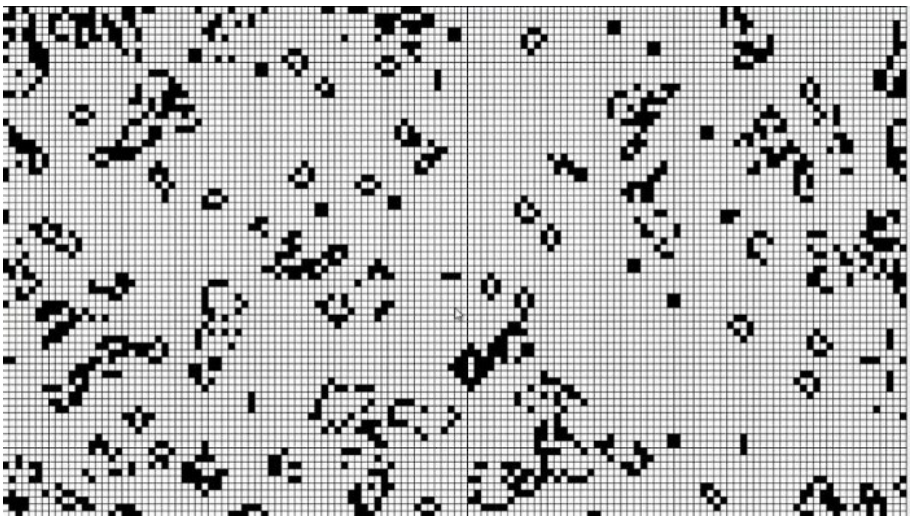


Figure 2.1: A still image of the Game of Life simulation made by John Horton Conway.[20]

While generative art is often coupled with use of computers, Kenneth Martin showed that it does not have to be so as seen in "Chance,Order,Change" seen in fig. 2.2. By choosing geometrical figures and defining some rules of proportion, he would randomize the layout of the artwork[16]. The rules decide the boundaries of randomness.

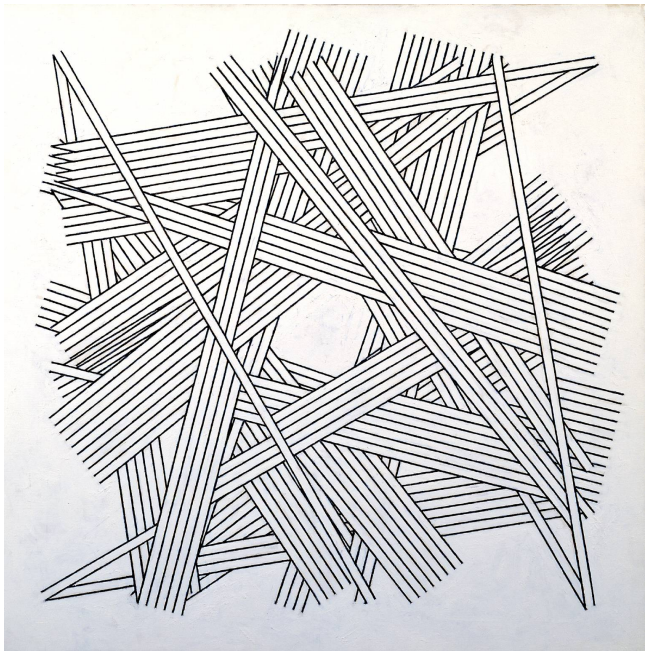


Figure 2.2: "Chance,Order,Change" by Kenneth Martin (1978)[51]

Rules of randomness is essential in the definition of generative art as it seems to give the systems a degree of autonomy [16]. The rules are in relation to the conscious decisions of the artist, which becomes a self interpretation of ideas formulated through limitations of randomness.

> "If the cybernetic spirit constitutes the predominant attitude of the modern era, the computer is the supreme tool that its technology has produced. Used in conjunction with synthetic materials it can be expected to open up paths of radical change and invention in art.... The interaction of man and computer in some creative endeavour, involving the heightening of imaginative thought, is to be expected." - Roy Ascott

2.3 Generative Design

Generative design - in continuation of generative art - is the process of designing through the use of a system that will find a variety of solutions within a defined

solution space. Just like generative art is about exploring randomness with constraints, generative design is about using constraints to have an algorithm explore different solutions that fits within the boundaries of the constraints.

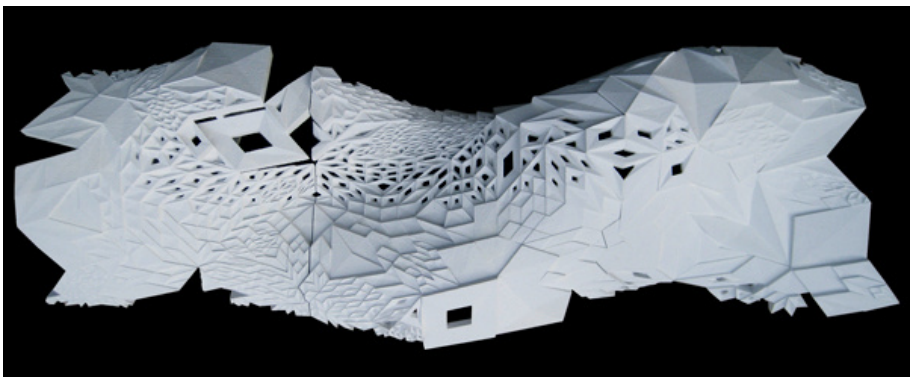


Figure 2.3: Example of a 3D structure as designed with the use of generative design.[19]

The power of generative systems comes from its innate random nature, as it seeks to explore every possible solution within the solution space. The systematisation of randomness is a way for designers to explore design possibilities as randomness itself can show new, possibly even revolutionary ideas. The question is - what are the constraints.

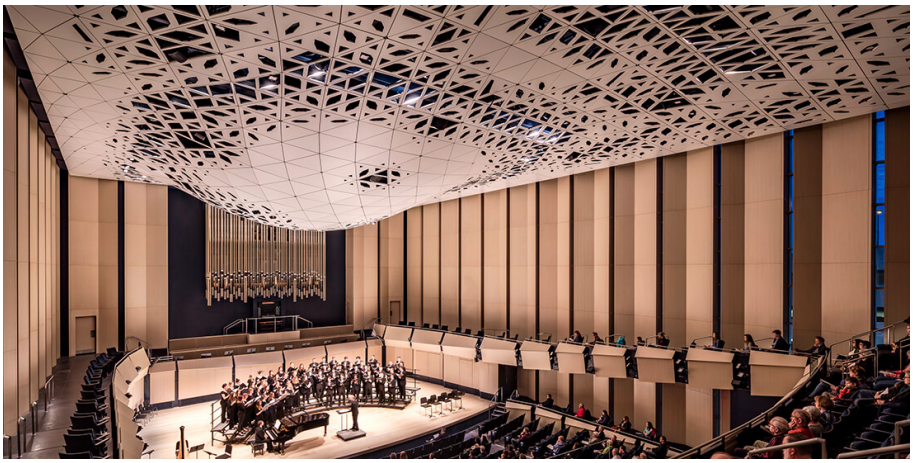


Figure 2.4: Acoustic 3D design for University of Iowa's Voxman School of Music concert hall made with generative design. [14]

2.4 Computer Vision

Deep dream as made by Alexander Mordvintsev from Google, is the result of a deep neural network that produces psychedelic images based on its knowledge on

the contents of pictures. This means that any image fed into the deep dream network, will be correlated with the knowledge of the network. Said in another way, if the network was trained on images of cats, it will look for cat-like features (ears, paws, eyes) in the image and merge those features into the image. An example of this can be seen in fig. 2.5. The Google deep dream network can also be used to do style transfers between images, which can be seen in fig. 2.6 A video of the journey through the depth of the deep dream network can be seen on Youtube[24].

Figure 2.5: An image made by Google Deep Dream where the network has found animalistic features in the night sky[22].

Figure 2.6: An image made by Google Deep Dream where the network has combined the style of one image with the content of another imagedeepdreamgenerator.

Another project by Mordvintsev shows the implementation of a Compositional Pattern Producing Network along with a Convolutional Neural Network that produces "Light Paintings" of the images used for training. Examples can be seen in fig. 2.7.

Figure 2.7: 9 pictures produced by Mordvintsev's "Light Painting" CPPN.

Chapter 3

Theory

3.1 Our Auditory and Visual Perception

Our perception of the world is the end of a long chain of neural events in the brain that gives us an illusion of an instantaneous event. The experience that we see the world as it is, when it is, is an illusion of the brain, guessing by it's best ability to make probabilistic inferences about the world.

Visual illusions helps illuminate some of these properties as they seek to challenge our perceptual limits. A quite famous illusion, the Ponzo illusion makes it seem as though the two gray lines are of different sizes as seen in fig. 3.1. Fig. 3.2 shows the "turning the tables" illusion that show two tables that are the same size, although it is hard to believe. Lastly the Kaniza illusion shows one of the five Gestalt principles known as closure as it looks as if there is a white triangle on top of 3 circles and another outlined triangle seen in fig. 3.3.

> "..when functioning properly, our perceptual system is supposed to distort the world we see and hear"[28, p. 97]

3.1.1 Gestalt Principles

In continuation of the Kaniza illusion, which exemplified the gestalt law of closure, it is one of the five laws of the Gestalt principles. According to the Gestalt Laws of Organization, the mind arranges incoming groups of perceptual stimuli. Perceptual information that abide by the laws of Gestalt is characterized by simplicity, neatness, order, and consists of the least amount of features in it's structural parts. "Good Gestalt" is perceptual information that is easy to process because it abides by the Gestalt laws. While Gestalt theory is most often correlated with visual perception, it is also a part of auditory perception system as it pertains to the temporal structures of stimuli as well as the spatial.

Figure 3.1: Ponzo illusion[28]

- Law of Similarity - Elements that have similar characteristics is perceived as a set.
- Law of Proximity - Elements that are close to each other are perceived as a group.
- Law of Closure - Only segments of the whole is necessary for shape identification.
- Law of Continuity - Elements that are arranged on a continuous line are perceived to be more related than elements not on the line.
- Law of Figure/Ground - Elements are perceived as being either in the foreground, or the background.

[34]

These principles makes the perception systems and grant us the ability to understand and synthesize perceptual information very quickly. It is a combination of filtering information into features, and recognising them as structures that we can recognise as an object, whether it is physical or not. A person who notices something almost no one else does is recognised as "perceptive", this person may have an aptitude for noticing different features, and/or combining them differently.

> "How does the brain figure out, from this disorganized mixture of molecules beating against a membrane, what is out there in the world? In particular, how does it do this with music? It does this through

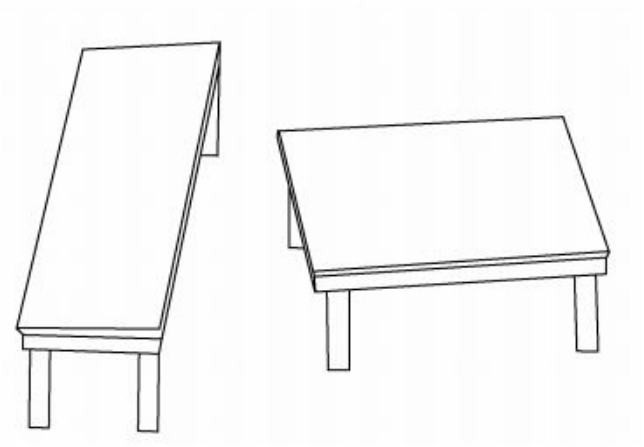


Figure 3.2: "Turning the tables" illusion[28]

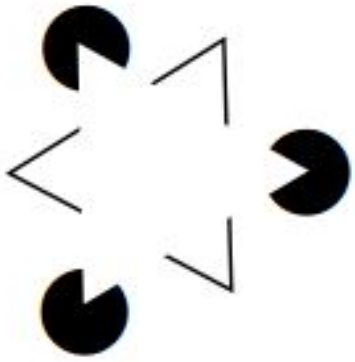


Figure 3.3: Kaniza illusion[28]

a process of feature extraction, followed by another process of feature integration."[28]

3.1.2 In the Brain

"Musical activity involves nearly every region of the brain that we know about, and nearly every neural subsystem. Different areas in the aspects of the music are handled by different neural regions—the brain uses functional segregation for music processing, and employs a system of feature detectors whose job it is to analyze specific aspects of the musical signal, such as pitch, tempo, timbre, and so on."[28, p. 83]

The neural networks of the brain consists of 100 billion neurons that are inter-

connected. That is the same amount of neurons as there are stars in the Milky Way galaxy. Usually a neuron has between 1000 and 10000 connections meaning that the connectivity of the brain has a potential of approximately 500 trillion unique neural connections. In the context of artificial neural networks, these numbers are astronomical and gives some perspective on the challenge of replicating human phenomena. As the number of neurons grows, so does the computational requirements as the amount of connections to be calculated grows exponentially:

$$Connections = 2(n * (n-1)/2)) \tag{3.1}$$

- For 2 neurons there are 2 combinations for how they can be connected
- For 3 neurons there are 8 combinations
- For 4 neurons there are 64 combinations
- For 5 neurons there are 1,024 combinations
- For 6 neurons there are 32,768 combinations[28]

The computational power of the neuron connections of the brain is - among many other things - used for feature extraction from the perceptual information. The sensory organs perform the initial filtering of information and decide where to look for features[21]. It can be described as a perceptual window because it is the lens through which we can detect the world. For vision, the retina filters wavelengths of light into what we know as visible light (380-740 nm), and for hearing, the cochlea filters waves of pressure into the audible frequency range (20 Hz to 20 KHz)[28]. The filtered information is transferred to the brain which can now extract features. When we talk about features, we are talking about any representation of the data that is showing a type of structure. A feature can be of a low abstraction level such as edges, circles etc. for vision, and volume, tone, timbre etc. for audio. A higher level abstraction feature consists of a combination of the lower level features that when combined can give the impression of a square or the sound of a trombone. This hierarchy of feature levels builds on itself as lower level features are combined to creature features of higher and higher abstraction, the peak of which is the emotional response - the *meaning*. As the features grow more and more complex, it involves more of the brain as all of the information is synthesized. The process of computing the simpler features independently to then later combine the features is called bottom-up processing[28, p. 101]. As the sensory information has been integrated and the features recognised, the sensory information seems to no longer be represented independently. At some point in the hierarchical extraction of meaning, the senses merge. [26]

> "The brain is a massively parallel device, with operations distributed widely throughout. There is no single language center, nor is there a

> single music center. Rather, there are regions that perform component operations, and other regions that coordinate the bringing together of this information."[28, p. 85]

Top-down processing is what happens when our centers for higher-level processing - mostly the frontal cortex - makes inferences about what is about to happen next based on a number of factors. For instance in music:

- what has already come before in the piece of music we're hearing;
- what we remember will come next if the music is familiar;
- what we expect will come next if the genre or style is familiar, based on previous exposure to this style of music
- any other information that we have synthesized, including spontaneous reactions to sudden movements etc.

Top-down processing can take control of the bottom-up processes in such cases and make us misperceive information as our experiences and reflexes can overrule our perception system[28, p. 103]. As neuroscientist at University College of London, Beau Lotto describes it: "Whenever we open our eyes, we never see what is there. We only see what was useful for us to see in the past."[13]

In this way, the human perceptual system is a hierarchy of filtration that starts by using the sensory organs to filter information from the world into the simplest representation: photon wave frequencies in the eyes, and pressure wave frequencies for the ear. The brain then makes calculated inferences about this information by extracting the most important information by applying another filtering which creates new features, and this process is repeated as the process ends in the oldest parts of our brain. The concept of higher level functions of the brain being connected to all of the senses also seems intuitive if we think about the metaphorical synonyms we use to describe experiences in different sensory systems: bright, deep, sharp, soft, vibrant, colorful, balanced. All of these terms can conjure up different contexts of emotional experiences, whether it be in light, taste, music, smell or touch, or even combinations of these.

> "At a deeper level, the emotions we experience in response to music involve structures deep in the primitive, reptilian regions of the cerebellar vermis, and the amygdala — the heart of emotional processing in the cortex."[28, p. 85]

3.2 Cross Modality

As we perceive the world around us with our senses, we may think that our perception is a result of the combination of all the sensory systems which handle information independently. However, the senses are not independent as information from two or more sensory modalities are integrated and synthesized. This means that the experience of one type of sensory information can change as another sense is stimulated[26].

The motion-bounce illusion shows two disks that are equal in size following two linear trajectories that intersect at a point. There are two perceptual options of how to understand this information. First the two disks can be perceived to continue their trajectory overlapping shortly. Second the disks can be perceived to bounce into each other, changing the trajectory. The experiment can be seen on website [12]. Without sound, only 20% of subjects thought the disks to be bouncing off of each other. But with the introduction of an impact sound at the moment of interaction between the disks, 60% of subjects perceived the disks to be bouncing[26].

Another audio-visual illusion is the McGurk effect, demonstrated by Harry McGurk and John MacDonald in "Hearing Lips and Seeing Voices" in 1976[33]. 98% of adults think they are hearing the word "DA" when they hear the sound "BA" and see the lip movements of "GA". McGurk and MacDonald also mention that the auditory features are similar for "BA" and "DA", and the visual features are similar for "GA" and "DA". Therefore it is assumed that the synthesis of similar features from the modalities result in a "DA" experience. A form of fusion of features. In another experiment in the theme of speech perception, subjects were found to to have an improvement of speech perception between 40% and 80% when they could also see the speaker's face under noisy conditions[26].

The ventriloquist effect was found by presenting subjects to conflicting spatial information in the visual and auditory modality. The subject was asked to point to where the perceived stimulus originated, and after presenting the subject to the visual and auditory stimulus independently, they were asked to point towards the stimulus while being presented with conflicting auditory and visual information. They found that the visual sense dominated the spatial perception as the subjects found the location of the auditory stimulus to be much closer to the visual stimulus, while their perception of the visual stimulus was hardly altered by the presence of a conflicting auditory stimulus[26]. The fact that one sensory modality is dominant over another in certain perception tasks, is believed to be the result of one sensory modality being better suited for perceptual decoding of a stimulus. The modality appropriateness hypothesis suggests vision as the origin of the concept of space, and the auditory system as the timekeeper. This is supported by an experiment where it was found that visual flicker fusion (the point where flicker

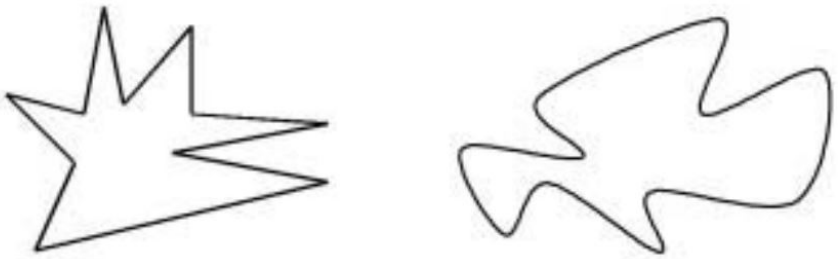


Figure 3.4: Two shapes, with each its name - Bouba, and Kiki. The sharp inflections of the shape on the left is alike the sharp inflections of the phonemic sound of Kiki, while the rounded contours of the shape on the right is alike the rounded phonemic sound of Bouba.[39]

is no longer perceived) occurs between 50 Hz and 100 Hz, while auditory tonal modulations could be detected at up to 400 Hz, suggesting that hearing is better at distinguishing temporal information [26].

> "..it has often been claimed, especially since Kant, that music is an art of time, if not *the* art of time"[9]

The famous study of the "bouba" and "kiki" effect shows the intuitive cross modal metaphoric understanding of "sharpness" and "softness" as seen in fig. 3.4. First developed by Wolfgang Köhler in 1929, the mapping of "Kiki" as the shape on the left, and "Bouba" as the shape on the right is found in 95% of people across cultures[39].

3.2.1 Synesthesia

Synesthesia which means "joined senses" is the term for the condition in which a sensory stimulus, elicits an additional response in another. The origin stimulus is also known as "the inducer", whilst the response to it is known as "the concurrent". Synesthesia is often described in conjunction with a question: "What color is the letter A?" or "What is the sound of orange?", and while these questions do engage the immediate concept of synesthesia, it doesn't fully grasp the phenomenological consequences it entails, as it highlights the philosophical problem of describing the connection between the physical world and the subjective world that we experience.

The opposite phenomenon to synesthesia is "anesthesia" which means "without sensation" gives a comparative meaning to what synesthesia is. Wassily Kandinsky, the Russian painter which allegedly had synesthesia, created paintings which he wanted people to understand as a multi sensory gestalt combining the senses in a cross modal phenomenon. When you look at Kandinsky's painting "Contrasts" fig. 3.5, which sound would you feel matches better with it? A harmonic, or disharmonic sound?

Figure 3.5: "Contrasts" by Kandinsky [1]

People with synesthesia known as "synesthetes", historically excel in the generation of creative ideas, and are more likely to follow artistic pursuits [44, p. 621]. Although there have been some debate on the premises of classifying synesthetes, and because the research of it has been reliant on subjective reporting, synesthesia has been shown to be not only a real description of the synesthetes' phenomenological experience of the world, but also a neurophysiological condition. In other words, synesthetes' brains are wired differently than the statistically average non-synesthete. It has been found that synesthetes excel in memory tasks related to their synesthesia[44, p. 707], and there is a strong association between synesthesia and vivid mental imagery[44, p. 730]. There are theories about manual induction of synesthetic experiences with LSD, and reports of people reporting a feeling of having synesthesia-like experiences after ingestion of LSD[31]. Although, there are some discrepancies in this theory as the drug induced synesthesia is reliant on serotonin, while "genuine" synesthesia is a neurological reaction[45]. Nonetheless, it is suggested that the subjective experiences of drug-induced and genuine synesthesia may be similar and has some psychedelic traits.

Origins

Development of the brain consists mainly of two processes: first the generation of new connections between neurons by producing new synapses, and second the strengthening of stimulated connections and pruning of unused connections[44, p. 46]. The formation and pruning of neural connections occurs in childhood until between the age of 7 and 9. These neural connections in early childhood include the connections between areas of the cerebral cortex that receives information from the sensory systems like vision and hearing. Unlike in adults, for newborns it has been found that sound amplifies the neural response to touch[44, p. 47]. In another study it has been found that while adults' neural response to human speech is only auditory, the neural responses of young infants are both activating the au-

ditory and visual cortex. This neural connection diminishes over the first 3 years of infancy suggesting a pruning of the multi sensory connection[44, p. 47]. In other words, evidence shows that we have all had the potential for synesthesia through a hyperconnectivity of the brain, but because of the neural pruning those neural connections have been lost. Synesthetes appear to have a genetic disposition for a less complete pruning stage of the neural development, leaving some neural connections behind giving them multi sensory experiences through a single sensory system.[44, p. 49]

> "The recent research (..) lend cogent support to the hypothesis that all individuals experience something like synesthesia as infants, with remnants of these cross-modal associations still observable in adulthood, either explicitly in synesthetes or implicitly in all other people."[44, p. 58]

Cognition Perception Divide

It is recognised in the field of synesthesia studies that there is a broad application for the synesthesia term that covers everything from abnormalities in connections of the sensory systems to the cognitive functionalities of metaphorical thinking [44]. For the applications of this project, the system development and inspiration is based on the crossing of sensory systems of the brains, as this is also the lowest abstraction level of feature integration in the brain. We must start from the bottom to build a bottom-up perception system.

3.3 Artificial Intelligence

Artificial Intelligence (AI), is the blanketing term for technology that allows a machine to emulate human behaviour. Machine Learning (ML) covers a subset of technologies that allows machine to improve it's behaviour over time as it gains more and more experience.

3.3.1 Neural Networks

An Artificial Neural Network (ANN) is a data structure of artificial neurons with connections between them. These connections are weighted meaning that a connection between two neurons is a gradient between activated and deactivated. In a typical ANN these neurons are organised into layers starting with an input layer, hidden layers, and an output layer. An ANN can have many different configurations with more hidden layers of varying sizes. A typical ANN is seen in fig. 3.6 with a 5-5-5 configuration with fully connected layers, meaning that every neuron in every layer has a connection with every other neuron in the preceding and next layer.

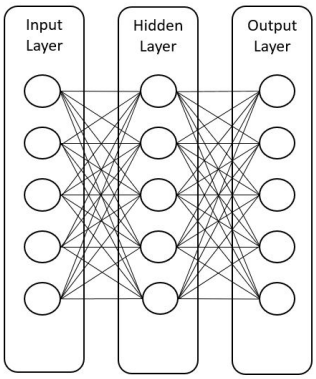


Figure 3.6: Typical ANN configuration with 3 fully connected layers

ANNs are inspired by the biological neural networks found in the brains of animals. And some research shows that some types of ANNs known as Convolutional Neural Networks (CNN) show some resemblance of gestalt principles in the individual layers of the network[10], which means that the nature of an ANN resembles the hierarchical structure of the human brain. The applications for ANNs are plentiful and are well integrated into everyday life as the technology powers everyday appliances like autonomous driving, voice recognition and ad placements for instance.

How does a neural network learn?

Learning in an ANN happens by changing the weights of the connections between the neurons. But how does it make a decision on when and how to do so? It does so by calculating the loss. Loss is the distance between the output of the network in its current state, and the desired output. The calculation of this is called a cost function. If we were to train a network to be able to tell if an image is an image of a cat or a dog, we would train it on images labeled as "cat" or "dog" and present it with a random image and ask it to tell us whether it thinks it is a cat or a dog. Then it can evaluate via the cost function whether it correctly predicted if it is a cat or a dog. This process is repeated for as many times as it takes until it has either learned to categorize cats and dogs, or until the success rate is satisfying. Each training cycle is known as an epoch.

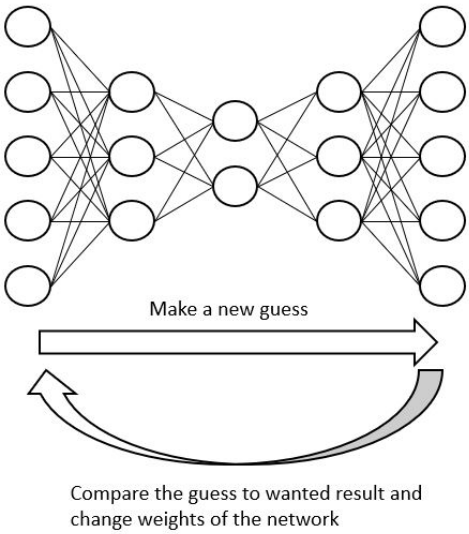


Figure 3.7: Every epoch the ANN makes a new guess which is used to guide the next guess

The optimization function is the method of determining how to update the weights of the network based on the results of the cost function. The activation functions of the neurons decide how to scale the value of the neuron before it is fed into the next weighted connection. The training data set and test data set is what is used as the input for network during training The training dataset is used as input to produce some output that can be compared against the test data set. In the configuration of an ANN, the parameters which needs to be determined before beginning the training are called hyperparameters.

The Neurons

Starting from the first layer in the network, this is where we place the input, this could be a row of numbers that represent pixel values of an image. This data is propagated through the network through what is known as feedforward. The value of each input neuron is propagated through the weights of its connections to the following layer. In every neuron of the following layer, the weighted values of all the connected neurons are then summed together and put through an activation function in the neuron.

The activation function of a neuron decides how activated the neuron will be by the incoming values. For instance the activation function rectified linear unit (ReLU) will multiply any value above 0 with 1 making it scale linearly with positive values.

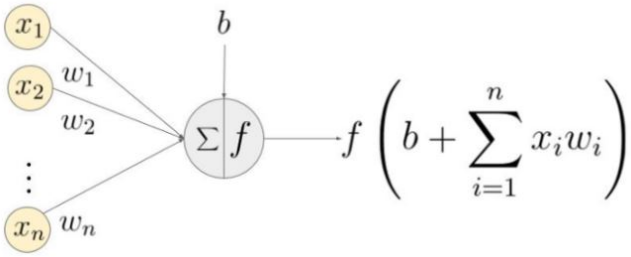


Figure 3.8: The weighted values (w1, w2.., wn) of the connected neurons (x1, x2.., xn) are summed along with a bias[8]

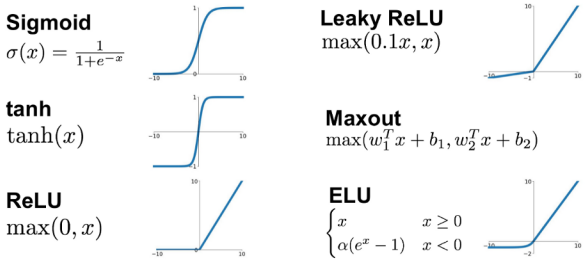


Figure 3.9: Examples of different activation functions showing different scalings of neuron values [18]

Autoencoder

As an ANN can have many configurations, we must choose one for the purpose of this project, and in the scope of replicating a human sensory center, this will be feature extraction. An autoencoder (AE) is an ANN configuration that seeks to do just that by training it to compress a variety of data with minimal loss. As it tries to perform compression without losing data, it will naturally find commonalities within the data set as it allows it to do feature representations as efficiently as possible. Practically, this means that within a dataset of numbers - let us say 64 - it will find a way to represent that specific combination of all the 64 with a smaller amount of numbers - let us say 9. Those 9 numbers will be a feature representation of those 64 numbers as seen in fig. 3.10.

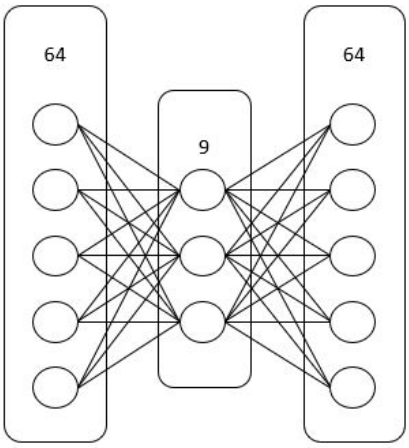


Figure 3.10: An autoencoder structure that compresses the 64 input values to 9 values in the middle, and uses the 9 values to rebuild the 64 output values. The output is compared to the input, and that teaches the AE to create efficient features for the dataset.

This process of feature extraction mimics the way our brain sorts through information and performs feature recognition. A trained AE is built to filter data and look for specific structures in a stream of data.

3.3.2 Genetic Algorithms

A genetic algorithm (GA) is an alternative way of searching for a solution by instead of evaluating the cost function of one ANN, it can evaluate the fitness of a variety of different ANNs and breed the best performing of them, thereby "training" the networks which has the best traits for the task. GA is a method of solution space search through creating a population of genomes and breed the most successful of them and add a random chance of mutation of their offspring. The method draws inspiration from Darwin's theory of natural evolution where "survival of the fittest" decides who can successfully breed and create new offspring[32]. But what does it mean to be fit for an algorithm? First we must lay the foundation define the terminology for describing neuroevolutionary practices. Since the context of the use GA in this project are ANNs, we will use this as a basis for specifying the terminology. As GAs are inspired by biological evolution, the terminology to describe the behaviour of algorithms are also borrowed from the bio-genetic vocabulary.

Terminology

Neuroevolution is the blanket term for methods that uses GA to evolve ANNs.

The gene is the individual variable that together with all the other genes makes up the genome. These genes in the context of ANNs are the activation functions,

weights, connections and all of the information that comprises an ANN architecture. **The genome** is the sum of all the genes that together make up an ANN. The genome can also be called the agent in the context of how the genome behaves in an environment.

The population is a number of genomes in a generation that restricts the genetic diversity. Optimally this would be infinitely big since it would yield the maximal genetic diversity, but due to limitations in computing power, this has to be restricted.

Genotype is the genetic encoding that describes a behaviour in a given environment. In biology the genotype is DNA which decides eye-color. **Phenotype** is the behaviour that the genotype elicits in the given environment. In biology, the phenotype of DNA is for instance eye-color. It is the behaviour of the genotype encoding.

There are also some genetic operators that needs defining, as they describe how the genomes are evolved using GAs.

Crossover

The crossover is the recombination of the genes from parent genomes in the existing population. Breeding successful genomes and recombining their genetic makeup has the potential to create better offspring. The genes can be recombined using different methods. As seen in fig. 3.11, single-point crossover slices a random point in the parent genomes and combines the genes from each side of the slice from each of the parents. Two-point crossover functions in the same manner as single-point, with one additional gene slice. Uniform crossover chooses a parent randomly for every gene to be combined in the child genome.

Mutation

The mutation operator is the random mutation of the genes in the genome that has the role of preventing stagnation in the population, when the genomes becomes too similar. Mutation diversifies the genetic population and makes the population able to explore new behaviours. There are some different methods for mutation operations as seen in fig. 3.12. Bit inversion will invert a random value in the genome, this could for instance be whether a neuron is connected to the next neuron or not. Order change will change the position of two genes in the genome, this could for instance be the position of two neurons in the network, each with different activation functions. Value change adds a value to a random gene in the genome, this could be the weight of the connection between two neurons. Lastly, gene expression change randomly adds or removes genes from the genome, this could be adding a new neuron to the network, or adding a new weight.

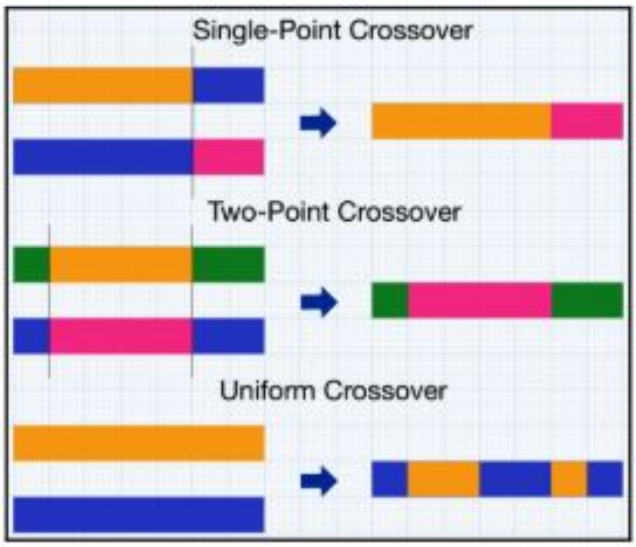


Figure 3.11: Three different genetic crossover methods. [36]

Fitness

A crucial operator for defining how successful a genome is, is the fitness function. This can be compared to the cost function of the ANN as it seeks to evaluate whether a genome is behaving well in a given environment. For example, if we wanted to evolve a navigation algorithm for autonomous driving, a fitness function could be the distance between a goal point and the behaviour of the algorithm. The closer the algorithm comes to the goal, the higher the fitness. The fitness function is used to evaluate every genome in the population every generation.

Generations

A generation is the evaluation cycle for every genome in the population. Before the evolutional training commences, the population is populated by genomes. Then their fitness is individually evaluated and ranked so the best percentage of the population can breed and mutate as seen in fig. 3.13. The middle genomes survives as they can still learn to exert better behaviour through mutation. The worst are removed to leave space for new genomes that may exhibit better behaviour.

NEAT

Neuroevolution of Augmented Topologies (NEAT) is an evolution of the above genetic algorithm principles. While ANNs typically have an architecture which remains static throughout training (only the weights are changed), NEAT seeks to implement GA methodologies to the development of an ANN in the search of

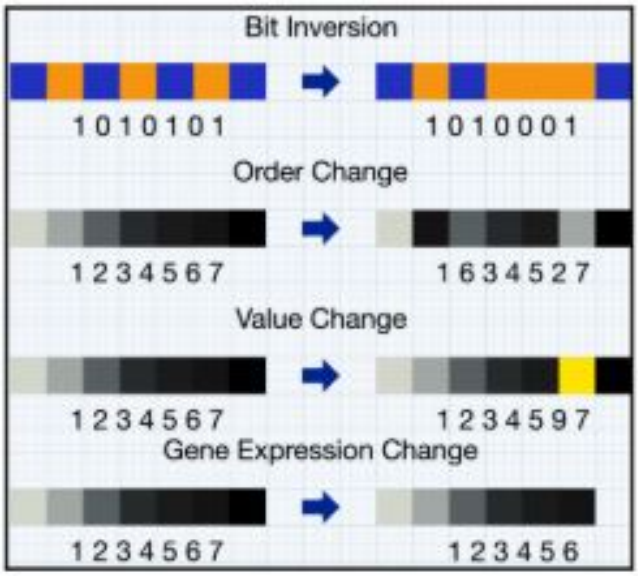


Figure 3.12: The different kinds of genome mutation that can occur. [36]

the best architecture. This means that the potential of the search becomes much larger in scale as it is not limited by the configuration of the ANN. The NEAT algorithm will colonize the population with very simple genomes that has a variety of topologies meaning that they have not only a variety of weights, but also a variety of neurons with asymmetric connections and different activation functions. These genomes gradually increase in complexity over generations which is known as complexification. As seen in fig. 3.14, the ANN has a typical symmetric topology with distinct layers that are only connected to the neurons in the following layer. An example of a NEAT topology can be seen below in fig. 3.14 where compared to a traditional ANN, the neurons are not organised into layers, but rather an arbitrary topology that can have any possible configuration with the predetermined input and output neurons[36, p. 17]. The goal of NEAT is to minimize the complexity of the genomes in the population, thereby limiting the search space and gradually increasing it as genomes become more and more complex.

A genome can have a worse fitness than its competition, it may contain a stronger base topology that may prove stronger with further evolution. To save such genomes from the evolutionary pressure of fitness competition, NEAT attempts to save such genomes by limiting the competition range in the population, meaning that not all genomes compete against the entire population, but a smaller bracket within it. This is called speciation, and makes sure that there is an allowance for less fit genomes to evolve. The age of the genomes are tracked by assigning the genomes an innovation number that is also used for matching genomes for breeding.

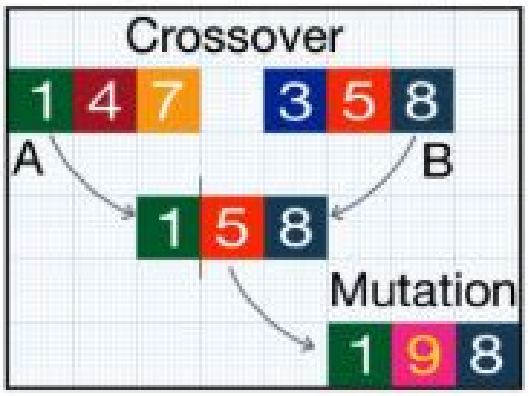


Figure 3.13: An example of the neuroevolutionary process of breeding the best genomes first and mutating it after which happens every generation.[36]

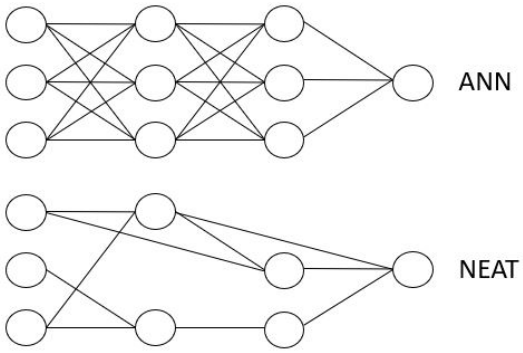


Figure 3.14: Two neural network topologies. Top: a typical ANN topology with layers of neurons that are all interconnected. Below: An example of a NEAT topology that can arise from the evolutionary method that can create arbitrary networks with neurons that can be connected to any other neuron.

The question we ask when we define the solution we're looking for in an ANN which guides the system design is: "which ANN configuration is best fitted for the task?". While the question we ask in designing the solution search for a NEAT system is: "how do we limit the search space to find the best configuration possible as fast as possible?"

HyperNEAT

The Fourier series explains how any signal can be decomposed into it's constituent part signals of different frequencies. This means that any signal can also be built by combining multiple signals together, which is how the HyperNEAT algorithm approaches a genetic encoding scheme.

> "To form an image on your retina, the lens in your eye performs Fourier transformations on the light that enters it," he explains. This tool is

> truly ubiquitous in nature, as our eyes and ears have subconsciously performed the Fourier transform to interpret sound and light waves for millions of years." - Ronald Coifman, Professor of Mathematics, Yale University[56]

In taking inspiration from the human brain, we must also do so with a structure that imitates the brain. Neuroscientists have found that spatial structure is essential to all tasks of the brain - from the perception of sensory information to cognitive and abstract thinking[36, p. 23]. The structure of the brain allow us to respond to patterns in signals by using designated neural structures activated by the patterns of the inputs. These neural structures are reused specific patterns of neural information which reduces the need for a much larger number of neural structures to handle specific information. This is only possible due to the hierarchical nature of neural information in the brain.[36, p. 23][47]

A problem with ANNs in reproducing human phenomena is the need for large scale ANNs to fully grasp the information presented, creating a bottleneck in computing power and memory. By instead describing information as a geometry with regularities and symmetries (which is observable in the physical world), this information can be heavily encoded.

As an evolution of NEAT, Hypercube-based NEAT (HyperNEAT) builds on the methodology by creating a substrate, which in the case of an image, is a two-dimensional grid setup as a coordinate system. This substrate can have any amount of dimensions which is why it is called a hypercube. Just like in an ANN there is an input substrate and an output substrate, and it can have a number of hidden layers of substrates between them. Another ANN is then used to find the weight between every point in the input substrate, and every other point in the output substrate. This is where the coordinate system grid is used to define the position of every pixel, in relation to all other pixels. The cartesian[1] distance between the input coordinate (x1 and y1) and the output coordinate (x2 and y2) is used as inputs (four in total) for the ANN to find the weight of the connection between the pixels in cartesian space.

Since the ANN to find these weights have a configuration of neurons that can have different activation functions, this ANN acts as a signal generator driven by cartesian distance of pixels, which then produces compositions of patterns in the output substrate. This is why this ANN is better known as a Compositional Pattern Producing Network (CPPN).[36, p. 24]

HyperNEAT is then a GA used to find the best configuration for a CPPN that produces the best weights for the substrate. It is a generative model for a signal generator that can describe the relationship between any two coordinates as a function.

[1] A Cartesian coordinate system is a geometry based on the axes in a coordinate system

$$w = CPPN(x_1, y_1, x_2, y_2)$$

Figure 3.15: The positions of the pixels in the coordinate space, has a spatial relation. That relationship is calculated the CPPN on basis of the coordinate positions[36]

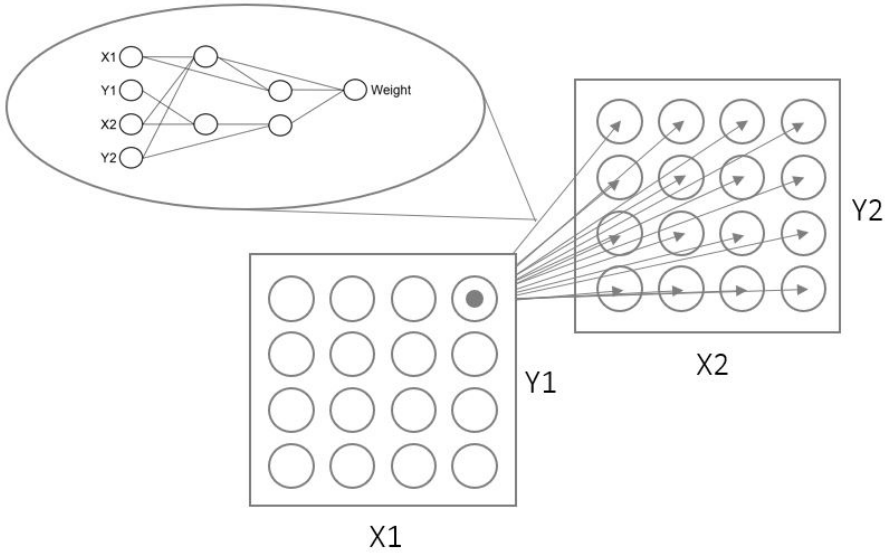


Figure 3.16: The CPPN calculates the relationship between the origin pixel(X1, Y1) and all of the destination pixels (X2, Y2), assigning all of them a weight.

Evolving a HyperNEAT configuration is different from training an ANN, because its job is not to find a structure in a dataset, but rather to learn how to approximate the structure of a dataset by comparing the behaviour of its encoding. In an ANN, the training dataset is used as input in the input layer, and then that data is propagated through the network and evaluated at the end. For HyperNEAT the input can be virtually anything within the frame of the substrate. The input used for the CPPN has impact on how the CPPN learns to deal with combining functions as all the pixels will be interrelated in the substrate. While the input does not directly relate to the output function of the CPPN, it has philosophical impact on how the CPPN will act in its use of the input to reach a good fitness. The output of a HyperNEAT network is the input multiplied by the sum of all the weights found by the CPPN, which can then be evaluated by comparing it with a fitness function. As the model has been evolved, and the most successful genome CPPN has been found, it can be used to create compositions. Since these compositions are the results of the CPPN that finds relationships between coordinates, the output can be up- or down-scaled in resolution infinitely by changing the size of the substrate. However, because we can consider the substrate an ANN where every point in the origin has a calculated weighted connection (by the CPPN) to all

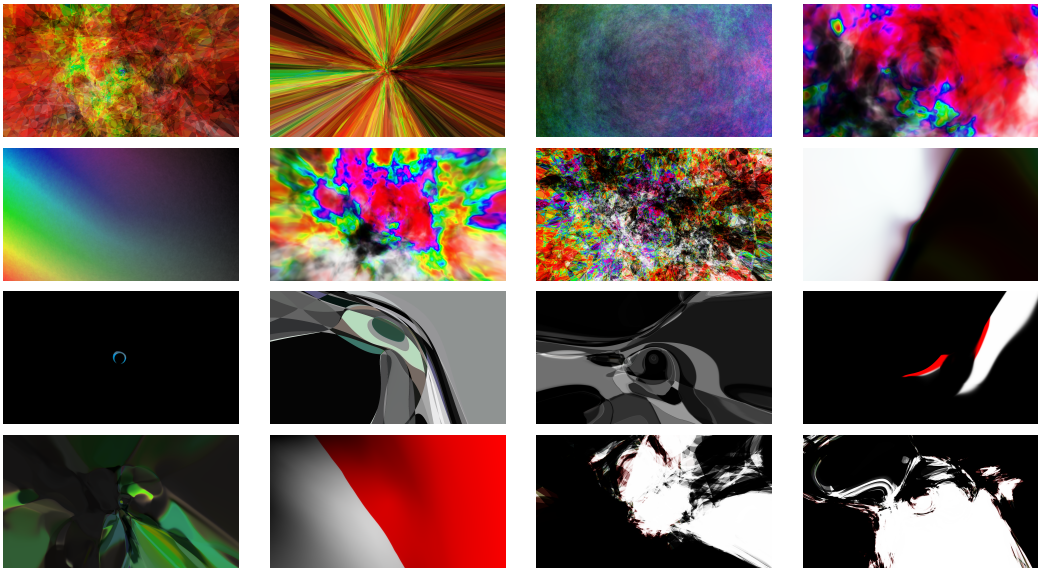


Figure 3.17: 4 x 4 grid of example images produced by a CPPN

destination points, the amount of computations for the resolution is squared. This does not need to be kept in memory which is essentially what allows the infinite scaling. The CPPN acts as a kernel for filtering information - an intelligent kernel that can scale infinitely with resolution and dimensions.

The CPPN is an encoding of a larger ANN substrate, that acts like an intelligent filter kernel. Phenomenologically speaking, the resulting images of this system is the equivalent of a "mind's eye" of our artificial brain. Example images from randomly initialised CPPNs can be seen in fig. 3.17.

Chapter 4

Implementation

Python[60] has been used as the main language for developing the software based on the neuroevolutionary approach inspired by synesthesia. Anaconda[2] was used as the Python environment and package organiser with Jupyter[3] as the development environment. Noteworthy libraries that made this project possible are: Librosa[29], MultiNEAT[35], Keras[52] and Tensorflow[4].

4.1 Imitating Human Hearing

Based on earlier work related to this project, an Autoencoder (AE) is used for feature extraction of music based on a Constant-Q Transformation (CQT)[57]. A CQT is an alternative to the Fast Fourier Transform (FFT) that instead of only dividing a signal into bands of frequency ranges, it also spaces those bands to closer mimic human hearing. The human hearing perception ranges from approximately 20 Hz to 20.000 Hz. That frequency range covers approximately 8 octaves since an octave of a tone is double the frequency. An "A" is 440 Hz, which means that A in the next octave is 880 Hz. As we perceive tonal information in relation to the original tone, a 44 Hz difference is perceived as bigger for tones in a lower octave than a higher octave as it makes up a bigger absolute difference (10% for 440 Hz and 5% for 880 Hz). This difference in relative frequency perception is why CQT is used to ensure that especially low frequency content does not get "squashed" by the amount of higher frequency content. This ensures that bass will be as impactful to the AE as it is to humans.

A training data set is built using the 15 second samples in a variety of musical textures that is used for the experiment videos, and consist of the following:

- Amen break
- Andre Aguardo . Through the Night

- Andy Stott - Execution
- Dauwd - Theory of Colours
- Duckmaw - Nicaraguan Dream
- Laurence Guy - Intro
- LEFTI - Diosa Del Amor
- Lorn - ARID
- Steve Gunn - Ancient Jules
- Noisia - Stonewalled - Hybris Remix

When the AE is trained with a complete set of CQT textures, it extracts features as it will save only the most important information to reproduce the signal with minimal loss. The features are an encoding of the information, and can be thought of as an auditory perception, filtering unnecessary information. The configuration of the AE is the same as used in previous work [57].

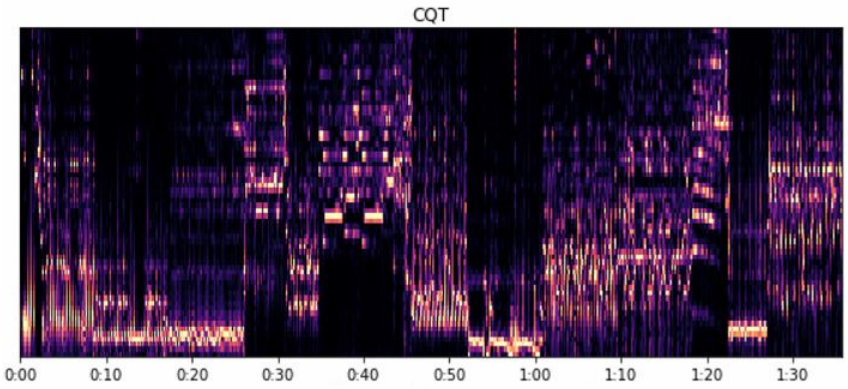


Figure 4.1: Spectrogram of the Constant-Q transformed music input.

A spectrogram of the CQT textures is shown in fig. 4.1, where it can be seen that there are intervals of signatures that are the different music tracks that combine to become the training dataset. As the feature extraction is used to build a 3 x 3 pixel image to use later with the CPPN, the latent space is set to a size of 9. As the CQT outputs 64 values, this means that the compression rate of the AE is 1:7.1. The model is trained for 500 epochs yielding a model that can compress the 64 inputs to 9 and recreate the 64 values with a loss of 0.0084. The output of the model can be seen in fig. 4.2 which compared to the input textures in fig. 4.1 shows a similarity of textures which means that the AE has learned to recreate the textures from the compressed representation.

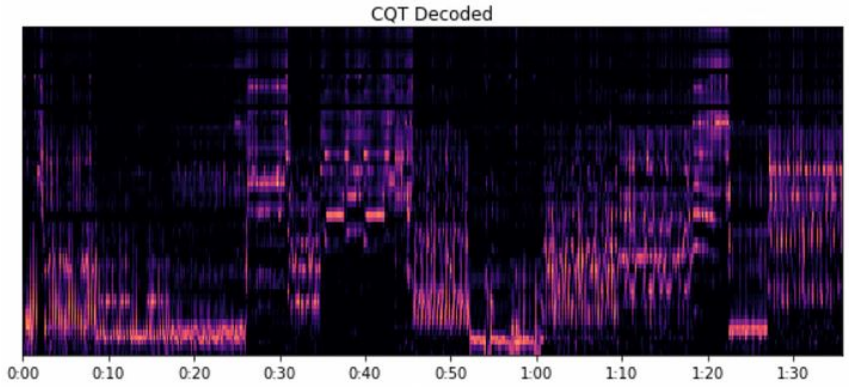


Figure 4.2: Spectrogram of the decoded output of the autoencoder after training.

4.1.1 Feature Extraction

The extracted features (the latent vectors) can be seen in fig. 4.3. These are the values that we use as the driver for the artificial vision to "neurally short-circuit" our artificial brain to imitate synesthesia. In phenomenological terms, this is the meaning the AE has derived from the music.

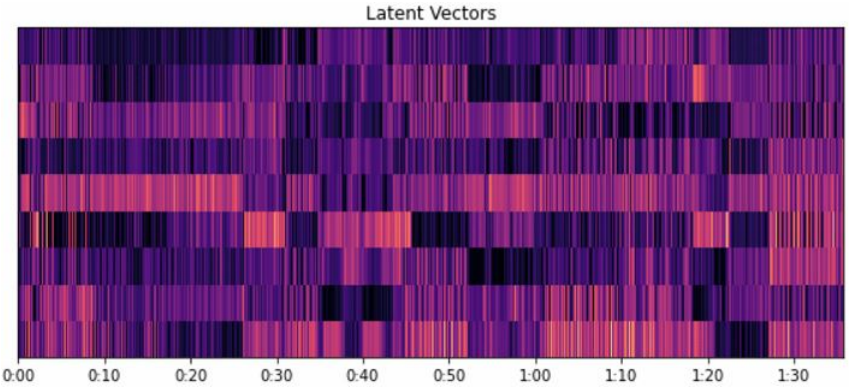


Figure 4.3: Spectrogram of the latent vectors found in the bottleneck layer of the autoencoder after training

Interpolation

The latent vectors are interpolated as it will make the movements more organic and calm. Interpolation is the creation of steps between two values. In this project it is used to limit how fast the values can change. It is done by finding the difference between the new values and the values from the last frame and dividing that by some number. In this project that number is set to 10 which means that every frame of the value can maximally move 10% of the distance every frame. Previous research found that control of interpolation speed can drastically change the

perceived expressiveness of musical mappings to light, and are very well suited as a meaningful parameter to adjust the movements of lights in real-time. [57]. An image of the interpolated latent space can be seen in fig. 4.4. For the experiment, the interpolation number is 10 which can be seen in the equation below.

$$InterpolatedValue = oldvalue + (newvalue - oldvalue)/10 \tag{4.1}$$

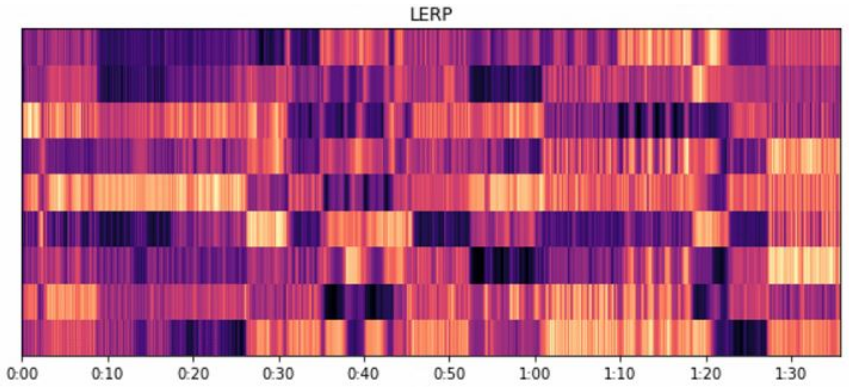


Figure 4.4: Spectrogram of the latent vectors with interpolation

4.1.2 Technical Performance Factors

The frame rate of the system has been optimised for 25 frames per second (FPS), because it is enough to fool the human vision perception of continuous movement. In service of development the FPS is limited because it creates optimisation problems for coupling the window size of the CQT and the sampling frequency of the audio. It also allows faster development iterations because the system has to produce an image every frame which makes the total number of images explode when making longer videos. Depending on the performance of the hardware computing the images, an image of 2560 x 1440 (2K) can take upwards of 2 minutes per image, which is why a lower frame rate also allows higher resolutions. For the purposes of developing a prototype, the FPS is set at 25. Fully optimising the prototype for higher frame rates are does not make sense as software migration away from Python is seen as a better prospect - Python is not efficient at signal processing.

There are some limitations to adjusting sample rate and window size for the CQT as we must obey some rules for sampling audio signals. Nyquist's Theorem states that to avoid aliasing, any signal analysis must be performed at double the frequency of the max frequency of the signal. Typically this frequency is 44100 Hz which is double the upper limit of human hearing of 22050 Hz. 44100 is therefore the lowest allowed sampling rate for the audio signals. The window size of the CQT defines the analysis window of the transformation. For the CQT functionality in the Librosa library, this number is a multiplicative of 2, meaning that the options

for window size are 512, 1024, 2048, 4096 etc. This number refers to the amount of samples to analyze per frame. This is fully in relation to the sample rate, which means that we can define a certain frame rate by increasing or decreasing the sample rate and window size. While choosing smaller window sizes may yield faster frame rates, it may also means that some low frequency information may be lost in the transformation, because the wavelength of the low frequency signals may be outside of the analysis window.

$$25FPS = 51200Hz/2048windowsize \tag{4.2}$$

4.2 Imitating the Retina

The human eye is the first filtering of light in the human body. As light is entering the eye through the pupil opening, through the lens and hits the retina, the light enters through the ganglion cells and the bipolar cells and finally - the rod cells, and the cone cells. The rods perceive luminance while the cones perceive colour [34]. As mentioned earlier, the cones are shaped to allow a natural fourier transformation of different wavelength of light [56]. The eyes' segmentation of visual information is best modeled by hue, saturation and value/brightness (HSB/HSV).

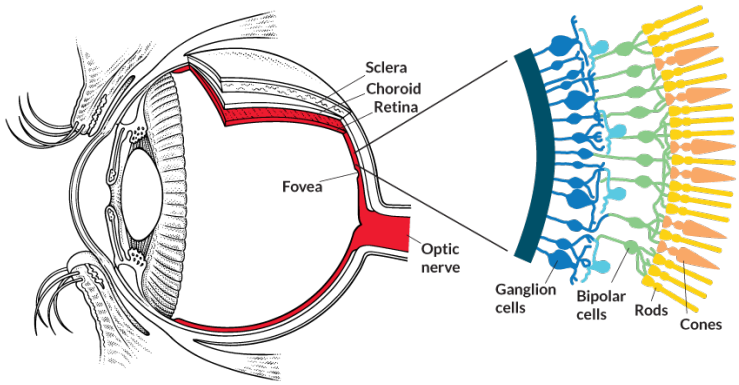


Figure 4.5: Illustration of the retina showing the placement of the rod and cone cells.[17]

As the retina consists of a segmentation between luminance in the function of the rods, and chromaticity in the function of the cones, in relation to HSV, the rods would function as a brightness channel, and the cones as a combination of the hue channel and saturation channel.[59]

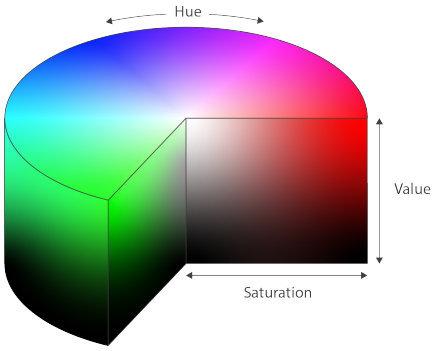


Figure 4.6: Hue, Saturation and Value cylinder [5]

4.3 Evolving the Visual Cortex

After the initial visual signal has been split into HSV, it proceeds to the artificial visual cortex that filters the signal in the search for features from which to derive meaning. As there are three channels, a CPPN is evolved for each of the channels emulating three filter kernels, filtering information in each it's respective channel.

The NEAT algorithm that builds the CPPN works by creating a population of a number of CPPNs and have each of them produce an image based on a white image input in the substrate. Phenomenologically this is the equivalent of being blinded by light as there is no filtering of visual information.

The output images are compared to one or more goal images by calculating the differences between the pixel values with root mean square error (RMSE) seen in fig. 4.3.

$$RMSE = \sqrt{\Sigma_{i=1}^{n} \left(\frac{output - goal}{n} \right)^2} \tag{4.3}$$

RMSE calculates the distance between the output of the CPPN and the goal. The closer RMSE is to 1, the more alike the images are (the distance is shorter). This is the fitness condition for evaluating the fitness of the genomes in the population. As they evolve they gain complexity, increasing the variety of images they can produce broadening the solution space. This also means that evolution of each generation becomes longer and longer as it takes more and more computations to create more and more complex solutions.

One example of a simpler evolutionary goal is to make it something without complicated shapes - for instance a picture of a sunset horizon as seen in Fig 4.7. The simple gradients of the sunset sky is easier for the CPPN to filter as it does not require a complex combination of functions to describe the relationship between pixels. The neuroevolutionary algorithm will evolve the genomes until they reach a goal threshold, or until it has evolved for X number of generations.

Figure 4.7: Gradients of orange and pink merge in the sunset sky

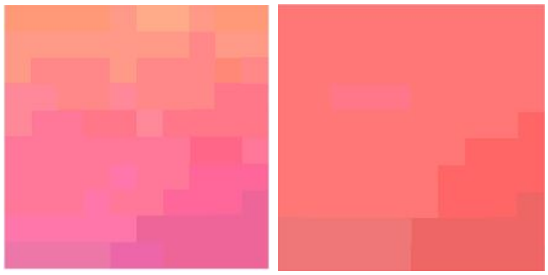


Figure 4.8: Left: Image of the down-sampled image used to compare the genomes against in the fitness function. Right: Image of the output after evolving all three channels of the image: hue, saturation and value.

For the neuroevolutionary approach in this project, the images are down-sampled to a manageable size for testing. An example is shown on the left in Fig 4.8 where the input image is down-sampled, split into HSV, and used as an evolutionary objective for the CPPN seen on the left in fig. 4.9, fig. 4.10 and fig. 4.11. The evolved CPPN is then able to produce the images seen on the right in fig. 4.9, fig. 4.10 and fig. 4.11 from an all white input. The fitness value for hue: 0.9573, saturation: 0.9795 and value: 0.9901. As these images are the results of three CPPNs - one for hue, saturation and value, the images can be merged back to a fully colored image, and can be seen on the right in fig. 4.8.

4.4 Merging the Senses

As we seek to merge the senses to let the AEs subjective experience of music stimulate the artificial visual sense, the latent space of the music must be integrated into the visual perception. Because of the evolutionary topology of the CPPN, it does

Figure 4.9: Left: Image of the input hue channel in gray scale. Right: Image of the output hue channel made by the CPPN in gray scale.

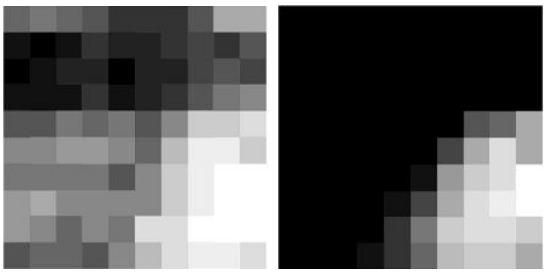


Figure 4.10: Left: Image of the input saturation channel in gray scale. Right: Image of the output saturation channel made by the CPPN in gray scale.

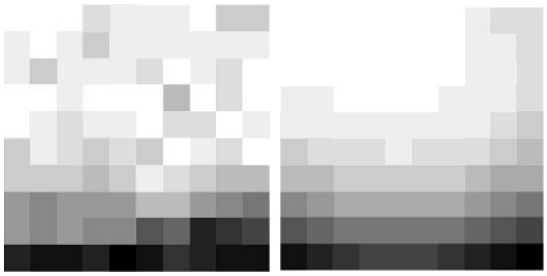


Figure 4.11: Left: Image of the input value channel in gray scale. Right: Image of the output value channel made by the CPPN in gray scale.

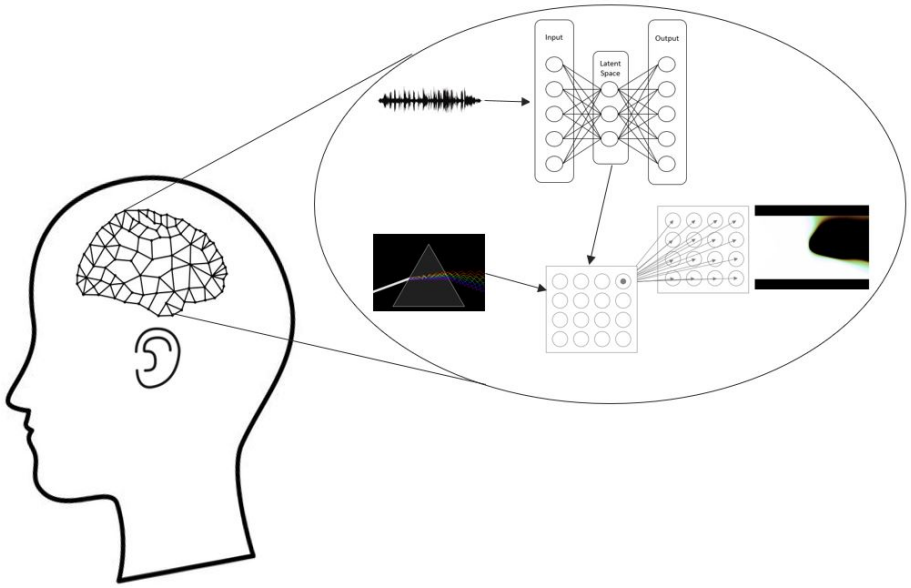


Figure 4.12: A model of the directional flow of information in the artificial synesthesia

not allow an integration of a static layer in the topology of the CPPN. We must therefore take a step backwards in the process to the input substrate. As the latent space is extracted from music, it is used to create a 3 x 3 pixel image consisting of 9 squares that each represent the latent vector values. These images are then used one by one to push the input substrate of the CPPN by resizing it to the same resolution and multiplying it with the image from the last image produced by the CPPN. This is a push to the feedback loop between output and input of the CPPN.

As the CPPN propagates the information from the input substrate, the squares disappear in the output as they have been filtered into the compositional potential of the CPPN. This manipulates the compositional range of the CPPN through

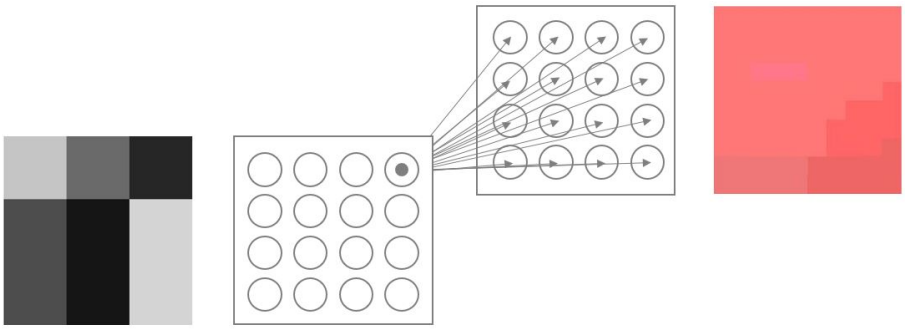


Figure 4.13: An image of the latent space vectors are used as input in the HyperNEAT substrate which creates a "sound driven composition" based on the evolved CPPN.

time synchronized with music, effectively using a subjective sound perception to stimulate the visual sense to create an artificial synesthesia.

4.5 Testing the Technology

The initial testing of evolving a CPPN is time consuming and computationally very heavy, especially for larger images. What it doesn't require in memory storage, it requires in computations. For an image of 100 x 100 pixels, since the CPPN describes the connection between two pixel points in an image, it has to run the CPPN 100.000.000 times to calculate all of the connections:

$$CPPNActivations = (X * Y)^2 \tag{4.4}$$

In the example of a 2k image (2560 x 1440), this number would explode to 13.589.544.960.000 connections. It can be imagined how this procedure could be time consuming when remembering that every generation of evolution in the NEAT algorithm has a population of genomes that each should produce an output that can be evaluated for fitness. While this is the inherent advantage as it allows very simplistic hardware to evolve very sophisticated solutions, it is also the limitation to how sophisticated we can allow the CPPN to be evolved in a shorter time span of a few months. Realistically the CPPN can be evolved to reproduce gradients and simple shapes in low resolutions of 25 x 25 pixels. After the CPPN has been evolved, it can be used to generate high resolution images. However, the MultiNEAT library in use for evolving the CPPN has a built-in function to create images that utilizes the CPPN in a manner that saves all of the connections in memory, meaning that it is unable to produce images of higher resolution higher than 100 x 100 pixels with 16 GB of RAM. This limits the potential of the CPPN encoding, and is contradictory to the entire point of developing a CPPN. This is an internal issue in the library that has not been sought to be solved as it would require an amount of software development and contact with the author of the code.

To show the the CPPN method, we can manually craft a CPPN structure that does not have any training, but is able to produce high resolution images that shows the compositional potential of a CPPN.

4.6 Manual CPPN Construction

An implementation of a CPPN using the Keras library is borrowed from Dennis Kerzig[62] and edited to fit our needs. As the ANN models built using Keras consists of layers instead of singular neurons seen in fig. 4.14, we lose some of the flexibility in arbitrary connection patterns that NEAT is able to. But what it

allows us to do is to make very big architectures and integrate the musical features directly into the CPPN, thereby using the sound to manipulate the compositions directly as seen in fig. 4.15. In this latent space layer, each neuron will be multiplied by one of the latent space values. This can show some of the potential of the CPPN encoding without relying on months of neuroevolutionary processes.

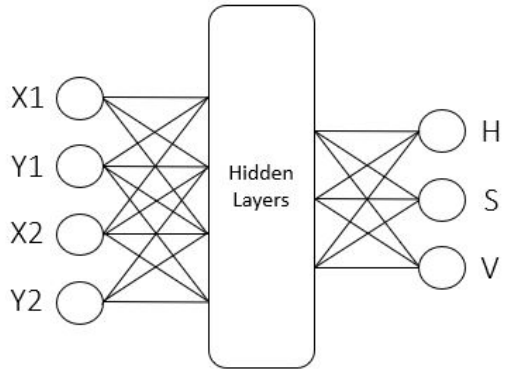


Figure 4.14: Model of a CPPN in a Keras implementation with an amount of hidden layers with an amount of neurons in each. This is predetermined by the user, while the activation functions for each layer is randomly drawn between ReLU, Sigmoid and Tanh.

As the CPPN model is built, we set a few variables to decide the architecture: the amount of hidden layers, the amount of neurons in the layers and the variance in the weight initialization. Every time the model is built, it will be unique as all the weights in the CPPN are randomly drawn which when combined with the random activation functions of each layer, yields a completely unique expression of the CPPN.

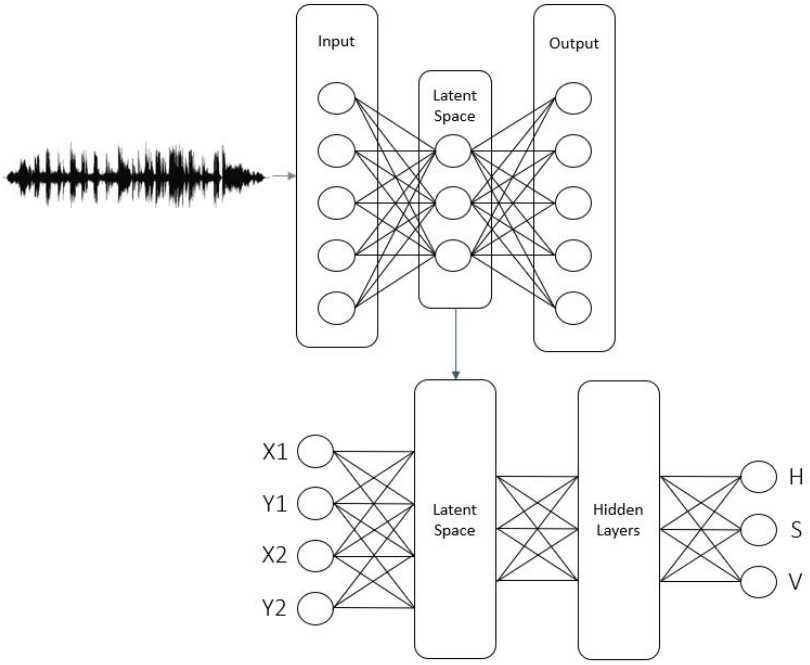


Figure 4.15: Model of a CPPN in a Keras implementation with the latent space from the audio AE implemented in the second layer.

Chapter 5

Test

An online survey was used as the basis for conducting an experiment with videos produced by the artificial synesthesia. The developed system was compared against two baseline visualisations to gather information on two points.

1. Do the movements of the compositions made by the system represent the music? This is evaluated by comparing scores with a video of the developed system, with visuals driven by Perlin noise, instead of music. This is to ensure that the system differentiates itself from the tendency in humans to see a connectedness, and meaningfulness in unrelated things. This is known as *apophenia*.

2. Does the visualisation aesthetics have potential to generate enhancing audiovisual experiences? This is evaluated by comparing scores with a video of a spectrogram driven by the music. The system should be able to differentiate itself from a generic visualisation style in a variety of compositions.

5.1 Preparations

5.1.1 Cross Modal Interaction

To evaluate the cross modal experience of synesthesia, the Likert scale definition was inspired by the classification of cross modal interactions made by Biocca & Choi, 2001 [15]. Their work on taxonomy of cross modal interactions inspired the assigning of the ends of the Likert scale to a "degree of integrated cross modality". The higher the score is, the more coherent the experience seems as the audio and visual experiences merge and enhance the effect of each other. A lower score means a perception of more separation between the perceived modalities, in other words, they do not feel connected or coherent.

Figure 5.1: An example of the perlin noise used as input for the CPPN.

5.1.2 Perlin Noise

As a comparative measure to verify the musical connection between visuals, an identical CPPN was used to create the same visuals, but based on Perlin noise as input instead of musical features.

Perlin noise is an award winning algorithm developed in the 1980's by Ken Perlin. While it is random since it is in nature a noise algorithm, it is found many places in nature and has been utilized proficiently for natural looking computer generated graphics. The simplest implementation of Perlin noise is a noise generator that creates random values based on the number(s) that has come just before it. This means that there is a time component involved that directs the noise to be random in relation to the value before it. This gives the animations a pleasant movement, as it has randomness in waves of directions, rather than jagged "unnatural" movement. The image in fig. 5.1 shows an example of the input values used as input for the CPPN to generate the movements in the image composition. An example frame of the implementation of Perlin noise with the CPPN can be seen in fig. 5.2.

5.1.3 Spectrogram

To evaluate the visual connection between different types of music, a spectrogram is used as a visualiser for the music as seen in fig. 5.2. The spectrogram is a typical visual representation of the frequency bands present in the music found by doing an FFT which is the most common feature form of sound. It is used as a baseline for this experiment for something to measure the "merging of audio and visual" against.

5.2 Experiment Setup

To evaluate whether our artificial synesthesia can produce cross modal experiences, the potential must be tested which is why the Keras implementation of the CPPN is used.

5.2.1 Participants

The only prerequisites for participating in the experiment is to have normally functioning vision and hearing. The Google Form was shared online directed at university students and adults. The total number of participants was 30.

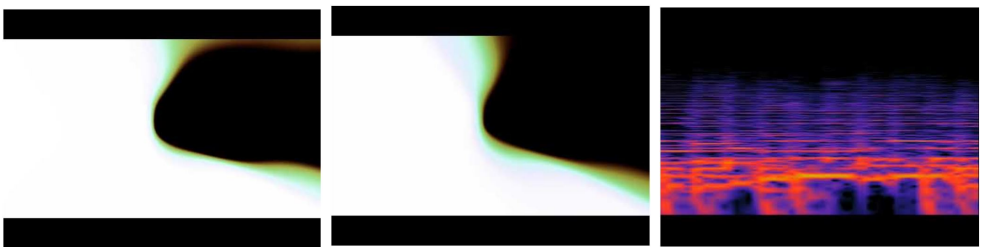


Figure 5.2: A still image from what the participants saw in the experiment. Left, the artificial synesthesia where the music is driving the CPPN. Middle: The CPPN being driven by perlin noise. Right: an FFT based spectrogram driven by music.

5.2.2 Online Survey

To test the cross modal potential of the system, it is compared to other visualisations in an experiment with 30 participants. 15 second videos are produced with the Keras implementation of the CPPN which is stimulated by the audio features from the artificial auditory sense. It is compared to videos of the same CPPN model, but uses perlin noise as input instead of audio features. The third comparative video is of a traditional audio visualisation scheme of an FFT based spectrogram. The videos presents a new type of CPPN visualisation every round to test if our artificial synesthesia system shows immediate connection to the music across a variety of different compositions.

The experiment was executed in 10 rounds and every round 3 videos are presented to the participant. The same music was used for all videos in each round. 1 video was of the artificial synesthesia developed in this project, 1 video of a CPPN visualisation based on perlin noise and lastly 1 video of an FFT spectrogram. The videos were randomised throughout the experiment. The participants were introduced to the Likert scale (seen in fig. 5.3) with this explanation:

- In this experiment, you will evaluate the merging of audio and visual in an audiovisual experience. This will be evaluated on a scale from 1 to 9.
- In the lowest end of the scale (1), the auditory experience and the visual experience are perceived separately as they don't seem to have any interaction or is even blocking the experience of each other. This is called substitution.
- In the middle of the scale (5), the auditory experience and the visual experience are perceived as unified as they both represent each other. This is called mapping.
- In the highest end of the scale (9), the auditory experience and the visual experience are perceived as not only unified, but also enhancing as one or both of the experiences enhance the experience of the other or both. This is called enhancement.

Figure 5.3: The Likert scale question from the questionnaire, which had the participants evaluate how they experience the sensory modalities of each video.

The participants were urged to use a PC/Laptop and headphones for participation as it grants better fidelity. Unfortunately Google Forms do not support full screen of videos, and therefore the participants were guided to view the videos in full screen by pushing a YouTube link which allows them to view it in full screen. Beneath every video, there is an open comment box for any comment the participant may have. At the end of the experiment, the participants were asked whether or not they viewed the videos in full screen and if they used headphones, along with an optional open comment box.

Chapter 6

Results

In the experiment survey with 30 participants, the results of the 3 types of videos were the following:

- The artificial synesthesia got a mean grade of 5,3 with a standard deviation of 2,1
- The Perlin noise based visualisation got a mean grade of 3,8 and a standard deviation of 1,9
- The spectrogram visualisation got a mean grade of 4,3 and a standard deviation of 1,9

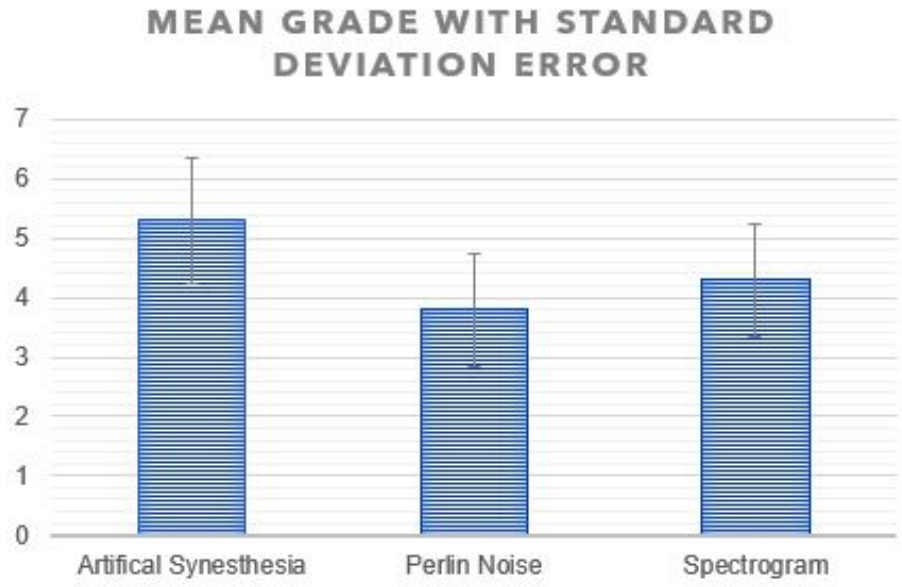


Figure 6.1: Graph of the mean grades with standard deviation error bars. The standard deviations are similar while the mean grades shows the Perlin noise as the least "merged" visualisation, and the Artificial Synesthesia as the most "merged" visualisation.

6.1 Wilcoxon

The Wilcoxon Signed Rank Test was run twice as the goal was to compare the artificial synesthesia to both the Perlin noise based visualisation and the spectrogram visualisation.

6.1.1 Comparing Artificial Synesthesia with Perlin noise

- Null hypothesis: artificial synesthesia is not perceived as more enhancing than the Perlin noise based visualisation.

The results of the Wilcoxon Signed Rank Test yielded a sum of 10. The critical value for alpha 0,005 with 30 participants is 98, which is why we can firmly reject the null hypothesis. The artificial synesthesia is perceived as more enhancing than the visualisations made by Perlin noise.

6.1.2 Comparing Artificial Synesthesia with Spectrogram

- Null hypothesis: artificial synesthesia is not perceived as more enhancing than the spectrogram.

The results of the Wilcoxon Signed Rank Test yielded a sum of 73. The critical value for 30 participants at alpha 0,005 is 98, which is why we can reject the null

hypothesis. The artificial synesthesia is perceived as more enhancing than the visualisations made by the spectrogram.

Chapter 7

Analysis

As the Wilcoxon tests rejects both of the null hypotheses, we can say that the artificial synesthesia performs a better cross modal integration of sound and light than the perlin noise based visualisations and the spectrogram. It was speculated that the grades of the three visualisations would align themselves after perlin noise being in the lowest end of the scale (substitution), the spectrogram in the middle of the scale (mapping), and the artificial synesthesia in the highest end of the scale (enhancement). The alignment seen in fig. 6.1, shows the confirmation of these speculations, and confirms that the combination of the non-linear musical mapping from the AE, combined with the movement of the compositional images can empower the cross modal experience. As the image composition was randomised between every round, there was no correlation between the music being played and the aesthetic of the visualisation. In spite of that fact, the artificial synesthesia is still able to communicate a musical merging with the visuals within 15 seconds. The compositional layout would change every round, and therefore it would present the participant to a new version of the merging of music and visual. These findings gives reason to believe that the artificial synesthesia has potential to enable designers to explore dynamic designs that is contextualised with music.

Chapter 8

Discussion

8.1 The Experiment

In the design of the survey experiment, there were some discussion about some variable parameters for testing the potential of the system.

8.1.1 Choice of Videos

The length of the video samples were discussed as rendering longer videos would take much more time, which meant that it would inadvertently mean a reduction in image resolution. The length of the videos meant that the participants would have enough time to intuitively understand the music in the visuals, but not enough time to "figure out" the complexities of behaviour powering them. This is a worst case for the system because there are structures of the visualisations of the musical features that stretch over much longer than 15 seconds. Because of the way the musical features are extracted, the individual features has a relationship to the complexity of the entire training data set, meaning that it derives features in relation to the structure that it fits within. This will show a variance in visualisation movements during a longer stretch of visualisation. Additionally the musical feature extraction is non-linear, which makes for a more complex mapping that is interrelated to the context of what the AE has learned. In other words, an increase in bass will not necessarily yield an increase of one of the latent space values. The representation of the bass will be put into a context of the AEs musical understanding.

There is a practical element to choosing a shorter sample as it allows a multitude of visualisation to be tested for quantitative evidence of visualisation power of the system, across a multitude of pairings of musical and visual textures. If the videos would be too long, it would risk the participants getting impatient with the survey and giving disingenuous answers. Further experimentation of the system

should be more directed at producing aesthetically coherent textures with longer musical samples.

8.1.2 Defining the Scale

The Likert scale used to evaluate the cross modal experience of the videos was defined by a polar opposition between "Substitution" and "Enhancement". Although the scale was thoroughly defined in the survey, it is speculated that people's experience with likert scales could draw them towards grading based on their personal preferences of the visuals rather than what was defined in the survey. One participant even admitted to this in the comments, which would suggest that another type of evaluation would be safer in the future. However, the occurrence of the alignment of the mean grades suggests that the majority of participants showed a good grasp of the concepts presented when defining the likert scale.

8.1.3 The Aesthetics

The videos made for the experiment were not aesthetically paired with music, which surely carries great potential for enhancing the experience. The potential disparity between musical and visual aesthetics has likely caused some of the variety of the grades which increased the standard deviation. As discussed earlier, the scale was not meant to be a way for the participant to grade their preferences, however, a successful matching of visual and musical style will be perceived as a more enhancing experience, and it is ultimately a subjective matter of taste. This shows that the worst case scenario for the system was employed which solidifies the potential of the system.

8.1.4 Iterations of Design

The results from the experiment shows that the system is a viable foundation for continuing development on AI based lighting design. It can be considered an end-user test as a feedback loop in the Design Thinking methodology.

8.2 Taking Control of AI

As a generative algorithm, the sky is the limit for the HyperNEAT algorithm in terms of possibilities of compositions, and different combinations of configurations yielding wildly different results. One of the challenges with these systems is figuring out how to gain meaningful control over the system. Control in this context meaning that the outcome can be expected based on some manipulation to the system. Because the systems almost take on their own behaviour and act in unexpected ways in development, some control is needed. The randomization

of the Keras implementation of the CPPN would often yield black images, or images with red being a very dominant color. Black images happens when all of the channel values goes to 0 which means that the randomised network gradient is cause all the neurons to die. The probability of this happening is greater with deeper CPPNs as there is a higher probability of hitting an "unlucky" combination of activation functions and weights that will descend the neurons into a vanishing gradient meaning that the neurons will descend towards a value of zero. Red being a dominant color also makes a lot of sense in this case as red occupies both 0 and 1 in the Hue channel. This means that red will be a dominant color if any exploding gradient or vanishing gradient tendencies happen. The Keras implementation of the CPPN can not be trained with back propagation and gradient descent, because the results that it needs to calculate the gradient descent is the result of the behaviour that it exerts - not the values it produces. This means that a GA implementation is needed to train/evolve it as it is a much better suited search heuristic for controlling the CPPNs behaviour. And while it is possible to integrate the Keras CPPN into a GA heuristic, it is still limited in its configuration options which makes it less suited for neuroevolution.

GA methodologies' challenge is how to narrow the solution space to something that is within a realistic time frame. Humans' brains have evolved from single cell organisms spanning 3.5 billion years. While we are only trying to replicate simple versions of singular modules of the human brain, the HyperNEAT approach to make CPPNs is very time consuming because of the rather frugal mutation and breeding scheme. The NEAT implementation of CPPN has great potential but requires significantly better evolution heuristics, to explore the solution space more efficiently. The fitness function used may also not be the best for the purpose as it may be much more interesting to look for structural similarities in the images instead of comparing pixel values. There are multiple other image distance comparison functions that should be tested for more efficacious neuroevolution of the CPPN (Hamming, Hausdorff). While comparing pixel values may yield technically more similar images, a lot of subtleties may be lost in the last few percentages of dissimilarity. Said in another way, it may be better to compare images with a perception-based comparison function that looks for structures in an image, instead of comparing pixel values. Basing it on gestalt theory would maybe empower the subtleties in the images that have a bigger impact (than pixel values) on human perception.

Another option is to implement novelty search as the search heuristic for "good behaviour". Instead of evaluating a genome's success by a fitness function, its success can be evaluated by how different its behaviour is compared to the population. Recent research shows that evaluating success by setting a goal and using it to "breed good behaviour", is not as efficient as simply breeding for novel behaviour in the population and keep track of the behaviour that comes closest to

the goal. Novelty search is not only practically interesting, but also potentially massively revolutionizing how we think about gaining knowledge and innovation. As the traditional way of evaluating genome success is by measuring its fitness (its objective), novelty search will instead measure novel behaviour as to allow a much more aggressive exploration. Just as human innovation doesn't come from knowing the destination, so too doesn't a genome in a GA population[27]. A very interesting presentation on this topic by Kenneth Stanley can be seen at [25].

Future implementations of this research would include novelty search and graphic processing unit (GPU) integration. There is a library available for integrating CPPN in GLSL[1][63] which would allow real-time rendering of high resolution substrates. The structure of HyperNEAT and CPPN lends itself very well to GPU operations, and as GPU markets have been saturated with GPUs for ML purposes, the software architectures for this kind of use is widespread. Google Colab offers free GPU acceleration for ML projects[6]. GPU integration and software migration to GLSL would be a powerful addition as the type of processing done when applying the CPPN to a substrate is perfect for parallel processing and multithreading. This means that with a trained/evolved HyperNEAT model, it can be visualised in real-time - potentially at very high resolutions. As GPU technology continues to progress, this can scale with new GPU technologies and performance.

Because the HyperNEAT algorithm is inhabiting a hyperspace that can be of any dimensional layout. This also means that it can do more than map pixels in images. It can also map voxels in 3D space for instance, meaning that the CPPN can be evolved to describe functional relationships of lighting in a space. In other words, the system can potentially be evolved to learn characteristics of 3D lighting design.

The AE used in this project has been used to extract features from music. The feature extraction part of the system is open-ended meaning that any data could be analysed for feature extraction in this system. One could imagine using daylight features to generate dynamic lighting scenes evolved by pictures of sunset skies. The core of the system is the pairing of time and space. The system is a merging of a temporospatial nature as it uses features in time as a driver of spatial compositions.

8.2.1 Morphology of Compositional Movements

In the extraction of features in the artificial auditory sense that we have built, there is a level of abstraction that has been indirectly chosen as we chose to use a 64-9-64 AE to compress auditory features. The size of the latent space was chosen to be 9 because it would fit neatly into a 3 x 3 grid, which was important as the latent space vectors were saved as images used by the CPPN. However, there is a discussion to

[1]GLSL is a shader language based on WebGL used for graphical work.

be had about the feature complexity of the choice of latent space size. Supposedly, lowering the latent space size would increase the loss of the AE meaning that it would try to find features that are easier for it to generalize with over the entire dataset. What happens if the latent space size is 1? In other words, if the AE could pick just one feature to describe the dataset, what would it be?

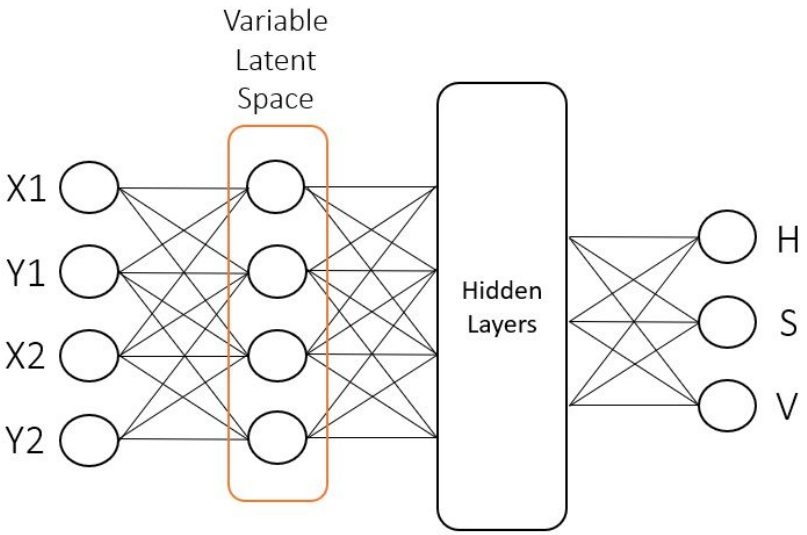


Figure 8.1: By varying the latent space size, the compositional morphology will change as the impact of every neuron in the latent space will be greater with less neurons. The bigger the latent space is, each neuron will have less impact, but will produce more complex animations.

The complexity of the feature extraction is also propagated through to the CPPN as the integration of the auditory latent space pushes the compositional combinations. There may be some interesting interactions between the complexity of the feature combinations and the compositional integration of them which may change the dynamics severely changing the expressive behaviour of music and light. Along with the interpolation type and speed of the latent space, this is subject to testing for interesting interaction design.

Another factor of the system is the neuron types used in the artificial auditory sense. As we are dealing with time series, a continuity sensory information has relations to what has come just before it. However, the neuron configurations used in this project does not take temporal continuity into consideration which is also discussed in previous work [57]. This may drastically change the connection between movements in music and movements in light. Additionally, the depth of the feature extracting AE is believed to have a significant impact on the abstraction of the features.

8.3 Interaction Parameters

If this system is to be made into a design tool, there are some considerations specifically around how many and which interaction parameters to give the user. Ultimately it would hinge on a focused analysis of the end user, and the specific purpose of the software. However, there are some inferences that we can make in that regard. From the previous research using the AE to produce musically driven visuals [57], we found that controlling the interpolation speed of the musical features from the AE has tremendous impact on the perceived expressiveness of the visualisations. This is most likely the case in this integration as well because it represents the velocity at which we present temporal information visually to the audience. The velocity at which the visuals change has a great impact on the perceived "tone" and should be matched by the user. It is comparable to using perlin noise instead of random noise.

Depending on the scope of further development of artificial synesthesia, there are some parameters that look very interesting in terms of generating empowering, meaningful interactions. As was just mentioned above, the amount of features extracted from the music change greatly in abstraction as there are more or less features, and the animation complexity that it yields in the CPPN will change drastically. Integration of a functionality for dynamically changing the latent space size of the AE and the latent space integration layer in the CPPN would be an interesting contender for shaping meaningful interactions with the system.

8.4 Cybernetic Design

Cybernetic theory has great potential for framing the future work of this project. A stronger methodology for describing systems, and systems communication can alleviate some decision making processes in re-framing the perspective of why things are designed a certain way. Cybernetic design and interaction design should also be integrated in future works specifically for aiding in the definitions of objectives and how to control the feedback systems of an adaptive technology like neuroevolution.

8.5 The Future

There are many possible improvements to the system, which may include layering of this system with other intelligent signal processing systems. The integration of a cybernetic point of view on the further development would be highly valuable in the scope of making design decisions for making a meaningful lighting design tool for dynamic lighting design.

8.5.1 Reference Projects for Development

Some interesting architectures has been developed for CPPN implementation in a visual encoding scheme. The image reconstructive abilities of the CPPN has been improved by Tesfaldet et. al. 2019 by letting the CPPN output Fourier coefficients instead of RGB, which yielded higher reconstructive fidelity[53].

As mentioned earlier in the discussion, the data structure and processing in the CPPN is very well suited for graphical processing as shown by Snelgrove et. al. 2018 by integrating a CPPN structure in OpenGL. This allows the use of GPU to power the pixel calculations in GLSL which can be done in real time at very big resolutions, depending on the complexity of the CPPN [46]. 0

Chapter 9

Design with Artificial Synesthesia

To show a practical implementation of the dynamic compositions made by the developed system, a virtual architectural environment was designed in Unreal Engine[55].

9.1 The Virtual Design

The compositions produced by the artificial synesthesia can be used in delineations of vertical and horizontal lighting (Søndergaard et. al., 2015) that makes lighting volumes to create lighting atmospheres[58]. Mapping the compositions to architectural spaces to control the movements of light in luminaires can communicate dynamic temporal structures[50].

The pixel experiment, as seen in fig. 9.1, is the foundation for a virtual implementations using a dynamic lighting system. In combination with the system developed in this project, it can create interesting movements in the perception of imagery through pixel mapping to virtual pixels in space.

Inspiration was also found in the work on harmonic visualisation by Jeong & Kim, 2019 where the luminaires are suspended from the ceiling and are lowered and raised to form waves according to the arousal of the music seen in fig. 9.2[23].

32 point lights in diffuse white glass spheres are suspended from the ceiling in a perlin noise wave formation in a 4 x 8 grid. The perlin noise wave formation changes the perceived luminaire size and formation as the observer moves around the space. The colors of the luminaires are controlled by reading images produced by the artificial synesthesia. The images are split into its HSV components and sent to Unreal Engine with OSC[1]. The HSV values are received in Unreal and mapped to the luminaires depending on their spatial position. Examples can be seen in fig. 9.3.

[1] *Open Sound Control* is a standardised protocol for sending messages between software

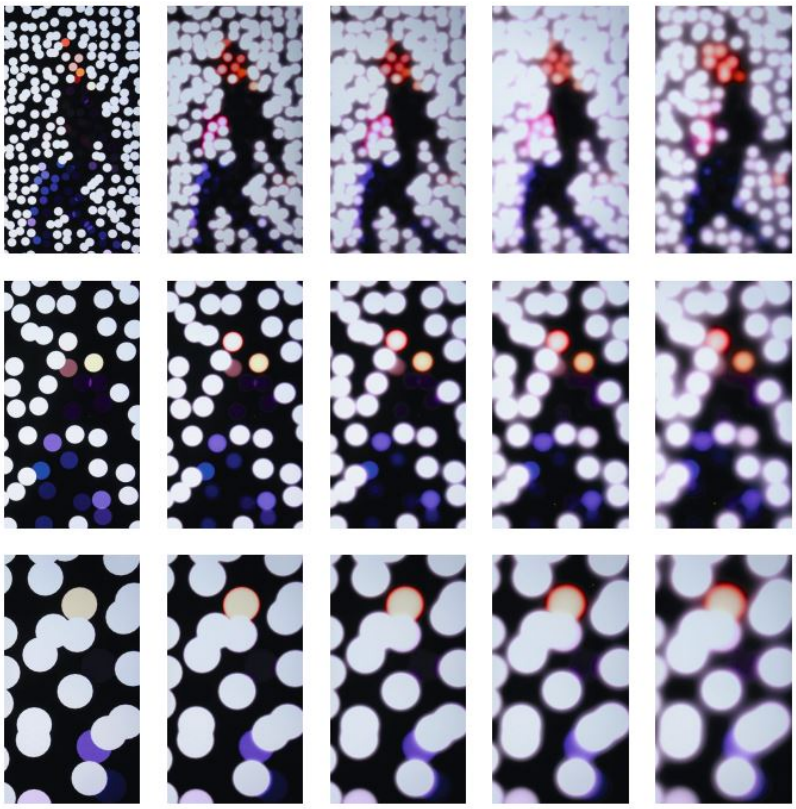


Figure 9.1: Comparison images from the pixel experiment as made by Søndergaard et. al. 2015. The grid of images compare different layouts of pixel placements, with different pixel sizes. [50]

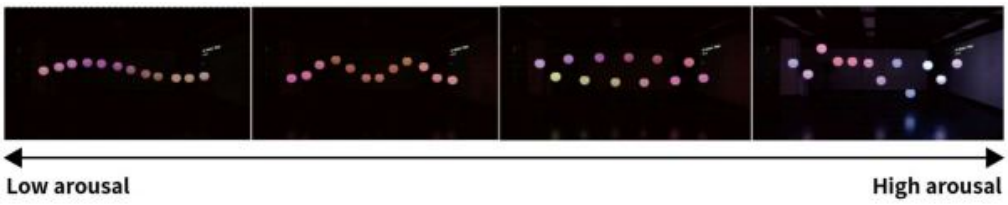


Figure 9.2: The artwork installation by Jeong & Kim changes the luminaire heights to form waves of arousal.[23]

A central theme to all of the implemented theory of light and sound is *waves*. A wave is relational to its surroundings, and signifies the natural dynamic structures of the universe. It is the basis of all of the perceptional stimuli that we process, and is fundamental to the way we perceive the world. Just like Perlin noise, while random and chaotic in nature, the relational nature of things makes structure out of chaos. It is *natural*.

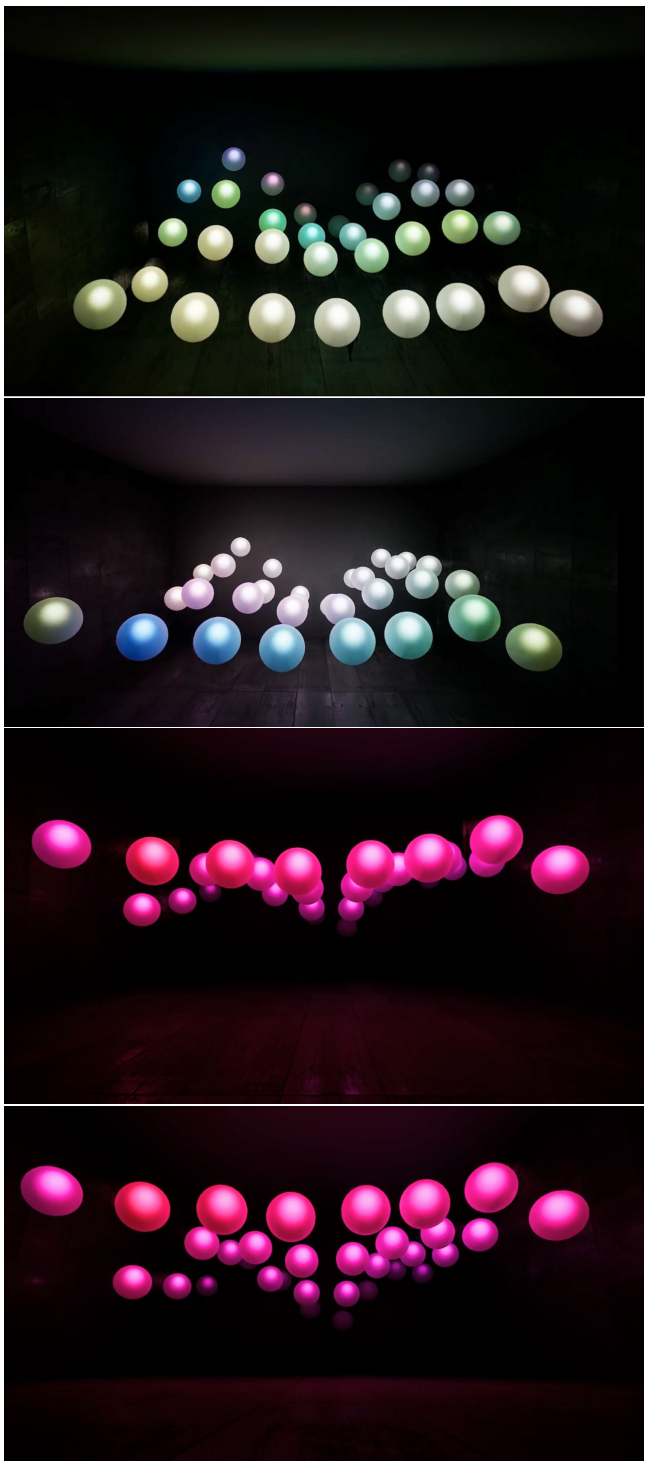


Figure 9.3: Images from the virtual installation made in Unreal Engine showing the wave formations of light representing the features of the music.

Chapter 10

Conclusion

A system has been built by replicating the neurophysiological phenomenon synesthesia. The system consists of artificial neural networks that functions as artificial sensory perceptions (vision and hearing), which are then neurally short circuited to create artificial music-vision synesthesia. It is a one-way short circuit where a stimulus to the auditory sense creates a response in the visual sense. An evaluation of the system shows that it can communicate musical features through movement of the compositions that it produces, and invites further development.

The developed system is a tool for generative dynamic lighting design that allows exploration of compositions driven on intelligent music recognition. It is a merging of auditory and visual gestalt that creates movements in time and space. It is an augmentation of lighting design, that explores lighting structures in time. Tapio Rosenius, in his talk about light for ambient communication[30] talks about the focal hierarchy as described by Lou Michel: people, movement, brightness, high contrast, vivid color, strong patterns, meaning and combinations of the former. From here Tapio emphasizes *movement* and *meaning* as strong communicators in lighting design as movement is part of nature, and with a meaning, it can move people as well as light. As Lou Michel also notes, the combinations of the focal accents is something that demands our visual attention. These concepts can be generated through the artificial synesthesia in the merging of meanings and movements of light and sound. The system is built on an AI structure that finds patterns in time to create movements in light. Movements through time can be found in all of nature - the colors of the horizon, the blooming of flowers, the swells of the tides, any of which can be used with the system. It is a generative design platform.

The project shows a cybernetic approach to synesthesia that has the spirit of co-creation between designer and AI in mind. The machine's ability to look for new solutions empowers the designer to explore, and create dynamic designs. There are many interesting parameters to limit the solution space for the AI - it could be limited to a colour palette, contrast threshold, brightness threshold among many

others. Anything that can narrow the solution space for the evolutionary AI to give coherent meaningful answers.

Synesthesia is an enormously complex topic because its boundaries are ill-defined. Or rather, it is difficult to define its boundaries because it is based on statistic anomaly of subjective experiences. In some sense, we are all synesthetes - we are all born with synesthetic abilities. Intuitively, we can understand synesthetic feelings if we think about smell and taste - two very interconnected sensory pathways. Synesthesia is fascinating because it is the unusual pairings of sensory reactions that really shows us the differences between subjective experiences. It is a classic philosophical wondering - how can I be sure that what I understand about the color red, is the same as you? Or rather, how do I *know* that it does not *look* green to you, but because you have learned to recognise it as red, you have assigned the same metaphorical meaning to it. That synesthesia exists highlights how our brain makes sense of the world, and has inspired a lot of research into cross modal integration of information in neuroscience.

What is really in common between all of the concepts throughout this project is language - communication of the metaphors we use to describe ideas and concepts that has universal *meaning*. The meanings are what sits in the middle of all our sensory experiences, what gives our sensory information context. Humans have evolved to understand contexts and meanings, because it is how we interact with the world and each other. Meaning is how our brain decides what is important and what is not. In taking inspiration from synesthesia, the system developed is seeking to enable the exploration of meanings in time and space.

The true potential of the system is realised by defining boundaries for it. With stronger cybernetic contexts to define the solution space for the evolutionary AI, the potential of the system can be explored.

Acknowledgements

A very sincere thank you to George Palamas for excellent supervision, and thank you to Stefania Serafin for guiding the cross modal integration of the project with her expertise. Thank you to the Lighting Design Studio for the access to the media PC that aided in rendering and training AI models for the project.

Bibliography

[1] URL: https://en.opisanie-kartin.com/description-of-the-painting-by-wassily-kandinsky-contrast-sounds/.

[2] URL: https://www.anaconda.com/.

[3] URL: https://jupyter.org/.

[4] URL: https://www.tensorflow.org/.

[5] URL: https://wumbo.net/article/color/.

[6] URL: https://colab.research.google.com/.

[7] 2016. URL: https://www.youtube.com/watch?v=2k8fHR9jKVM.

[8] abhigoku10. *Topic DL01: Activation functions and its Types in Artifical Neural network*. 2018. URL: https://medium.com/@abhigoku10/activation-functions-and-its-types-in-artifical-neural-network-14511f3080a8.

[9] P. Alperson. "Musical Time" and Music as an "Art of Time."". In: *Journal of Aesthetics & Art Criticism* (1980), p. 407.

[10] A. Amanatiadis, V. G. Kaburlasos, and E. B. Kosmatopoulos. "Understanding Deep Convolutional Networks through Gestalt Theory". In: *2018 IEEE International Conference on Imaging Systems and Techniques (IST)* (2018). DOI: 10.1109/ist.2018.8577159.

[11] Roy Ascott. *Behaviourist Art and the Cybernetic Vision*. W.W. Norton & Company, 1966.

[12] Michael Bach. *Motion-Bounce Illusion*. URL: https://michaelbach.de/ot/aud-bounce/index.html.

[13] *Beau Lotto - Understanding Perception: How We Experience the Meaning We Create*. URL: https://futureofstorytelling.org/video/beau-lotto-understanding-perception-how-we-experience-the-meaning-we-create.

[14] Phil Bernstein et al. *Generative Design in Architecture and Construction Heralds Productivity*. 2020. URL: https://redshift.autodesk.com/generative-design-architecture/.

[15] Frank Biocca, Jin Kim, and Yung Choi. "Visual Touch in Virtual Environments: An Exploratory Study of Presence, Multimodal Interfaces, and Cross-Modal Sensory Illusions". In: *Presence: Teleoperators and Virtual Environments* 10.3 (2001), 247–265. DOI: 10.1162/105474601300343595.

[16] Margaret A. Boden and Ernest A. Edmonds. "What is generative art?" In: *Digital Creativity* 20.1-2 (2009), 21–46. DOI: 10.1080/14626260902867915.

[17] By. *Home*. 2019. URL: https://www.readingmybooks.com/2019/10/14/how-to-rewire-the-eye/.

[18] *Complete Guide of Activation Functions*. 2019. URL: https://mc.ai/complete-guide-of-activation-functions/.

[19] *Generative Processes in Architectural Design*. 2016. URL: https://kadk.dk/en/case/generative-processes-architectural-design.

[20] Paul Halpern. *It From Bit: Is The Universe A Cellular Automaton?* 2017. URL: https://medium.com/starts-with-a-bang/it-from-bit-is-the-universe-a-cellular-automaton-4a5b1426ba6d.

[21] Trevor Huff. *Neuroanatomy, Visual Cortex*. 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK482504/#:~:text=Thevisualcortexisthe,posteriorregionofthebrain.

[22] *Human*. URL: https://deepdreamgenerator.com/.

[23] Won-ung Jeong and Se-Hwa Kim. "Synesthesia Visualization of Music Waveform: 'Kinetic Lighting for Music Visualization'". In: *International Journal of Asia Digital Art & Design* (2019). URL: https://www.jstage.jst.go.jp/article/adada/23/2/23_22/_pdf.

[24] *Journey on the Deep Dream*. 2015. URL: https://www.youtube.com/watch?v=SCE-QeDfXtA.

[25] *Kenneth Stanley: Why Greatness Cannot Be Planned: The Myth of the Objective*. 2015. URL: https://www.youtube.com/watch?v=dXQPL9GooyI.

[26] A. Kohlrausch and S van de Par. "IS&T/SPIE Conference on Human Vision 34 and Electronic Imaging IV". In: *Auditory-Visual Interaction: From Fundamental Research in Cognitive Psychology to (possible) Applications* (1999), 34–44.

[27] Joel Lehman and Kenneth O. Stanley. "Novelty Search and the Problem with Objectives". In: *Genetic and Evolutionary Computation Genetic Programming Theory and Practice IX* (2011), 37–56. DOI: 10.1007/978-1-4614-1770-5_3.

[28] Daniel Levitin. *This is Your Brain on Music*. Penguin Books Ltd., 2019.

[29] *librosa*. URL: https://librosa.org/doc/latest/index.html.

[30] *Light for Ambient Communication | Parsons School of Design*. 2018. URL: https://www.youtube.com/watch?v=rjp6ne7EaU0.

[31] David P. Luke and Devin B. Terhune. "The induction of synaesthesia with chemical agents: a systematic review". In: *Frontiers in Psychology* 4 (2013). DOI: 10.3389/fpsyg.2013.00753.

[32] Vijini Mallawaarachchi. *Introduction to Genetic Algorithms - Including Example Code*. 2020. URL: https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3#:~:text=Ageneticalgorithmisa,offspringofthenextgeneration..

[33] Harry Mcgurk and John Macdonald. "Hearing lips and seeing voices". In: *Nature* 264.5588 (1976), 746–748. DOI: 10.1038/264746a0.

[34] Lou Michel. *Light: the shape of space: designing with space and light*. Wiley, 1996.

[35] *MultiNEAT*. URL: http://multineat.com/docs.html.

[36] Iaroslav Omelianenko. *Hands-on neuroevolution with Python: build high performing artificial neural network architectures using neuroevolution-based algorithms*. Packt Publishing Ltd., 2019.

[37] Paul Pangaro. *Cybernetics*. URL: https://www.pangaro.com/definition-cybernetics.html.

[38] Gordon Pask. *An approach to cybernetics*. Harper & Bros., 1961.

[39] V S Ramachandran and E M Hubbard. "Synaesthesia - A Window Into Perception, Thought and Language". In: *Journal of Consciousness Studies* 8.12 (2001), 3–34. URL: http://cbc.ucsd.edu/pdf/Synaesthesia-JCS.pdf.

[40] *Ratatouille - Synesthesia - HD - FX Animation by Michel Gagné*. 2015. URL: https://www.youtube.com/watch?v=rLXYILcRoPQ.

[41] *Recall the number from the colour*. 2013. URL: https://goneuro.wordpress.com/2013/11/28/recall-the-number-from-the-colour/.

[42] *Ryoichi Kurokawa - ad/ab Atom*. 2020. URL: https://www.youtube.com/watch?v=bvHiobhVfnw.

[43] *Sensology by Michel Gagné - HD 720P - Music by Paul Plimley and Barry Guy*. Michel Gagné, 2010. URL: https://www.youtube.com/watch?v=UVWKtXDvr04.

[44] Julia Simner and Edward M. Hubbard. *The Oxford handbook of synesthesia*. Oxford University Press, 2019.

[45] Christopher Sinke et al. "Genuine and drug-induced synesthesia: A comparison". In: *Consciousness and Cognition* 21.3 (2012), 1419–1434. DOI: 10.1016/j.concog.2012.03.009.

[46] Xavier Snelgrove and Matthew Tesfaldet. "Interactive CPPNs in GLSL". In: *32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.* (2018). URL: https://nips2018creativity.github.io/doc/interactive_cppns_in_glsl.pdf.

[47] Kenneth O. Stanley, David B. Dambrosio, and Jason Gauci. "A Hypercube-Based Encoding for Evolving Large-Scale Neural Networks". In: *Artificial Life* 15.2 (2009), 185–212. DOI: 10.1162/artl.2009.15.2.15202.

[48] *SYNESTHESIA - Animation by Michel Gagné / Music by Gheorghe Costinescu*. 2014. URL: https://www.youtube.com/watch?v=sao3NAapOAI.

[49] *Synesthetic Sensory Stimulation with Ryoichi Kurokawa*. 2013. URL: https://www.youtube.com/watch?v=_XAK248_apY\&t.

[50] Karin Søndergaard, Kjell Yngve Petersen, and Christina Augustesen. *Pixel Experiments*. 1970. URL: https://www.forskningsdatabasen.dk/en/catalog/2393817290.

[51] Tate. *'Chance, Order, Change 6 (Black)', Kenneth Martin, 1978–9*. URL: https://www.tate.org.uk/art/artworks/martin-chance-order-change-6-black-t03190.

[52] Keras Team. *Simple. Flexible. Powerful.* URL: https://keras.io/.

[53] Mattie Tesfaldet, Xavier Snelgrove, and David Vazquez. "Fourier-CPPNs for Image Synthesis". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019). DOI: 10.1109/iccvw.2019.00392.

[54] *The Future of Vision and Eye Care*. 2019. URL: https://medicalfuturist.com/future-of-vision-and-eye-care/.

[55] *The most powerful real-time 3D creation platform*. URL: https://www.unrealengine.com/.

[56] Rohit Thummalapalli. *Fourier Transform: Nature's Way of Analyzing Data*. 2010. URL: https://www.yalescientific.org/2010/12/fourier-transform-natures-way-of-analyzing-data/#:~:text=âĂœToformanimageon,wavesformillionsofyear

[57] Simon Borst Tyroll, Daniel Overholt, and George Palamas. "Proceedings of the 17th Sound and Music Computing Conference". In: *AVAI: a Tool for Expressive Music Visualization based on Autoencoders and Constant Q Transformation* (2020), 378–385.

[58] Philip Ursprung. *Herzog & De Meuron: Natural history*. Lars Muller, 2002.

[59] A. Vadivel, Shamik Sural, and Arun K. Majumdar. "Human color perception in the HSV space and its application in histogram generation for image retrieval". In: *Color Imaging X: Processing, Hardcopy, and Applications* (2005). DOI: 10.1117/12.586823.

[60] *Welcome to Python.org*. URL: https://www.python.org/.

[61] Norbert Wiener. "Cybernetics or Control and Communication in the Animal and the Machine". In: (1948). DOI: 10.7551/mitpress/11810.001.0001.

[62] Wottpal. *wottpal/cppn-keras*. URL: https://github.com/wottpal/cppn-keras.

[63] Wxs. *wxs/cppn-to-glsl*. URL: https://github.com/wxs/cppn-to-glsl.