Auris: Creating Affective Virtual Spaces from Music

Misha Sra
MIT Media Lab
sra@media.mit.edu

Prashanth Vijayaraghavan
MIT Media Lab
pralav@media.mit.edu

Pattie Maes
MIT Media Lab
pattie@media.mit.edu

Deb Roy
MIT Media Lab
dkroy@media.mit.edu

ABSTRACT

Affective virtual spaces are of interest in many virtual reality applications such as education, wellbeing, rehabilitation, and entertainment. In this paper we present Auris, a system that attempts to generate affective virtual environments from music. We use music as input because it inherently encodes emotions that listeners readily recognize and respond to. Creating virtual environments is a time consuming and labor-intensive task involving various skills like design, 3D modeling, texturing, animation, and coding. Auris helps make this easier by automating the virtual world generation task using mood and content extracted from song audio and lyrics data respectively. Our user study results indicate virtual spaces created by Auris successfully convey the mood of the songs used to create them and achieve high presence scores with the potential to provide novel experiences of listening to music.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Computing methodologies** → *Computer vision*;

KEYWORDS

Virtual Reality, Music, Deep Neural Networks, Generative Models

ACM Reference format:
Misha Sra, Prashanth Vijayaraghavan, Pattie Maes, and Deb Roy. 2017. Auris: Creating Affective Virtual Spaces from Music. In *Proceedings of VRST '17, Gothenburg, Sweden, November 8–10, 2017,* 11 pages.
DOI: 10.1145/3139131.3139139

1 INTRODUCTION

Light, color, texture, geometry and other architectural design elements have been shown to produce predictable and measurable effects on our minds, brains, and bodies [Ellard et al. 2015]. This suggests spaces that can mirror or transform feelings or serve specific purposes like improving learning or enhancing wellbeing can be designed. With Auris, we take a first step towards the design of such spaces in virtual reality (VR) by attempting to automatically generate virtual environments (VEs) that can affect our emotions.

Few scholars dispute the claim that listeners recognize and respond to emotions in music as indicated by findings from studies using behavioral, physiological, and neurological measures [Gagnon and Peretz 2003; Krumhansl 1997; Mitterschiffthaler et al. 2007]. Studies measuring physiological effects of music have shown changes in listeners' heart rate, skin temperature, skin conductance, and breathing. Listening to music has been shown to activate brain areas previously known to be associated with emotional responses. Expressive behavior such as laughing, crying, smiling as indicated by observations and electromyographic (EMG) measures of facial muscles are further evidence of emotive response to music [Juslin 2011]. Since our goal is to design emotive virtual spaces and music universally evokes emotional responses, we chose to use music as one of the inputs into our design pipeline.

A variety of consumer VR device setups are now available offering different features and capabilities. However, developing VR applications remains a difficult and time-consuming task that requires specialist skills. Building 3D models of objects is the first step of creating a VR world. To complete the design, one also needs to add user interactions, specify materials for all objects in the world, and consider lighting and other environmental elements. While the models and textures define the shape and look of a virtual space, lights and color define the mood of the 3D environment. Putting material and lights together to create a world that is aesthetically pleasing, regardless of realism, is thus a difficult task and requires knowledge and an aesthetic eye. One way to facilitate VE creation is to do it automatically from 3D scans of real world spaces [Sra et al. 2016].

In this paper we present Auris, a novel system to automatically generate VR worlds from music. The input to our system is a song (audio and lyrics) and the output is a VR world that encapsulates the mood and content of the song in the design of the space, the objects added, and the textures applied. Additional information about objects and lighting comes from an online study where we asked participants about their associations between places, objects, lighting and moods. While the virtual world is generated and textured automatically, interactive elements and lighting are currently added manually as needed. Using creative license we decided to transform the generated virtual landscapes into psychedelic and surreal places by pre-processing textures that are applied to objects through a DeepDream like neural network.

A user experiences the generated VE immersively through an HTC Vive head-mounted display (HMD). We demonstrate the output of Auris with two generated worlds that correspond to the two broad mood classes of happy and sad Table 1). The happy world is generated from the song The Bird and the Worm by Owl City and

VRST '17, Gothenburg, Sweden
© 2017 ACM. 978-1-4503-5548-3/17/11…$15.00
DOI: 10.1145/3139131.3139139

the sad world from the song Blue Prelude by Nina Simone. Specific contributions of our work include:

- The concept of using music data (audio and lyrics) to generate 3D virtual worlds.
- The implementation of the end-to-end pipeline to demonstrate the concept of a data centric approach to automatic generation of emotive VEs.
- Encoding of high level features like mood in the image generation process.

We envision musicians and listeners creating and sharing immersive musical experiences and in the future perhaps even allowing for their individual personalities to be incorporated into the creation process through physiological sensors.

2 RELATED WORK

In this section we review prior literature on the affective dimensions of architecture and music, and also explore data-driven content creation methodologies along with image generation techniques.

2.1 Music, Space & Emotion

Our motivation for using music to create affective virtual spaces comes from the fact that music can both convey emotion as well as influence listeners' emotions. There have been several studies that indicate that listeners respond affectively to music [Krumhansl 1997; Witvliet and Vrana 2007]. In the literature, different models of emotion have been used that lend themselves to different measurement techniques. Categorical models make use of distinct labels (e.g., happy, sad, etc.), whereas dimensional models are consistent with the use of rating scales (often for arousal and valence) [Hunter and Schellenberg 2010]. Background music is often used in film or video games to induce mood and emotional context [Cohen 1999]. Paiva et al. describe the architecture for an agent that can generate appropriate background music by matching the current mood of a virtual environment [Casella and Paiva 2001]. Chen et al. classify nature images into eight mood classes and extract emotion from music audio. They match the classified images with music segments based on similar emotions to present the listener with a 2D music visualization [Chen et al. 2008]. In Auris we attempt to embed moods extracted from audio into the spatial and visual design of a 3D virtual experience.

With the rise of affective computing and increased collaborations between architects and neuroscientists, we are beginning to see the design of real world spaces that evoke specific emotions or put occupants in predetermined moods by measuring and predicting the psychological effects of the built environment [Ellard et al. 2015]. The idea of using landscape design to promote the reintegration of nature and the healing process [Marcus and Barnes 1999] has been around for centuries. However, VEs have yet to be examined in the emotional and psychological dimensions [Naz et al. 2017]. We believe architecture is an important connection between the real and the virtual worlds. Auris takes a first step towards building VEs that encapsulate emotions in the spatial and visual design and in turn elicit emotional responses from users.

2.2 Data-driven Content Creation

With the availability of 3D models and sensors for generating 3D point clouds, data-driven content creation has attracted much interest in recent years. Different techniques have been proposed for purposes like scene arrangement [Fisher et al. 2012], modeling [Chen et al. 2014], generation [Sra et al. 2016], and interactive synthesis of virtual worlds [Emilien et al. 2015]. Aesthetiscope is an artwork whose grid of colors are dynamically generated from a poem or song, to illustrate the power and potential of going beyond literal understandings of text [Liu and Maes 2005]. Many successful procedural methods have been proposed for different domains (terrains, forests, cities, etc.), some have been combined to synthesize complex virtual worlds [Smelik et al. 2011] with automatic material suggestions for 3D scenes [Chen et al. 2015]. Our work shares the same spirit of the above methods and uses song audio and lyrics to generate 3D virtual worlds.

2.3 Mood-Based Image Generation

Even though there is a lot of work overall related to generative models for image generation [Denton et al. 2015; Goodfellow et al. 2014; Mansimov et al. 2015], there is not much on image generation using moods or emotions. Most of the work associated with images and moods has been done with discriminative models [Ng and Jordan 2002]. In EmoNets [Kahou et al. 2016], the authors use multimodal deep learning techniques to predict moods in videos. For sentiment analysis in the same videos, they use progressively trained and domain transferred Deep Networks by You et al. [You et al. 2015]. In Auris, we build a generative model using mood and text labels for the image generation task. The mood is extracted from song audio data and text labels come from song lyrics. Our generative model is a variant of the PixelCNN Decoders [van den Oord et al. 2016] model and explores conditional image generation with a new image density model.

3 AURIS OVERVIEW

We use music and data from users in an online study as input for generating virtual spaces. To demonstrate our pipeline, two VR worlds are generated corresponding to the two broad mood classes of happy and sad which are among the most frequently felt musical emotions according to survey studies [Juslin and Laukka 2004; Sloboda 1992; Zentner et al. 2008]. The Auris pipeline (Fig. 1) consists of the following steps.

1. Mood identified by feeding Mel-frequency Cepstral Coefficients (MFCC) features [Müller 2007] of the song audio through a gated recurrent neural network (RNN) [Chung et al. 2014].
2. Noun-phrases are extracted from song lyrics using the Stanford part-of-speech tagger. [Manning et al. 2014].
3. Images are generated by providing mood + noun-phrase pairs obtained in steps 1) and 2) to our Mood-Conditional PixelCNN (MC-PixelCNN) model trained on our image dataset.

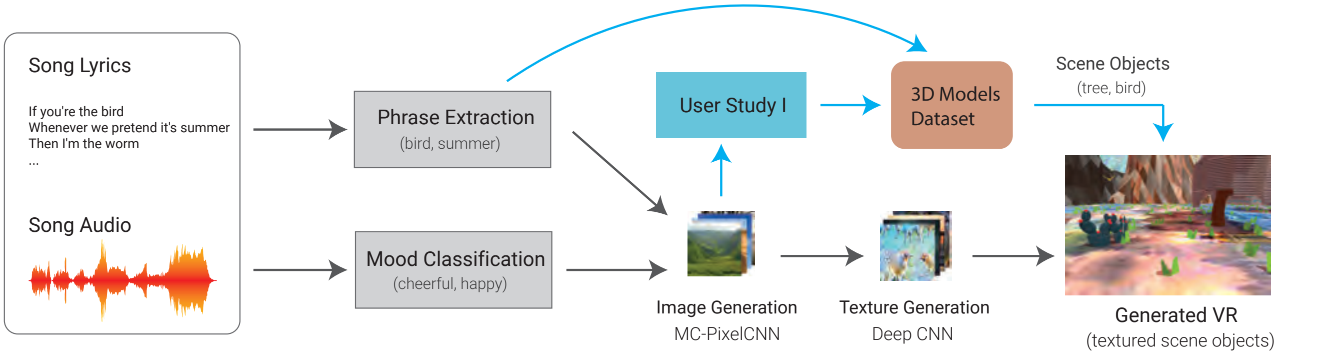

Figure 1: Illustration of the Auris pipeline to generate a VR world from one song. The input to the system is a song (audio and lyrics). Noun-phrases are extracted from the lyrics and mood from the audio. The mood and noun-phrases e.g., 'happy + bird' are used to generate images using a trained MC-PixelCNN model. The generated images are fed through a DeepDream based convolutional neural network to create surreal textures which are applied to objects in an automatically created VR world. The noun-phrases are used to select objects from a tagged 3D model dataset to add to the scene. Additional information about which objects to add to the scene comes from data collected in User Study I.

(4) Textures are created by feeding the output from step 3) into a DeepDream[1] like deep convolutional neural network (CNN) codenamed Inception [Szegedy et al. 2015].
(5) A genetic algorithm is used to procedurally create a VR world where the objects are textured with the output from step 4). Objects added to the VE are chosen from a database of tagged 3D models. Search terms used are the noun-phrases extracted from song lyrics in step 2) and feedback from participants in User Study I.

To summarize, we extract the mood of a song from its audio and the contents of a song from its lyrics. Using the mood and the content, we generate 2D images with MC-PixelCNN, transform the images into textures with DeepDream, and apply the textures to objects in a 3D scene. The 3D scene is made up of objects from the song lyrics and objects that users in our online study suggested should be part of places associated with specific moods.

3.1 Song Audio

3.1.1 Mood Dataset. The first step of our approach consists of accessing song audio data needed for mood prediction and for building our image dataset (Section 3.3.1). We use the Million Song Dataset (MSD) [Bertin-Mahieux et al. 2011] that comes as a collection of meta-data such as song names, artists and albums, together with MFCC features , loudness, and tempo extracted with the The Echo Nest API [2]. Automatic emotion analysis from song audio is a complex and challenging task, and beyond the scope of this work. Instead, we use metadata tags added by millions of listeners through a community voting process on LastFM[3], to build our song-mood dataset as described in [Van Zaanen and Kanters 2010]. To select songs to be included in our dataset, we require each song to be tagged at least twice with at least one mood tag to avoid retrieving songs with no tags or with tags that are not synonyms of tags in our list. This results in a subset of songs from the MSD with an associated mood tag. While using LastFM tags simplifies our task, we acknowledge that the tags may not accurately represent emotions in the songs, considering individual differences in expressing and describing emotions. For mood classification, similar to prior work [Lu et al. 2006; Yang and Lee 2004; Yang et al. 2008], we adopt Russell and Thayer's arousal-valence emotion plane [Russell 2003; Thayer 1990] as our taxonomy and define four mood classes happy, angry, sad, and calm, according to the four quadrants of the emotion plane, as shown in Figure 2. These four mood classes (see Table 1) are also among the most frequently felt emotions by music listeners [Juslin and Laukka 2004; Sloboda 1992; Zentner et al. 2008].

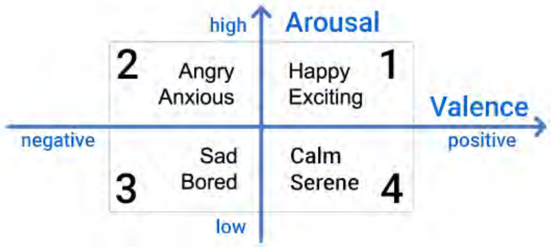

Figure 2: Russell and Thayer's arousal-valence emotion plane. We define four emotion classes according to the four quadrants of the emotion plane.

3.1.2 Mood Classification. We need to predict the mood of any given song in order to encode the mood information for generating images for that song later in the pipeline. We extract MFCC features for each music segment from the MSD. These features are fed into an RNN to predict the broad mood category of a song (Table 1).

[1] https://en.wikipedia.org/wiki/DeepDream
[2] http://the.echonest.com/
[3] http://www.last.fm/

We use a gated recurrent network (GRU) [Chung et al. 2014] as it is computationally less expensive than Long Short Term Memory (LSTM) networks and performs better than a standard RNN [Cho et al. 2014; Chung et al. 2014]. At each time step t, the GRU unit takes a row of the MFCC feature segment x_t and a hidden state h_t as input. The internal transition operations of the GRU are defined as $h_t = GRU(x_t, h_{t-1})$ The final hidden state (h_T) is fed to a fully connected layer followed by a softmax layer. The output of the softmax layer is a distribution over our mood classes. To learn the parameters of the network, we minimize the categorical cross entropy loss using Adam Optimization [Kingma and Ba 2014]. After the classifier is trained using the mood dataset, we can input a new song and get its mood as output.

Emotion recognition from music is a difficult task because emotions are subjective and a universal expression for emotion is not feasible because descriptors for the same emotion can vary between individuals. Defining computational models of emotions is an active area of research. MFCC, spectral shape, harmonic change or chromagram are common features used for mood classification. Most features are, however, better at predicting arousal than valence, which is more challenging to predict [Kim et al. 2010].

Table 1: Mood classes according to the quadrants of the Russell and Thayer's arousal-valence emotion plane (Fig. 2).

Mood Category	Mood Tags
Angry	aggression, aggressive
	angst, anxiety, anxious
	anger, angry, choleric, fury
Happy	upbeat, gleeful, enthusiastic
	cheerful, festive, jolly
	happy, happiness, happy music
Sad	depressed, blue, dark, gloom
	sad, sadness, unhappy
	grief, heartbreak, sorrow
Calm	meditative, contemplative, quiet
	comfort, serene, peaceful
	calm, soothe, mild

3.2 Song Lyrics

3.2.1 Lyrics Dataset. Using the subset of songs that have been mapped to mood classes, we use the MSD to extract track information like song name, artist name, trackID etc. This data is used to perform an automated song search on Flash Lyrics[4] where we look for an exact match of artist and song name. Non-english song lyrics are filtered out. After this search we get a total of ∼75, 000 songs that are mapped to a mood category and have full song lyrics.

3.2.2 Noun Phrase Extraction. We need to extract noun-phrases from song lyrics to use with moods for building our image dataset. Additionally, we use the extracted noun-phrases to select matching 3D models to be included in the VE generated from that song. During pre-processing, we remove stop words (e.g., the, is, at, which, and on), other infrequent words, and tokenize the sentences. A tokenizer splits text into a sequence of tokens, which roughly correspond to "words". The Stanford POS tagger [Manning et al. 2014] is employed to read the lyrics and assign parts of speech tags to each token such as noun, verb, and adjective. We perform Noun-Phrase or NP-Chunking on the tagged results of the POS. For example the sequence, "the yellow dog" is tokenized and tagged by the POS as "the (DT for determiner), yellow (JJ for adjective), and dog (NN for noun)." The final list of words after de-duplication forms our candidate descriptive labels. We use a threshold t (where t refers to the number of songs in which a particular word or phrase occurred in the entire lyrics dataset and was empirically established as $t = 10$) to remove inherent noise in the dataset and noise generated from the chunking technique. The final list of noun-phrases is used, along with mood data and the image dataset, to train our MC-PixelCNN generative model.

3.3 Images

3.3.1 Image Dataset. We chose to build our own image dataset as pre-existing datasets for visual mood analysis like the IAPS [Machajdik and Hanbury 2010; Zhao et al. 2014] only encode mood and we wanted the images to encode both mood and content from lyrics. Combining the extracted phrases with broad mood classes (for example, 'happy-blue-sky' or 'sad-ocean'), we collect 25 images per mood-phrase pair using Google Image Search to build our dataset. Using the top 5000 words or phrases, we retrieve images for two mood tags (happy, sad). Search terms that do not provide a minimum of 100 results per mood-phrase pair are removed. Thus, a dataset containing ∼250, 000 images is created where the images represent song lyrics and mood combinations. Manual curation is performed randomly on the collected images to verify the overall quality of the collected images before using the data to train the MC-PixelCNN. The dataset includes photos of nature, people, objects, animals, indoor scenes with occasional clip art. The dataset is built once to train the MC-PixelCNN model. Once the model is trained, our system can take mood-phrase text pairs extracted from any song as input and generate new images that encode that mood-phrase information.

Moods are often associated with colors, lighting, and places. There is a positive relation between certain colors and moods. For example, red is more often associated with exciting-stimulating, black with powerful-strong-masterful or blue with tender-soothing [Wexner 1954]. Similarly, there exist systematic influences on mood from lighting encountered in everyday interior conditions [McCloughan et al. 1999]. However, studies of the impact of full-spectrum lighting on mood have given controversial results [Küller et al. 2006]. Our system attempts to create an association between mood and a virtual place through the generation of textures from music mood and content data.

3.3.2 Image Generation. The PixelCNN model [van den Oord et al. 2016] focuses on description or tags for conditional image generation. Our MC-PixelCNN architecture is a new and modified version that models the conditional distribution of natural images given two latent vector representations of text labels from song lyrics data and mood extracted from song audio. In the original PixelCNN implementation [Oord et al. 2016], every pixel depends on all the pixels above and to the left of it. Hence, the joint distribution

[4] https://www.flashlyrics.com/lyrics

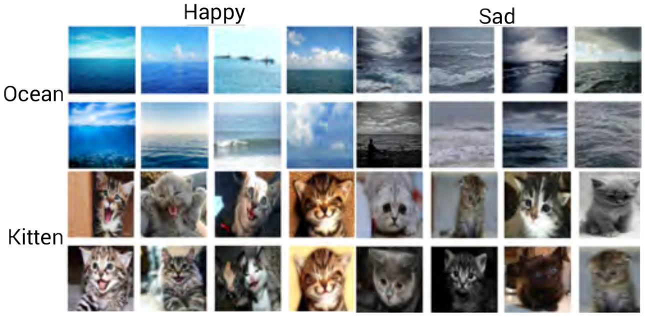

Figure 3: Example images generated by our MC-PixelCNN for mood-phrases like happy-ocean or sad-kitten.

of pixels over an image x is modeled as a product of conditional distributions. Formally, it can be defined as follows.

$$p(x) = \prod_{i=1}^{n^2} p(x_i|x_1, ..., x_{i-1}) \tag{1}$$

where x_i refers to a pixel in the image. In order to ensure the dependency condition (in Equation 1) is satisfied, we used masked convolution filters. A stack of such filters is applied over an input image $I \in \mathbb{R}^{N\times N\times 3}$. Recent work by Oord et al. shows that a softmax distribution tends to perform well even though the data is inherently continuous [Oord et al. 2016]. Therefore, for each pixel 256-way prediction is performed for the three color channels (R, G, B) successively, i.e., conditioned on the previous color channel and predicted sequentially. The output of the MC-PixelCNN is $I_{out} \in \mathbb{R}^{N\times N\times 3\times 256}$. The same GRU is used as in the PixelCNN decoders [van den Oord et al. 2016].

$$z = tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x), \tag{2}$$

where $*$ denotes a convolution operation, $\odot$ denotes element-wise multiplication, k is the layer index, f and g denote filter and gate respectively, and W is a convolution filter.

The conditional distribution $p(x|h)$ of images is modeled by representing the text phrases from song lyrics (P) and the mood extracted from song audio (M) as latent vectors h_P and h_M, respectively:

$$p(x) = \prod_{i=1}^{n^2} p(x_i|x_1, ..., x_{i-1}, h_P, h_M) \tag{3}$$

. Formally, the implementation computes the following:

$$y = tanh(W_{k,f} * x + U_{k,f}^T h_P + V_{k,f}^T h_M)$$
$$\odot\sigma(W_{k,g} * x + U_{k,g}^T h_P + V_{k,g}^T h_M)$$

where k is the layer number, h_M is a one-hot encoding that specifies a mood class and h_P is a phrase representation computed by summing the GloVe word-vector representation [Levy et al. 2015]. This representation incorporates co-occurrence statistics of words that frequently appear together in a text document.

3.4 Textures

3.4.1 Deep Texture Generation. This step is not necessary as the MC-PixelCNN output images can be directly used as textures in the VR scene. We take a playful approach to the visual aesthetic of the generated scenes by processing the output images from the MC-PixelCNN through a DeepDream based CNN. DeepDream creates surreal textures (Fig. 4) that we apply to objects in our generated VR scenes. We believe these textures provide a more nuanced and perhaps playful interpretation of the literal representation of mood and content in the images generated by the MC-PixelCNN (Fig. 3), even more so after manipulating tiling values in Unity (Fig 7) .

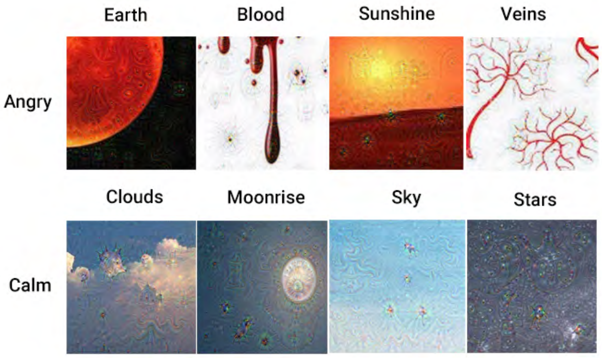

Figure 4: Textures created by our DeepDream based CNN. The input images were generated by the MC-PixelCNN using the mood and nouns as shown, extracted from two songs. The input ’angry’ song was Drop The World by Lil Wayne and the ’calm’ song was Constellations by Jack Johnson.

For creating surreal textures, we use a pre-trained Inception model and let the network make decisions about selecting which features amplify. Each input image is run through DeepDream 24 times, zooming into the image at each iteration to enhance the features that are detected by the layers of the CNN. A single input image results in 24 output textures (Fig. 6). Since higher layers extract more sophisticated features, complex structures and objects emerge when we pull out textures from the layers just before the classification layer.

3.5 VR World Generation

This is the final step in the pipeline. Using the noun-phrases extracted from song lyrics, we select objects from a tagged 3D model dataset to add to the scene. Additional information about which objects to add to the scene and whether the scene is indoors or outdoors comes from the results of User Study I (Section 4.1). In the study we asked participants to describe places and things they associated with happy and sad moods. Once we know which objects will be part of the scene, we texture them with the output of DeepDream. The type of lighting to use in the scene is based on feedback from participants in User Study I.

3.5.1 World Design. Each generated VR world (Fig. 5 is composed of four types of elements: (i) location or setting, (ii) scene objects, (iii) atmospheric elements, and (iv) interactive elements. We chose to use low poly 3D models as they were freely available and allowed for faster and easier building of a tagged 3D model dataset.

3.5.2 Location or Setting. These are the base elements the VR world is built upon. In our examples, each world has at least two

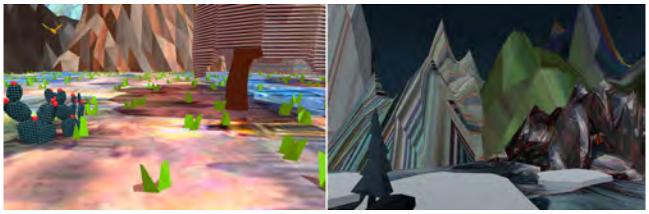

Figure 5: Left) 'Happy' VR world generated from The Bird and the Worm by Owl City. The image shows a kaleidoscope of colors and patterns in this bright outdoor scene. Right) 'Sad' world generated from Blue Prelude by Nina Simone. The image shows the textured surreal looking mountains in this night scene.

such elements, the sky and the ground terrain. For indoor spaces these would include the floor, walls and ceiling. These elements are large and visible to the user from anywhere in the scene and influence the feel of the virtual space. We use a Perlin noise [Perlin 2002] based terrain generator that allows us to alter the terrain size, cell size and noise scale to create a variety of terrains from flat lands to rolling hills to steep mountains. To simplify object placement in our generated scenes we set the noise value to zero for a flat terrain. We realized a little late that we could have used music to generate our terrain which would have added another layer of mood-based connectivity between our input and output. Current time constraints do not allow for adding this to the system presented in this paper but we plan to include it in our next iteration.

3.5.3 Scene Objects. In User Study I we asked participants to list descriptors of places and things they associated with happy and sad moods. Objects from the compiled results, together with objects corresponding to noun-phrases extracted from song lyrics, are picked from a tagged 3D model dataset to add to the VR scene. Similar to [Sra et al. 2016] we use a genetic algorithm (GA) [Whitley 1994] with elitism to model the optimization function for the placement of objects in the scene using a set of created rules that define spatial relationships between them. The rules take into account orientation relative to the center of the VR world, which is the place where the user begins the VR experience when they put on the HTC Vive. While our existing 3D model dataset is not exhaustive, we believe the variety of scenes we can generate will only increase as the size of this dataset increases.

3.5.4 Atmospheric Elements. At the time of writing this paper, these elements were added manually to the generated scene. This includes elements like lighting, particle effects, fog etc. For our generated worlds, the type of lighting to add to a scene was based on data from User Study I where predominantly the 'happy' scene was described with words like 'sunny', 'bright', and 'daylight' and the 'sad' scene was described with words like 'dark' and 'night'.

3.5.5 Interactive Elements. The generated worlds have the potential to provide a novel spatial and immersive musical experience by including the song audio in the scene. While some animation is added automatically e.g., moving textures on objects, others like objects pulsing to the beat of music are added manually. We believe visual indicators of music playing in the scene can add dimensionality to the musical VR experience.

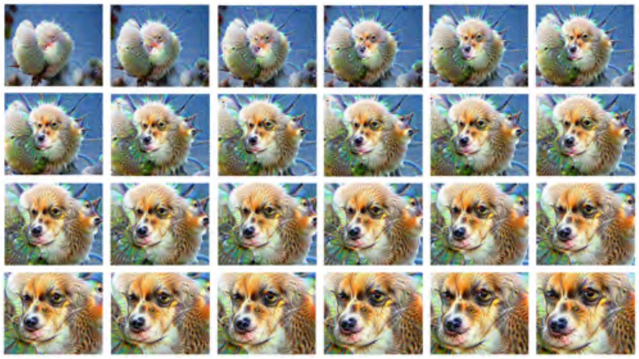

Figure 6: Textures created by feeding images generated by the MC-PixelCNN into a DeepDream based CNN. In this example an image of a cotton ball against a bright blue sky was generated using mood 'happy' and phrase 'cotton' from the song The Bird and the Worm by Owl City which is the song used to generate the 'happy' VR world (Fig. 5a).

3.5.6 Texture Mapping. Light and color are inextricably linked as perception of color depends on the type of light and the interactions between the color and light. Light also impacts non-visual psychological processes like mood and cognitive performance [Knez 2001]. In our generated worlds, the colors and patterns come from the textures applied to objects in the scene. For the two example scenes presented in this paper, 42 images were generated for the 'happy' song and 21 images for the 'sad' song by the MC-PixelCNN. These generated images were fed into a DeepDream based CNN to output 24 psychedelic textures per input image (Fig. 6).

In Unity, a material was created by randomly selecting one texture from the set of 24 textures and automatically applied to an object. When creating a new material, we altered the X and Y tiling values to create variety. When the tiling is set to 1:1 (Fig. 7), the

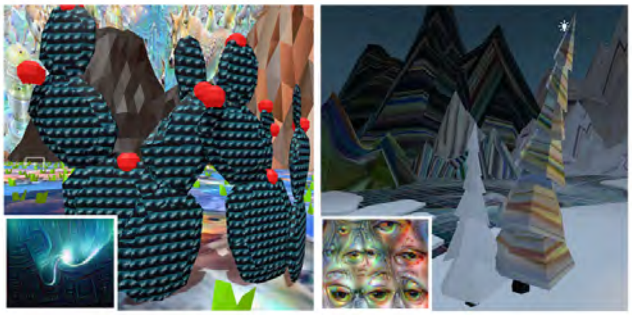

Figure 7: To create visual variety in the scene, texture tiling is altered by modifying the X and Y values. Left: X and Y values are both set to 1 to create a repeating tiled texture. Inset shows the original texture before tiling. Right: X is 0.1 and Y is 1 to create a horizontally scaled texture which preserves the colors but not the details.

original texture is visible as a checkered pattern. Modifying X and Y values results in horizontal or vertical stripes (Fig. 7).

4 EVALUATION

We conducted two user studies. The first study (55 participants) evaluated the quality of the images generated by our MC-PixelCNN to validate whether the images successfully encoded mood + phrase data. We used the same study to collect data from users about their associations between places, objects, lighting and moods. The second study (12 participants) was done to validate the generated VR scenes and to understand if they successfully conveyed the mood of the songs used to create them.

4.1 User Study I

An email with a url to the online study website was sent to our department’s internal mailing list which includes students, staff, and faculty. Fifty-five participants responded and completed the online study over two days. No demographic data was collected. The study consisted of two parts: an evaluation of 20 images generated by the MC-PixelCNN and a question asking participants to describe or list places or things they associate with a particular mood.

To evaluate the output of our MC-PixelCNN, participants were shown 20 generated images, one by one, for each of the 4 mood groups (sad, happy, calm, angry) in a random order along with the mood + phrase pair used to generate that image. Participants were asked to rate how well the images matched the displayed moods on a 1-5 Likert scale where 1 meant ‘poor match’ and 5 meant ‘perfect match’. Results (Fig. 8) show that our model successfully generated good images for ‘happy’ and ‘sad’ moods but did not do well for ‘angry’ images and therefore we only generated example VR worlds for happy and sad moods. Looking at the images generated for anger + noun-phrase, we learned that encoding anger in color was easier than encoding anger in an object. We believe our results can be improved by building a curated mood image dataset for training the MC-PixelCNN model. Collecting data about mood associations from a larger set of participants and for a broad variety of mood terms would also help improve image generation results. Participants were asked, *“What are the elements that constitute a happy place for you?"* for both a happy and a sad place. They responded with answers like ‘outdoors,’ ‘sunny day,’ ‘smile,’ ‘mountaintop,’ ‘trees,’ ‘dark night,’ etc. These responses supplemented the noun-phrases extracted from the lyrics to help select objects from a 3D model dataset to place in the VR scenes. Responses were also used to inform the design of lighting in each scene.

4.2 User Study II

To validate whether our system was able to create virtual worlds that conveyed the mood of the songs used to create them, we conducted a VR user study with 12 participants.

4.2.1 Method. Twelve volunteers (Ages 22 - 58, Mdn = 34, 8 Female) were recruited via email for the study conducted with an HTC Vive in a 2.5x2.3m tracked space. The study setup included 2 songs, The Bird and the Worm by Owl City (happy) and Blue Prelude by Nina Simone (sad) and two corresponding VR scenes.

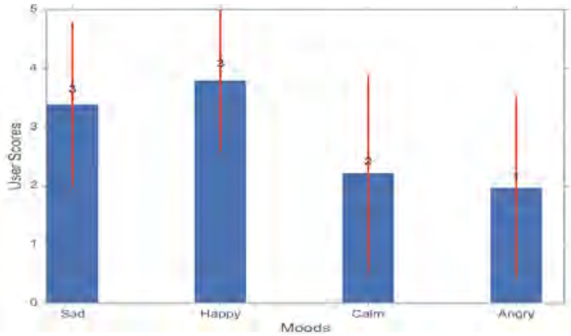

Figure 8: Evaluation of images generated by the trained MC-PixelCNN model. Results show good image generation for happy and sad moods and poor image generation for angry moods.

4.2.2 Procedure. The study procedure took 23 minutes on average, and included listening to a song, experiencing a generated virtual world, filling out questionnaires and providing open-ended feedback. Before starting, participants were asked to sign a consent form and they received instructions about the study. After the initial orientation, participants were asked to listen to a happy or sad song, following which they filled out a questionnaire with one question “*Thinking about the song you just heard, please describe how it made you feel."* Participants were then asked to view the world that was created from a song other than the one they just heard. If they listened to the happy song, they viewed the world generated from the sad song and vice versa to prevent any bias from the listening experience affecting the VR experience and to be able to compare the responses independently from the two stimuli.

Before starting the VR experience, we demonstrated how to use the hand-held controller for navigation. Participants were told that they could use a combination of walking and teleportation to explore the virtual space. An experience was considered complete when the participant had spent at least 7 minutes in the VE, which is slightly longer than 2X the length of either song. The lowest time spent in VR was 6 minutes by P12 while the longest times spent were 17 minutes by P5 and 14 minutes by P1. At the end of the exploration, participants were asked to fill out another questionnaire with one question “*Thinking about the VR space, please describe how it made you feel."* We used the Witmer and Singer presence questionnaire (PQ) to evaluate the subjective experience in VR [Witmer and Singer 1998] and it was filled out last. Since our goal was to create a pipeline for generating VR worlds from music, we used presence to evaluate if the generated worlds provided a satisfying VR experience.

5 DATA AND ANALYSIS

5.1 Presence Questionnaire

We grouped the questions in PQ into three classes for factor analysis and took the average of 7-point rating scale responses across all questions in a single category. The responses from each category were converted to a 3-point scale: high [5-7], neutral [4], and low [1-3] for analyzing the distribution of the participants across this scale using a chi-square (χ^2) test (Fig. 10). Values between 3-4 and

4-5 were treated as low and neutral respectively. Factor analysis explains three loaded factors that collectively affect presence: Spatial Presence, which is related to the sense of being in the VE, Involvement, which describes the environment's ability to stimulate the senses and encompasses emotional responses to the VR experience, and Realism, which is related to the consistency of information in the VE with the real world [Witmer and Singer 1998]. The reported overall rating of presence across all participants was 5.11/7, SD = .81) derived from the average of ratings for the three factors. Average ratings were high for Spatial Presence (M = 5.87/7, SD = .63) and Involvement (M = 5.22/7, SD = .51) and medium for realism (M = 4.25/7, SD = .91). We can infer that though the participants were engaged and present in the virtual world, they perceived the visual representation and interactions as not realistic. Since the objects in the VEs are textured with psychedelic images output by DeepDream and movement in the VE is done using a combination of walking and teleportation, this result is expected.

5.2 Emotion Questions

In response to the emotion question, one participant wrote, "arousal high, valence positive - definitely a happy place" (P4, Happy World) and another wrote, " feeling homesick to places i have never been to ... simple life" (P7, Sad World). Since we asked users to describe how they felt, a mapping of the text responses to emotions would help quantify and compare the responses. Emolex[5][Mohammad 2012] is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). We tokenize each user's text response and use Emolex to associate the response with a distribution over emotions. Since the songs and VEs correspond to happy and sad moods in our user study, we look at the user's response and its association with "joy" and "sadness". Based on this distribution we are able to classify if the user's response for a song and a scene is associated with happy or sad moods. We find that 84% of the users felt "joy" listening to the happy song and experiencing the VE generated from it. Though 84% of the users felt "sadness" listening to the sad song, only 33% of them felt "sad" viewing the sad scene (Fig. 9). Users, instead, described the scene as 'alone', 'beautiful', 'relaxed', 'cold' etc.

5.3 Discussion

All 12 participants reported that moods conveyed by the VR scenes matched those of the songs in general but they were more closely matched for the happy scene and less so for the sad scene. We think the muted colors and abstract patterns conveyed the mood to a far greater extent than the scene composition of camping alone in a faraway place surrounded by mountains. To create better sad scenes, we would need to collect a lot more data on people's associations with sadness and sad places and maybe even design newer ways to visually encode that data into VR scenes. Surprisingly, participants gave the VR experience a high presence score even though there was only one sensory modality enabled and minimal interaction. We expect presence to go higher after we enable song audio in each scene and add interactivity to the experience, based on user feedback. An interesting test would be to see if participants can identify the song from which the world they view is created. As described in the data and analysis section, participants rated Spatial Presence and Involvement factors high, while Realism was rated low (Fig. 10. We discuss these factors with respect to our system.

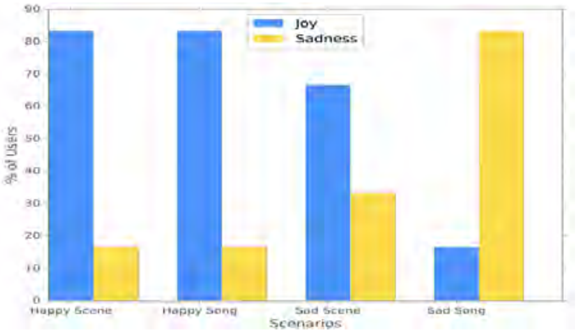

Figure 9: 84% of the users felt "joy" listening to the happy song and viewing the happy world. While 84% of the users felt "sadness" listening to the sad song, only 33% of them felt the same when viewing the sad scene.

Spatial Presence: Participants felt spatially present in our system. We believe the main factors that contributed to the spatial presence were the design and lighting (outdoors sunny day and outdoors dark night). Participants described the scenes as "imaginative," "childlike," "a dream," and "fantasy" probably due to the surreal textures on familiar objects like mountains and trees, and the scale of the spaces.

Involvement: A high rating for involvement suggests that participants were likely engaged in exploring the VR worlds, despite the lack of audio and pre-defined tasks. The were asked to explore the worlds and we implemented full access to all parts of the world via teleportation. Participants were free to climb the mountains or get on top of trees or clouds. Two participants enjoyed the falling sensation that came from standing on a cloud and waiting for it to move from under them and repeated this many times. 8/12 went to each peak in the scene and explored every aspect of the world.

Realism: Unrealistic movement mechanism (teleportation) and inability to grasp objects and interact with them caused the realism to be rated low during the study. We observed several participants reaching out with their hand to touch and interact with objects they were viewing in the HMD. All participants walked in the tracked space to walk in the virtual world and only when they reached the boundary did they engage the teleportation mechanic. This leads us to believe that a walking based navigation system like redirected walking [Razzaque et al. 2001] may have led to the reported perceived realism.

We ran a short pilot study with five participants, after enabling audio in each scene and adding simple interactive elements, to learn whether the generated VEs could provide novel musical experiences. Qualitative feedback was positive and almost invariant. Subjects were excited and expressed a desire for real-time changes to the scene beyond the pulsing objects and the animated textures. We plan to incorporate these in our next version and to run a larger study that explores this immersive form of experiencing music. Our

[5] Details at: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

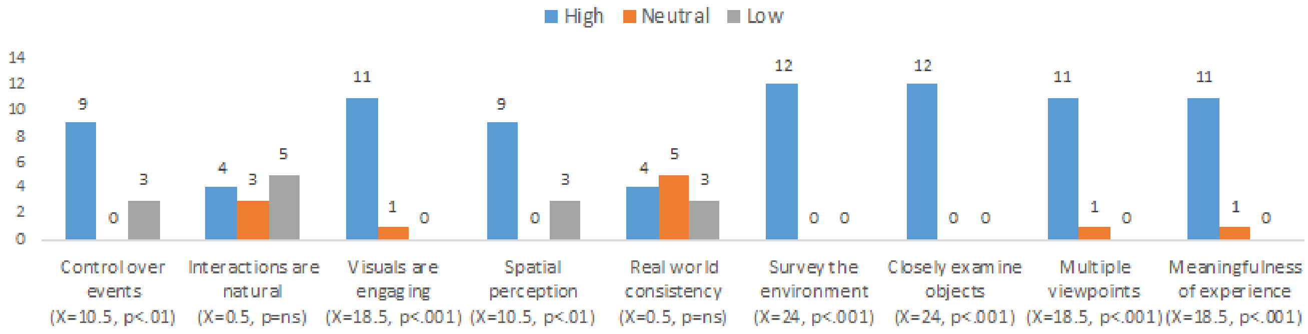

Figure 10: The distribution of participants across questions from the presence questionnaire. Below each question are the chi-square test ($\chi^2_{(2,N=12)}$) results. For factor analysis, we group these questions into three classes of Spatial Presence, Involvement, and Realism.

high level goal was to create a visual and spatial music experience in VR. Since in the first user study, participants experienced the generated worlds and the songs separately, we ran the pilot to see if the generated VR worlds did indeed create an interesting musical experience. Our hope was that this study could help provide us with some ideas for future work in designing novel musical experiences. The positive feedback suggests that newer ways to convert music into 3D experiences may be worth exploring.

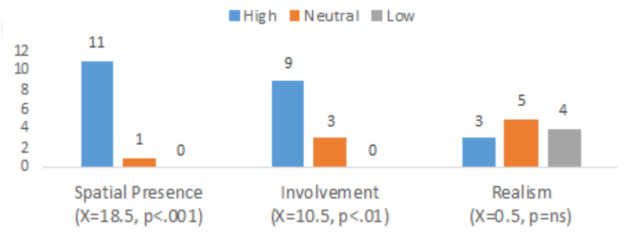

Figure 11: The distribution of participants across three classes. Below each category are the chi-square test ($\chi^2_{(2,N=12)}$) results.

5.4 Affective Virtual Spaces

While the emotions elicited by the happy song and scene correlated well, the same was not true for the sad song and the sad scene. We think this is because participants found the lonely dark world beautiful, dreamy and relaxing and enjoyed the solitude instead of feeling lonely and sad. Even though we designed our world with objects and lighting that people associated with sad places, the resulting outcome did not successfully encapsulate the sadness. This could mean that more people agree on what constitutes happy places but they differ on sad places. This also makes automatically generating sad places a more challenging task than happy places and something we will explore further.

6 LIMITATIONS AND FUTURE WORK

We generated our image dataset from freely available online images so the relationship between the image and the tagged mood may not as accurate as a curated dataset. Since we use a CNN to generate surreal textures, it is possible that the generated virtual worlds will start having a similar look despite different MC-PixelCNN generated images and texture tiling. This can be countered by adding DeepStyle [Gatys et al. 2016] to the pipeline, using different shaders, or by using a combination of colors and patterns. With the prevalence of digital cameras, we have witnessed an explosion of digital photos on the Internet. In contrast, the growth of free 3D models with metadata has been relatively slow. While many techniques have been proposed to enrich the set of available 3D models, their availability is still quite limited. This restricts the types of VR worlds we can generate as we have a very small database of 3D models from which we extract a subset of models to use in each scene. There is much room for improvement as well as opportunity for further development in creating affective virtual spaces. An immediate enhancement would be to encode audio data directly into the scene by using it for terrain generation. Another easily implemented modification would be to allow multiple users to share the same virtual world. A promising direction would be to employ more sophisticated procedural map generation techniques for creating the virtual world and automating the addition of atmospheric elements. An interesting challenge would be to estimate lighting data from the generated images in order to automate the addition of lights to the 3D scenes.

7 CONCLUSION

This work is a first step towards understanding how we may embed emotive content automatically in a VE. In future iterations of Auris, larger online studies could better inform the associations people have between moods, places and objects which could help model new ways of encoding that data into the Auris pipeline. A curated image dataset of mood-based images would improve the trained model and generate better mood related images. Having prior knowledge about the context of the generated VE could also able customized integration of directly referential associations into the pipeline to further enhance the mood conveyed, e.g. the color or image of a spicy pepper could inform the design of a 'fiery' dragon in a grim fantasy VR experience.

REFERENCES

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset.. In *ISMIR*, Vol. 2. 10.

Pietro Casella and Ana Paiva. 2001. Magenta: An architecture for real time automatic composition of background music. In *Intelligent Virtual Agents*. Springer, 224–232.

Chin-Han Chen, Ming-Fang Weng, Shyh-Kang Jeng, and Yung-Yu Chuang. 2008. Emotion-based music visualization using photos. *Advances in Multimedia Modeling* (2008), 358–368.

Kang Chen, Yukun Lai, Yu-Xin Wu, Ralph Robert Martin, and Shi-Min Hu. 2014. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. *ACM Transactions on Graphics* 33, 6 (2014).

Kang Chen, Kun Xu, Yizhou Yu, Tian-Yi Wang, and Shi-Min Hu. 2015. Magic decorator: automatic material suggestion for indoor digital scenes. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 232.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

Annabel J Cohen. 1999. The functions of music in multimedia: A cognitive approach. *Music, mind, and science* (1999), 53–69.

Emily L Denton, Soumith Chintala, Rob Fergus, and others. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in neural information processing systems*. 1486–1494.

Colin Ellard and others. 2015. *Places of the Heart*. Bellevue Literary Press,.

Arnaud Emilien, Ulysse Vimont, Marie-Paule Cani, Pierre Poulin, and Bedrich Benes. 2015. Worldbrush: Interactive example-based synthesis of procedural virtual worlds. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 106.

Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. 2012. Example-based synthesis of 3D object arrangements. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 135.

Lise Gagnon and Isabelle Peretz. 2003. Mode and tempo relative contributions to ?happy-sad? judgements in equitone melodies. *Cognition & Emotion* 17, 1 (2003), 25–40.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

Patrick G Hunter and E Glenn Schellenberg. 2010. Music and emotion. In *Music perception*. Springer, 129–164.

Patrik N Juslin. 2011. Music and emotion: Seven questions, seven answers. *Music and the mind: Essays in honour of John Sloboda* (2011), 113–135.

Patrik N Juslin and Petri Laukka. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research* 33, 3 (2004), 217–238.

Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, and others. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10, 2 (2016), 99–111.

Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proc. ISMIR*. Citeseer, 255–266.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Igor Knez. 2001. Effects of colour of light on nonvisual psychological processes. *Journal of environmental psychology* 21, 2 (2001), 201–208.

Carol L Krumhansl. 1997. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 51, 4 (1997), 336.

Rikard Küller, Seifeddin Ballal, Thorbjörn Laike, Byron Mikellides, and Graciela Tonello. 2006. The impact of light and colour on psychological mood: a cross-cultural study of indoor work environments. *Ergonomics* 49, 14 (2006), 1496–1507.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3 (2015), 211–225.

Hugo Liu and Pattie Maes. 2005. The Aesthetiscope: Visualizing Aesthetic Readings of Text in Color Space.. In *FLAIRS Conference*. 74–79.

Lie Lu, Dan Liu, and Hong-Jiang Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing* 14, 1 (2006), 5–18.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 83–92.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit.. In *ACL (System Demonstrations)*. 55–60.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793* (2015).

Clare Cooper Marcus and Marni Barnes. 1999. *Healing gardens: Therapeutic benefits and design recommendations*. John Wiley & Sons.

CLB McCloughan, PA Aspinall, and RS Webb. 1999. The impact of lighting on mood. *International Journal of Lighting Research and Technology* 31, 3 (1999), 81–88.

Martina T Mitterschiffthaler, Cynthia HY Fu, Jeffrey A Dalton, Christopher M Andrew, and Steven CR Williams. 2007. A functional MRI study of happy and sad affective states induced by classical music. *Human brain mapping* 28, 11 (2007), 1150–1162.

Saif Mohammad. 2012. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 587–591.

Meinard Müller. 2007. *Information retrieval for music and motion*. Vol. 2. Springer.

Asma Naz, Regis Kopper, Ryan P McMahan, and Mihai Nadin. 2017. Emotional Qualities of VR Space. *arXiv preprint arXiv:1701.06412* (2017).

Andrew Y Ng and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems* 2 (2002), 841–848.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759* (2016).

Ken Perlin. 2002. Improving noise. In *ACM Transactions on Graphics (TOG)*, Vol. 21. ACM, 681–682.

Sharif Razzaque, Zachariah Kohn, and Mary C Whitton. 2001. Redirected walking. In *Proceedings of EUROGRAPHICS*, Vol. 9. Citeseer, 105–106.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

John A Sloboda. 1992. Empirical studies of emotional response to music. (1992).

Ruben Michaël Smelik, Tim Tutenel, Klaas Jan de Kraker, and Rafael Bidarra. 2011. A declarative approach to procedural modeling of virtual worlds. *Computers & Graphics* 35, 2 (2011), 352–363.

Misha Sra, Sergio Garrido-Jurado, Chris Schmandt, and Pattie Maes. 2016. Procedurally generated virtual reality from 3D reconstructed physical space. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. ACM, 191–200.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

Robert E Thayer. 1990. *The biopsychology of mood and arousal*. Oxford University Press.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, and others. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*. 4790–4798.

Menno Van Zaanen and Pieter Kanters. 2010. Automatic Mood Classification Using TF* IDF Based on Lyrics.. In *ISMIR*. 75–80.

Lois B Wexner. 1954. The degree to which colors (hues) are associated with mood-tones. *Journal of applied psychology* 38, 6 (1954), 432.

Darrell Whitley. 1994. A genetic algorithm tutorial. *Statistics and computing* 4, 2 (1994), 65–85.

Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments* 7, 3 (1998), 225–240.

Charlotte VO Witvliet and Scott R Vrana. 2007. Play it again Sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial EMG, and heart rate. *Cognition and Emotion* 21, 1 (2007), 3–25.

Dan Yang and Won-Sook Lee. 2004. Disambiguating Music Emotion Using Software Agents.. In *ISMIR*, Vol. 4. 218–223.

Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. 2008. Toward multi-modal music emotion classification. In *Pacific-Rim Conference on Multimedia*. Springer, 70–79.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. *arXiv preprint arXiv:1509.06041* (2015).

Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494.

Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 47–56.