Automatic Transfer of Musical Mood into Virtual Environments

Sangyoon Han
POSTECH
Pohang, Korea
han0209@postech.ac.kr

Amit Bhardwaj
POSTECH
Pohang, Korea
amitbhardwaj@postech.ac.kr

Seungmoon Choi
POSTECH
Pohang, Korea
choism@postech.ac.kr

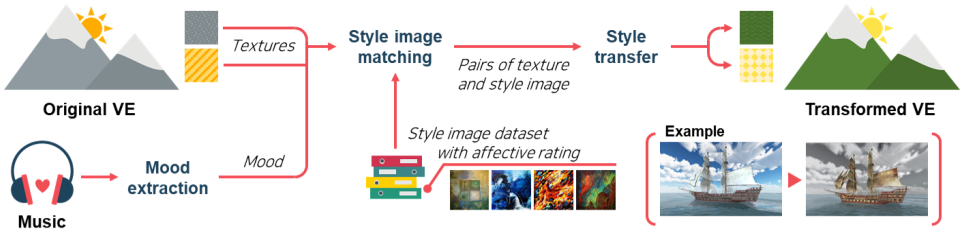

Figure 1: Procedure of our musical mood transfer to virtual environments.

ABSTRACT

This paper presents a method that automatically transforms a virtual environment (VE) according to the mood of input music. We use machine learning to extract a mood from the music. We then select images exhibiting the mood and transfer their styles to the textures of objects in the VE photorealistically or artistically. Our user study results indicate that our method is effective in transferring valence-related aspects, but not arousal-related ones. Our method can still provide novel experiences in virtual reality and speed up the production of VEs by automating its procedure.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Applied computing** → Sound and music computing;

KEYWORDS

Virtual environment, mood, music, transfer, affect

ACM Reference Format:
Sangyoon Han, Amit Bhardwaj, and Seungmoon Choi. 2018. Automatic Transfer of Musical Mood into Virtual Environments. In *VRST 2018: 24th ACM Symposium on Virtual Reality Software and Technology (VRST '18), November 28-December 1, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3281505.3281534

1 INTRODUCTION

A virtual environment (VE) creates an immersive experience that is not possible in physical reality. However, developing VEs is a labor-intensive and time-consuming task requiring complex processes and considerable expertise in 3D modeling and the design of material, sound, effects, interaction, and levels. Particularly in level design, some processes are often repeated and modified to suit the intent of the VE designers, and here automatic methods for VE generation can greatly improve productivity. In fact, several systems have been proposed to automate VE generation: scene arrangement [10], modeling [6], generation [20], interactive synthesis of virtual worlds [8], and a procedural method [18]. They, however, do not deal with affective aspects, such as mood and emotion, even though VE designers often attempt to express such attributes.

Our approach tested in this paper is re-stylizing a VE by transferring texture (representing mood) from an image to the VE using image style transfer [11]. The resulting VE can induce substantially different emotional responses. The texture and color scheme of a VE are highly related to its affective aspects [7], and many effective methods on image emotion assignment exist in multimedia and computer graphics [12, 15–17].

The general flow of our method is illustrated in Figure 1. First, we extract mood from music using the random forest classifier that was trained on a large song dataset with affective annotations of valence and arousal (VA). Second, for each texture used in a VE, we select an abstract painting image that has the most similar VA scores to those of the musical mood. For this, we labeled abstract painting images with VA scores by a user study with 30 participants. Lastly, the selected image is transferred to the texture in the VE to match the mood. These two steps are repeated for all textures in the VE. We can play the music during user interaction with the re-stylized VE or use it as just a source for mood selection. We also evaluated our method by a user experiment.

Our work shares the same idea with Sra et al. [21], which provides many useful guidelines for automatic generation of VEs. They presented a system that generates a VE automatically using mood and content extracted from the sound and lyrics of a song. Particularly, they fully automated the process of generating textured 3D objects corresponding to the extracted mood. To identify the mood, they used a categorical model of emotion, which is limited in expressing the diversity of affect. In contrast, we use the continuous VA model, which allows us to produce a greatly larger number of VEs from a single VE and songs. Hence, our work can be regarded as a more focused study in that aspect.

VRST '18, November 28-December 1, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6086-9/18/11…$15.00
https://doi.org/10.1145/3281505.3281534

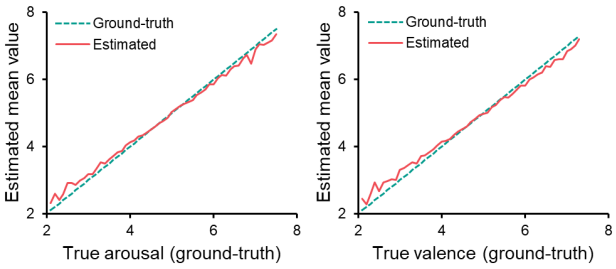

Figure 2: Performance of random forest predictors.

Figure 3: Examples of abstract style images.

2 MOOD EXTRACTION FROM MUSIC

The goal of this step is to estimate the VA scores that represent the mood of a given music piece. We trained a predictor using machine learning on the DEAM (MediaEval Database for Emotional Analysis in Music) dataset [1]. This dataset contains more than 1800 songs and their affective annotations of valence and arousal, both dynamic and static. In this work, we consider only static annotations, which provide the average and standard deviation of the VA ratings for the whole song on a nine point scale. The dataset also provides music features for each song, extracted by openSMILE [9]. We consider a 25-s excerpt, and each 500-ms interval has 261 music features. Thus for each song, we use 50 consecutive feature vectors, each of which has 261 music feature values.

Next, we train a predictor of VA annotations on the selected features using random forest classifier [4]. In the random forest, many decision trees are trained, and the final decision is made by majority voting. The input feature data are divided into two parts: 80% for training and 20% for testing. Before training the predictor, we preprocess the data by excluding those whose valence and arousal scores have less than five songs, and identify 70 and 65 unique values of each, respectively. This step is performed on valence and arousal separately, so we have two sets of data, one for valence and the other for arousal. This ensures that each annotation is represented by sufficient examples in both of the training and test data. As such, we train two random forest predictors for arousal and valence. Each random forest consists of 5000 decision trees.

The random forest predictors built on the training data predict the annotations of the test data with an accuracy up to 84%. In order to illustrate the performance of the predictors over the entire range of the annotations, we plot the estimated mean values of arousal and valence against their respective true (ground-truth) values in Figure 2. The estimated and true scores match quite well over the entire range. Thus, the random forest predictors have good performance in extracting mood from music.

3 TEXTURE TRANSFER

Given a VE and the VA scores VA_{target} representing the musical mood, this step aims to transform the VE to have the similar mood to VA_{target}. We attempt to achieve this goal by transferring a *style image* that has the mood close to VA_{target} to a *content texture* that is overlaid on an object in the VE. This is repeated for all the texture images (e.g., sky sphere and objects) used in the VE.

3.1 Labeling Images with VA scores

We first need a set of images that are used as *style images*. We use a dataset of abstract paintings from [13]. They do not have contextual influence (e.g., of objects, actions, and metaphors), and the mood is determined only by the style and color scheme in the painting. To label each abstract painting with VA scores, we conducted a rating experiment for 280 images in the dataset with 30 volunteers (male 24, female 6, age 18–30 years). A participant looked at each image on a 27-inch monitor without a time limit and entered two scores for valence and arousal on a scale of -100–100 using a graphical interface. Each participant rated each image once. The participants were paid USD 10 each after the experiment. Then for each image we computed the average and standard deviation of the VA scores across the participants. Some examples are presented in Figure 3. We denote each abstract image by I_i and its VA scores by VA_i.

3.2 Texture Transfer Procedure

For each texture T_j used in the VE, we need to match a style image I_i for style transfer. To this end, we define two similarity measures. One is the mood similarity S_{mood} using the z-score distance:

$$S_{mood}(I_i, M) = \frac{1}{1 + \left\| \frac{V_i - V_{target}}{\sigma_{V_i}}, \frac{A_i - A_{target}}{\sigma_{A_i}} \right\|}, \quad (1)$$

where σ_{V_i} and σ_{A_i} are the standard deviations of the valence and arousal score of I_i, and M represents the music with VA_{target}. $0 \le S_{mood} \le 1$, and high S_{mood} means that the moods of I_i and M are close on the VA space.

The other is the color histogram similarity S_{color} to favor a style image that has a similar color scheme to that in the original texture T_j. This is to avoid the uncanny valley that can be caused by large differences in the color scheme. We use color histograms constructed based on the CIELAB color space. We assume that changing hues is mainly responsible for the uncanny valley, but changing luminosity is not. So we use only the a^* channel (green-red) and the b^* channel (blue-yellow) to construct a color histogram. This histogram is also normalized using its L1 norm and denoted by $hist^{a^*b^*}$. So we do not impose constraints on luminosity, and this allows dynamic luminosity changes. Then S_{color} is defined using histogram intersection as

$$S_{color}(I_i, T_j) = \sum_{k=1}^{K} \min\left(hist_k^{a^*b^*}(I_i), hist_k^{a^*b^*}(T_j)\right), \quad (2)$$

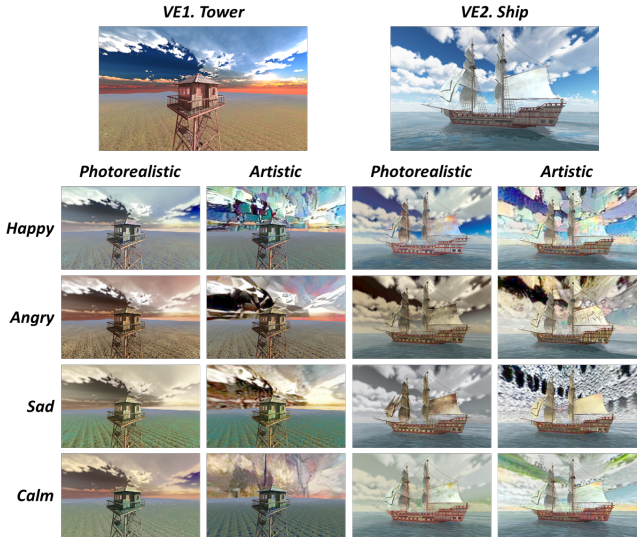

Figure 4: Original and transformed VEs for the user study.

where $hist_k^{a^*b^*}$ is the k^{th} bin of the color histogram, K is the number of bins, and $0 \leq S_{color}(I_i, T_j) \leq 1$.

Then the final similarity measure S is

$$S = \omega_{mood} S_{mood} + \omega_{color} S_{color}, \quad (3)$$

where ω_{mood} and ω_{color} are weights of two similarity measures. We choose a style image that maximizes S. To solve this optimization problem efficiently, we first find ten style images I_i with the highest S_{mood} values. We then select one that maximizes S with $\omega_{mood} = 0.1$ and $\omega_{color} = 0.9$, where the emphasis to S_{color} is because the style images have already been sorted out for S_{mood}.

Now we have style images for all the textures used in the VE. For style transfer, we use the two most popular methods of artistic [11] and photorealistic transfer [19].

4 USER STUDY

We conducted a user experiment to evaluate the quality of the VEs that our mood transfer method makes.

4.1 Methods

We selected four music pieces representing the four quadrants of the VA space, respectively, from the DEAM dataset: Happy–"The Unfinished Sympathy, Teenage Kicks", Angry–"Violins is not the answer, Drop", Sad–"Ben Holmes and Patrick Farrell, Prelude #2 in A Minor (Chopin)", and Calm–"Riding Alone For Thousands Of Miles, satellite". We also used two VEs (a tower and a ship; Figure 4) implemented in Unity3D. The two VEs were transformed artistically or photo-realistically using the moods of the four music pieces. Thus, a total of 18 VEs ((8 transformed + 1 original) × 2 environments) were prepared, as shown in Figure 4.

The experiment consisted of two sessions: 1) VA rating for the music pieces and VEs, and 2) congruence rating for pairs of a music piece and a VE. Before the experiment, the experimenter explained concepts associated with the VA model to participants and helped them to understand the experimental procedure clearly.

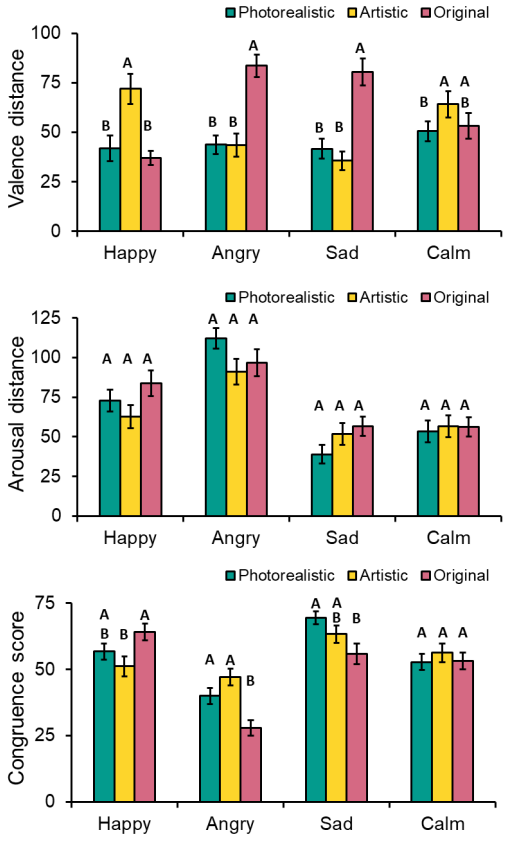

Figure 5: Average distances in valence and arousal between VE and musical mood, and average congruence scores. In each plot, the data of the two VEs were pooled. Error bars represent standard errors. Means marked by the same letters for each mood did not have significant differences by Tukey's tests performed on the data of the same mood.

In Session 1, participants listened to the four music pieces by wearing noise-canceling headphones and explored the 18 VEs by wearing an HMD (Oculus Rift), and then rated VA scores in the 2D space (each axis -100 to 100). In Session 2, participants looked and listened to the 24 combinations of music pieces and VEs (4 music pieces × 2 VEs × (1 original + 2 transformed using the music pieces)) and rated overall congruence on a scale of 0–100. In both sessions, the orders of experimental conditions were randomized per participant. Participants could appreciate the VEs while moving the viewpoint to five predetermined locations using an Oculus remote. They were given a 5-min break after Session 1 and a half of Session 2. The experiment was finished in 75 minutes on average.

Twenty two participants (20 male and 2 female; 20-30 years old, mean 25.0) with normal visual and hearing abilities volunteered for this study. Each of them were paid USD 13 after the study.

4.2 Results and Discussion

To evaluate the extent to which the VEs represent the target musical mood, we looked at distances in the VA space. For each pair of a

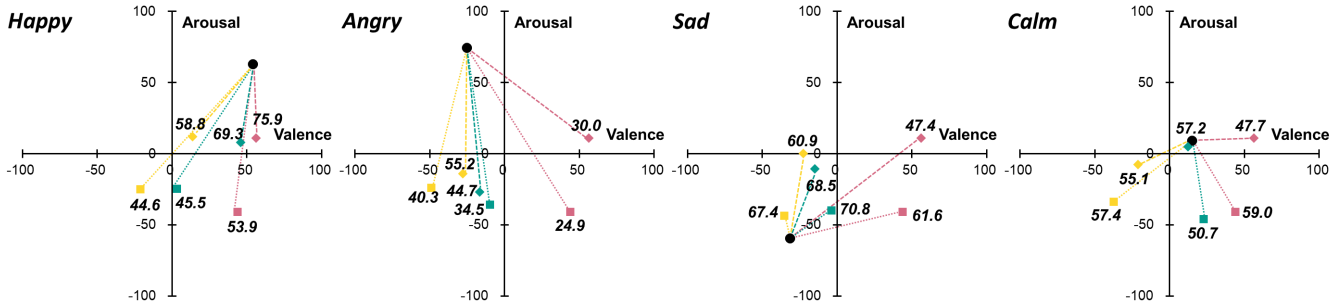

Figure 6: Transitions of the mean VA scores of VEs. •: music (target mood), ♦: original VE1, ♦: photo realistically transferred VE1, ♦: artistically transferred VE1, ■: original VE2, ■: photo realistically transferred VE2, ■: artistically transferred VE2. The number on each point is the mean congruence score of that pair.

music piece and a VE, we computed two 1D distances in valence and arousal between their VA scores, as shown in Figure 5. Congruence scores are also presented in Figure 5 (bottom). We also calculated the mean VA scores of music pieces and VEs to obtain further insights to overall transition tendency (Figure 6). We performed three-way ANOVA on valence distance, arousal distance, and congruence score using style transfer method (original, artistic transfer, and photorealistic transfer), mood (happy, angry, sad, and calm), and VE (VE1 and VE2) as the independent variables.

For *valence* distance, style transfer method was the only statistically significant factor ($F(2, 42) = 3.996, p = 0.0258$). Figure 6 shows that style transfer clearly shifted the valence of an original VE to the negative direction and that artistic transfer did so more than photorealistic transfer. In particular, the valence of photorealistically transferred VEs were relatively well aligned with the valence of the target music, except when the target music was for happy. These tendencies are also faithfully reflected in the valence distance plot of Figure 5. Photorealistic transfer resulted in closer valence distances to the target mood than artistic transfer when the target mood was happy and calm, and similar valence distances for angry and sad.

We note that all the target musical moods happened to have a much more negative valence than the original VEs, except for the case of happy music. This seems to be a main reason for why all the valance shifts by style transfer were negative. Besides, the uncanny valley, which has a negative impact on valence [5], can be a source. We observe some cases that can cause the uncanny valley in the transformed VEs. For example, the green grass is turned blue (VE1, Happy), and the blue sky is turned green (VE2, Calm). Our transfer algorithm tries to prevent the uncanny valley by comparing the color histograms, but it seems to be incomplete. We need another study to see if our mood transfer method can add positive valence.

Dominance, which refers to whether you feel in control, powerful, or overwhelmed [3], can be related to our result that artistic transfer led to more negative valance. Dominance is another dimension used to describe affect, and it plays an important role in determining the mood in environmental perception [2]. As dominance decreases, valence tends to decrease [14]. Unlike viewing artistic paintings, people are unfamiliar with exploring artistic environments. We guess that the participants could have been overwhelmed by the artistic environments, which resulted in decrease in dominance, and subsequently in valance.

For *arousal* distance, style transfer method was not significant, but mood and VE were ($F(3, 63) = 11.84, p < 0.0001$ and $F(1, 21) = 10.61, p = 0.0038$). This can also be seen in Figure 5 (middle) and 6. Thus, our present method is not effective in injecting the arousal-related characteristics of music into a VE.

The two VEs were static without any change, and the only user interaction supported was viewpoint movement. These are reflected in the low arousal scores of the VEs in Figure 6. Our texture-based mood transfer method, which affects only the appearances of objects, seems to be insufficient to modulate the arousal of such static VEs. We may need to control more dynamic factors, e.g., level of interaction, object animation, and VE's context.

For *congruence*, style transfer method was not significant, but mood and VE were ($F(3, 63) = 30.35, p < 0.0001$ and $F(1, 21) = 8.41, p = 0.0086$). This indicates that the congruence between a VE and the mood of the music played with the VE is mainly determined by the pair itself (see Figure 5, bottom), but not heavily by the textures of the objects in the VE.

5 CONCLUSIONS

We have proposed a complete pipeline for automatic transformation of a VE by transferring abstract images representing the mood of a musical piece to the textures of the VE. The performance of our system was evaluated by a user study. Our method was shown more effective in transporting aspects related to valence than arousal.

Our approach has several contributions. (1) It enables change of the mood of the VE. VE designers can use our system to emphasize their intentions after creating a VE. (2) It improves the reusability of VEs. A single VE can be transformed to an unlimited number of VEs using a wide variety of music. (3) It imposes texture consistency on objects in the same VE, even for objects with initially diverse textures. (4) It was the first trial to transfer images affectively while excluding the context in the images. Meanwhile, we will need to test whether our method can transfer positive valance and find more effective ways for arousal transfer.

ACKNOWLEDGMENTS

This research was supported by a grant (NRF-2016M3C1B6929724) for Convergence R&D over Science, Technology, and Liberal Arts by the National Research Foundation of Korea. w

REFERENCES

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

[2] Iris Bakker, Theo van der Voordt, Peter Vink, and Jan de Boon. 2014. Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current Psychology* 33, 3 (2014), 405–421.

[3] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[5] Marcus Cheetham, Lingdan Wu, Paul Pauli, and Lutz Jancke. 2015. Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. *Frontiers in psychology* 6 (2015), 981.

[6] 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).

[7] Colin Ellard et al. 2015. *Places of the Heart.* Bellevue Literary Press,.

[8] 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.

[9] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[10] 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.

[11] 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.

[12] Li He, Hairong Qi, and Russell Zaretzki. 2015. Image color transfer to evoke different emotions based on color combinations. *Signal, Image and Video Processing* 9, 8 (2015), 1965–1973.

[13] 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.

[14] Giulia Miniero, Andrea Rurale, and Michela Addis. 2014. Effects of arousal, dominance, and their interaction on pleasure in a cultural environment. *Psychology & Marketing* 31, 8 (2014), 628–634.

[15] Naila Murray, Sandra Skaff, Luca Marchesotti, and Florent Perronnin. 2011. Towards automatic concept transfer. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*. ACM, 167–176.

[16] Chuong H Nguyen, Tobias Ritschel, Karol Myszkowski, Elmar Eisemann, and Hans-Peter Seidel. 2012. 3D material style transfer. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 431–438.

[17] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 860–868.

[18] 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.

[19] Cameron Smith. 2016. neural-style-tf. https://github.com/cysmith/neural-style-tf. (2016).

[20] 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.

[21] Misha Sra, Pattie Maes, Prashanth Vijayaraghavan, and Deb Roy. 2017. Auris: creating affective virtual spaces from music. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. ACM, 26.