Scene-Aware Background Music Synthesis

Yujia Wang
Beijing Institute of Technology
wangyujia@bit.edu.cn

Wei Liang*
Beijing Institute of Technology
liangwei@bit.edu.cn

Wanwan Li
George Mason University
wli17@gmu.edu

Dingzeyu Li
Adobe Research
dinli@adobe.com

Lap-Fai Yu
George Mason University
craigyu@gmu.edu

ABSTRACT

Background music not only provides auditory experience for users, but also conveys, guides, and promotes emotions that resonate with visual contents. Studies on how to synthesize background music for different scenes can promote research in many fields, such as human behaviour research. Although considerable effort has been directed toward music synthesis, the synthesis of appropriate music based on scene visual content remains an open problem.

In this paper we introduce an interactive background music synthesis algorithm guided by visual content. We leverage a cascading strategy to synthesize background music in two stages: *Scene Visual Analysis* and *Background Music Synthesis*. First, seeking a deep learning-based solution, we leverage neural networks to analyze the sentiment of the input scene. Second, real-time background music is synthesized by optimizing a cost function that guides the selection and transition of music clips to maximize the emotion consistency between visual and auditory criteria, and music continuity. In our experiments, we demonstrate the proposed approach can synthesize dynamic background music for different types of scenarios. We also conducted quantitative and qualitative analysis on the synthesized results of multiple example scenes to validate the efficacy of our approach.

CCS CONCEPTS

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

KEYWORDS

scene sentiment, background music synthesis, music transition

ACM Reference Format:
Yujia Wang, Wei Liang, Wanwan Li, Dingzeyu Li, and Lap-Fai Yu. 2020. Scene-Aware Background Music Synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413894

*Corresponding author

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413894

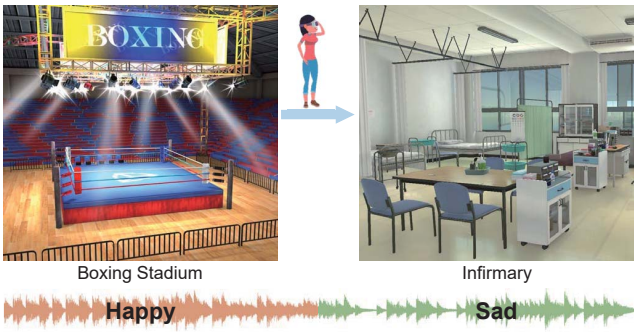

Figure 1: Seamless background music synthesis in accordance with the scene emotion expression during navigation in different virtual environments.

1 INTRODUCTION

Creating an immersive virtual environment that combines both visual and hearing rendering to enhance the multimedia experience is important for games, virtual/augmented reality (VR/AR), and multimedia applications [42]. Throughout the past few decades, significant research effort has focused on improving the visual fidelity using multimedia techniques [48] or high-quality graphics rendering. Compared to visual rendering, the state-of-the-art in audio synthesis based on visual content lags behind. Research in computer vision and graphics enables the generation of appropriate sound according to the events in video progression. However, the synthesis of music based on visual content remains an open problem. Devising algorithms and tools to automatically synthesize real-time desirable music in virtual environments will find useful applications. As shown in Fig. 1, different virtual scenes may be associated with different emotions. Desirably, matching background music should be assigned; and background music should also transition seamlessly between virtual scenes.

The notion of synthesizing background music (BGM) for different environments is widely studied in the context of cognitive psychology. Music has been proven to be a sensory stimulus that can affect human mood and behavior [26] and BGM is a very important element to provide diverse multimedia experience to customers [34]. For example, a shopping mall might play positive and uplifting background music to motivate customers to shop, closing more sales [51]. Current practices commonly resort to composing unique music by hiring professional musicians, or retrieving music that matches the scene followed by cropping, stitching, etc. During the music composition process, the musicians need to pay close

attention to the scene sentiment. As the complexity of the scene increases, producers need to consider the natural transitions between multiple emotions during the music composition process. Consequently, BGM synthesis is time-consuming and costly to produce. Moreover, it does not cater to the different time lengths of each user's music playback and ensure the seamless transition between different music.

We propose a computational approach to facilitate and automate the synthesis of BGM for different places. As illustrated in Fig. 2, given the panorama images of a virtual environment, the core idea of our approach is to provide a desirable BGM for users when navigating different regions to engage them with the audiovisual-based activities in the environment. To overcome the challenge of the non-trivial mapping from the input visual domain to the output audio domain, we devise a cascade visual-aware background music synthesis approach comprising a *Visual Analysis* stage and a *Background Music Synthesis* stage. In the first stage, our approach extracts emotional features from the detected salient objects and the color tone of the scene. In the second stage, guided by the visual analysis results, our approach performs the real-time music composition optimization, considering music transition factors (i.e. chords sequence, pitch, tempo, and key signature).

With the increasing ubiquity of smartglasses for AR interactions, our approach can also be extended to synthesize a matching BGM for real-world scenarios, such as during a museum exhibition tour. In order to satisfy the music preference of different customers, our approach could also take into account the user's preferred playlist during the music synthesis.

Our main contributions in this paper are as follows:

- proposing a computational approach to synthesize background music driven by visual analysis of a scene;
- formulating the music transition problem as an optimization problem considering scene sentiment, tempo, and other music-related properties as constraints;
- demonstrating the proposed approach for different applications and validating its performance through quantitative and qualitative experiments.

2 RELATED WORK

2.1 Background Music Industry

The music industry, as part of the broader "experience design", impacts our everyday life. Music is tightly integrated with different virtual-immersive environment technologies for different purposes, e.g., education [15], gaming [37]. The main goal of background music is to create distinct and cohesive musical identities. Its creation is usually contracted to music companies, such as Music Concierge and Mood Media.

Synthesizing BGM for different virutal scenes needs a lot of human effort and requires professional skills [16]. Imagine creating BGM for a hotel. The music producers have to investigate the themes and functionalities of different areas (e.g., lobby, bar, restaurant, gym) of the hotel, so as to create matching background music for different areas [30]. The full playlist of music choices of music library can sum up between $1,000$ to $8,000$ [1]. The number mostly depends on how many different sections the client wants to divide their playlist into, and the number of different areas at the location. Consequently, the producer may spend lots of time on music composition and editing, incurring considerable costs.

Our work aims to overcome the challenge of synthesizing BGM for scenes. Inspired by the effort of designing AI musical assistants, such as AI drummer [31] and music generator for fearful virtual environment GhostWriter [38], we explore the possibility of developing a computational approach to synthesize matching, dynamic, and personalized BGM guided by the visual scene.

2.2 Visual Sentiment and Music

Psychological studies have shown that human emotions can be affected by visual information [25, 26] as well as music signals [4]. Psychologists found that visual and auditory signals can influence us in two main ways: i) we often subconsciously match our body motion with what we see and hear; ii) audio signals (e.g., music) can trigger us to associate the environment we are in with a certain context and emotion [6].

Visual perception is influenced by image sentiment. Psychologists have found that human attention generally prioritizes emotional contents [45] and low-level features such as color [44]. This process enables us to analyze a scene through regions attended selectively and its color tone. In computer vision, graphics, and image processing, many machine learning and deep learning methods have been proposed for human attention detection and salient object detection [14], image sentiment recognition [43, 54], and color palette extraction [41].

Music theories and practices indicate that music conveys sentiment-related messages, where tempo, pitches, and rhythms are linked to expressing emotions [32] and can influence human's cognition and behaviors [11]. A biosensing prototype [29] has been proposed to transform emotion into music, facilitating social interaction and human-computer interaction in VR games.

Inspired by these studies, we incorporate deep learning techniques to automatically infer the visual sentiment by performing visual analysis on images, which guides the background music synthesis. Our approach allows both general and professional users to synthesize their desired background music.

2.3 Background Music Processing

Visual to Sound. There are studies about the association between vision and sound. Researchers in multimedia, computer vision, and graphics have worked on enhancing the audio-visual diversity of their surrounding environments. For example, techniques have been proposed to generate natural sound for the wild [56], scenes [35], 360° panorama images [22], and music videos [27]. In order to enable "drawing scenes by ear", researchers working on image sonification have proposed approaches to represent visual data by means of sound, which could be applied for blind assistant systems. For example, color information could be represented by audio attributes, e.g., instrument [2], pitch and loudness [10].

However, little research has evaluated the relationships between scene visual sentiment and background music. In this paper, we present a novel approach to synthesize BGM for different scenes, which facilitates the development of human-computer interaction interfaces and virtual world applications promoting seamless interaction.

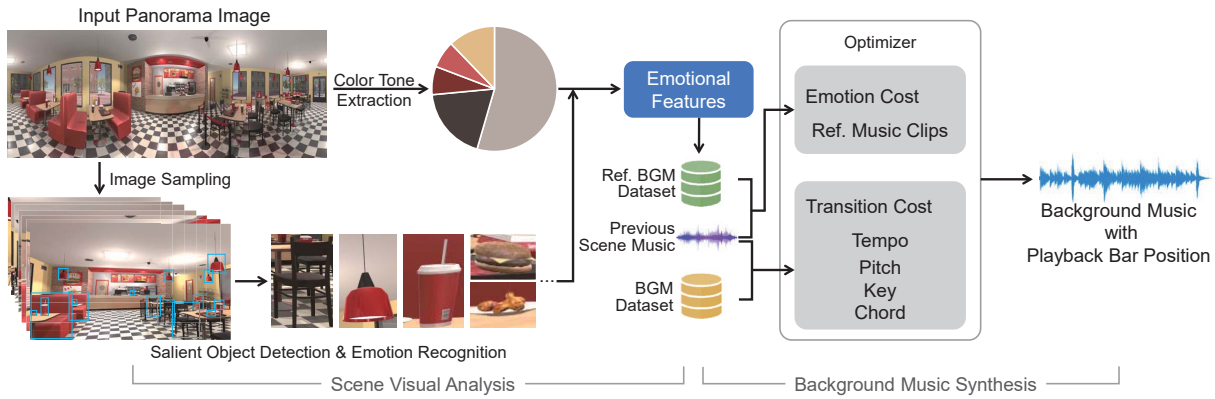

Figure 2: Overview of our proposed approach for scene-aware background music synthesis. It consists of two major steps: *Scene Visual Analysis* and *Background Music Synthesis*.

Music Synthesis and Transition. Music greatly contributes to the immersive multimedia experience of a virtual world. An appropriate choice of BGM can add emotional depth to the experience and help communicate with the listeners [32]. The idea of synthesizing music by concatenating existing music fragments based on euphonious transitions has a long history [33]. A considerable amount of research has been done to automate music retrieval [27, 52], music exploration [3], music playlist creation [5], and music search evaluation [20], which allows effective music recommendations based on specific conditions such as emotional states.

To concatenate the retrieved music natural and seamless, several factors considered during music transition have gained popularity in such fields, such as harmonic similarity [33], chunk similarity [40], pitch [24], tempo and volume [39], and the circle of fifth for the key signature [18]. We draw inspirations from existing music retrieval and transition algorithms and synthesize plausible background music based on visual emotional observations.

3 SCENE VISUAL ANALYSIS

As demonstrated in [55], human are able to perceive and understand scenes in terms of high-level semantics (i.e. emotion expression). The semantic information carried by scene images is conveyed by salient objects [14] and color tone [44]. Inspired by these studies, our algorithm is designed to extract two kinds of emotional features in *Scene Visual Analysis*, i.e. salient object emotional features and color tone features. Such features are used to guide the BGM synthesis to match the semantic information of the virtual scene.

3.1 Object Detection

Our *Scene Visual Analysis* is devised upon robust object detection considering that people have a remarkable ability to attend selectively to some salient objects in an area. For example, people attend to emotional stimuli (i.e. an object that elicits an emotional response for the observer), such as special exhibits in a museum or window displays in a clothing store. Researchers have been incorporating higher-level perceptual properties of images to predict salience [14], and their models have encoded high-level concepts such as faces, interacting objects, and text.

Beginning with a panorama image of the virtual scene and a camera viewing horizontally from the center of the rendered image, we rotate the viewpoint horizontally 36° to capture different segments of the image. We use the proposed network with short connections in [21] to detect salient objects, which is based on VGGNet pre-trained on the MSRA-B dataset [28]. If there are any overlapping objects from one slice to the next, we count the detected objects as the same object. *Scene Visual Analysis* in Fig. 2 depicts the object detection process of the virtual burger store such as the chair, dome light, and food.

The hyper-parameters used in this work contain: learning rate (1e-8), weight decay (0.0005), momentum (0.9), loss weight for each side output (1). Please refer to [21] for more details.

3.2 Feature Extraction

Scenes typically convey emotions via the constituent objects and color tone. Analyzing the scene emotional states helps synthesize vivid BGM, enriching a user's auditory experience.

Object Emotional Features. To capture high-quality emotion information, we apply a deep learning-based visual sentiment recognition model, an AlexNet-styled network [9]. We modify the order of the pooling and normalization layers of the original architeccture. Their network is pre-trained on the dataset collected and released in [53], whose images are labeled with either positive or negative sentiment.

To be consistent with the emotional categories of psychology research in visual and auditory perception, our approach seeks to refine the emotion recognition of the scene images. We do so by replacing the last two-neuron fully-connected layer with a five-neuron layer, which is fed to a softmax that computes the probability distribution over the target classes representing 5 emotion types (i.e. calm, happy, sad, angry, and fearful). We also fine-tuned the pre-trained model on the dataset proposed in [54]. The network was trained using stochastic gradient descent with a momentum of 0.9 and a starting learning rate of 0.001.

For each detected object, we use the extracted feature from the second to the last fully-connected layers as the visual emotion representation, which is represented as a 4, 096-d vector.

Color Tone Features. A good palette extracted from the image is one that closely captures the underlying colors the image was composed of, even if those colors do not appear in their purest form on the image itself, i.e. blended or covered by other colors for example. To extract the color palette of the virtual scene panorama image, our approach computes and simplifies the convex hull enclosing all the color samples [41], providing more general palettes that better represent the existing color gamut of the image.

The palette size is set as 5. Each color in the palette is represented by its RGB value and its corresponding proportion in the five extracted colors. Fig. 2 illustrates the detected color tone of the virtual burger store. The color tone feature, concatenated with the object emotional features, will be used in our background music synthesis (Sec. 4).

4 BACKGROUND MUSIC SYNTHESIS

Based on the emotion representations extracted from the scene image, our approach synthesizes emotional and seamless background music so as to yield attractive multimedia experience for users. Abrupt music transition has been shown to interrupt one's immersive audiovisual experience [7]. To address the smooth music transition challenge, we formulate the music generation process as an optimization with various constraints, including emotional states, music chord progression, tempo, and etc. We will discuss how we define these constraints or cost terms in detail.

4.1 Optimization Formulation

We synthesize scene-aware background music by optimizing against the approximated total cost function (Equation 1). We specify our optimization process on bars, which are segments of time in music corresponding to a specific number of beats. Let $M(n) = (b_1, b_2, ..., b_n)$ denote the currently playing music track, which consists of n bars $\{b_n\}_{n=1}^{N}$ assembled in a sequential order. As the user moves to a different area, the upcoming music should match the currently playing music at the current bar position. When a user moves to a new scene I^*, we choose the optimal upcoming M^* and m^*that minimize the emotion and transition costs:

$$\mathrm{C}_{\mathrm{total}}(M, m, I^*, M^*) = \mathrm{C}_{\mathrm{E}}(I^*, M^*) + \lambda \mathrm{C}_{\mathrm{T}}(M, m, M^*, m^*) \quad (1)$$

where $m \in [1, n]$ and $M = (b_1, b_2, ..., b_m)$ represents the current music that has played m bars. In the following sections, we introduce two cost terms. $\mathrm{C}_{\mathrm{E}}(\cdot)$ is the emotion cost term for evaluating the emotion consistency between the new scene I^* and the upcoming music (M^*). $\mathrm{C}_{\mathrm{T}}(\cdot)$ is the transition cost term, which measures the transition difficulty from the current music to the upcoming music, constraining the pitch, tempo, and chords progression during the optimization, at a specific location from m in M to m^* in M^*. λ is a regularization factor to balance these two terms.

4.2 Dataset

We created two datasets from the Internet to achieve the goal of background music synthesis:

- *Reference BGM Dataset*: contains 300 combinations of scenes and the corresponding background music. The dataset covers a variety of scenes expressing different emotion types (e.g., calm, happy, sad, angry, and fearful). To extract emotional representation from each scene in this dataset, we use the same method as in Sec. 3.
- *BGM Dataset*: includes 1,000 copyright-free music midi files, which are used to synthesize the final output background music. The dataset covers the same emotion types as those of the *Reference BGM Dataset*.

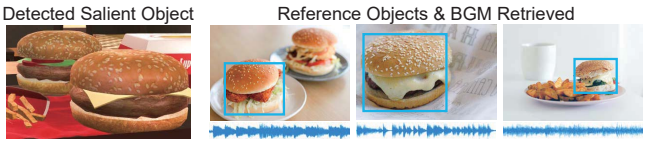

Figure 3: Retrieval of reference music clips that carry similar emotion as that of the query. The similarity is computed based on visual emotional features.

4.3 Emotion Cost

Directly evaluating the emotion consistency between images and the background music is difficult even with the state-of-the-art deep learning techniques. This is due to the significant gap between the visual domain and the audio space and the difficulty of collecting sufficient high-quality multi-modal data [36]. To tackle this challenge, we devise our approach with a *Reference BGM Dataset*, which provides an intermediate, emotion-rich audio representation that makes matching visual contents and music features feasible.

Reference Music Retrieval. For a scene I, our approach retrieves top N pieces of reference music $\{r_i\}_{i=1}^{N}$ from the *Reference BGM Dataset*. The similarity between I and a reference BGM r_i is computed using the mean squared distance (l_2) over their 4,096-d visual emotion features of the detected salient objects and the color tone features of the scene panorama image (Sec. 3). Specifically, we average the sum of l_2 between every two salient objects of I and the reference image, as well as the detected color palette of I and the reference image (corresponding to r_i); the shorter the distance is, the higher the similarity is.

For each reference music r_i, we compute its importance $h_i \in [0, 1]$ using its similarity to I within the visual emotion space (i.e. salient object emotion and color tone), normalized over the similarities of all the retrieved reference music $\{r_i\}_{i=1}^{N}$. Fig. 3 presents some examples of the retrieved reference BGM, which carry similar emotion as that of the query. In our implementation, we use $N = 3$.

Music Emotion Recognition. Before describing the details of the emotion cost term in Equation (1), we use a music emotion recognition model [23] based on which the cost term is defined. We trained the model on the CAL500exp dataset [46], which contains 3223 items annotated with a dependent emotion tag (i.e. calm, happy, sad, angry, and fearful).

Based on the emotion recognition model, each type of emotion style embedding of the reference music set $\{r_i\}_{i=1}^{N}$ is computed as a weighted sum: $\sum_{i=1}^{N} h_i \cdot p_{l_j}(r_i)$, where $\sum_{i=1}^{N} h_i = 1$, and $p_{l_j}(\cdot) \in [0, 1]$ is the classification score of the emotion label l_j. Specifically, l_j is the index of 5 different labels, i.e. $l_j \in \{\text{calm, happy, angry, sad, fearful}\}$.

$C_E(\cdot)$ drives the upcoming background music M^*, retrieved from the collected *BGM Dataset*, to be consistent with the inferred emotion states of the reference music set $\{r_i\}_{i=1}^{N}$, thus matching the conveyed emotion of I. It is defined as:

$$C_E(\cdot) = \sum_{j=1}^{5} \sum_{i=1}^{N} |h_i p_{l_j}(r_i) - p_{l_j}(M^*)|. \quad (2)$$

4.4 Transition Cost

A transition from music M to another music M^* sounds natural when the pitch and tempo at the bar b_{m^*} of M^* are similar to those of the previously played bar b_m of M; and if the key change and the chord progression from M^* is harmonious to M. We compute the transition cost $C_T(\cdot)$ by combining these four cost terms:

$$\begin{aligned} C_T(\cdot) = {} & w_t D_t(b_m, b_{m^*}) && \rightarrow \text{ tempo distance} \\ & + w_p D_p(b_m, b_{m^*}) && \rightarrow \text{ pitch distance} \\ & + w_k D_k(b_m, b_{m^*}) && \rightarrow \text{ key distance} \\ & + w_c D_c(M, M^*) && \rightarrow \text{ chord progression.} \end{aligned} \quad (3)$$

4.4.1 Tempo, Pitch, and Key Distance Cost. We compute the average BPM (beats per minute), a measure of tempo, of each bar and then take the absolute BPM difference between pairs (b_m, b_{m^*}) to get $D_t(\cdot)$. We estimate pitch differences by computing chroma features [13] for each beat in a bar, and then compute the average cosine distances $D_p(b_m, b_{m^*})$ between bar pairs. For the key signature distance between M and M^*, we follow the method proposed by Clough et al. [12], which can be easily calculated by the musical theory tool of "Circle of Fifths".

4.4.2 Chord Progression Cost. In order to keep the music playback harmonious between different clips, the chord sequence of M^* should harmoniously progress with the bars of M that have been played. We apply a text-based Long Short-Term Memory (LSTM) [19] network based on which the cost term is defined.

Network Architecture. LSTM network is a redesign of Recurrent Neural Network (RNN) to help explore and store information for longer periods of time, which is applied to infer the chord progression based on the played bars of M^*. Fig. 4 depicts the model architecture. We use four different types of layers, i.e. LSTM layers, dropout layers, dense layers, and the activation layers.

We use a three-layer LSTM structure, each of which consists of 256 hidden units. The output of each LSTM unit is used as an input to all of the LSTM units in the next layer and so on. Dropout of 0.3 is added after every LSTM layers. In the last layer we use "Softmax" activations. The weights of the network are updated iteratively.

Training Data. To train our model, we collected a music chord progression dataset, which consists of 1k chord sequences from the McGill Billboard Chord Dataset [8] and the Internet. The dataset covers different music genres, such as classic, pop, jazz, and blues. We further convert all chord sequences under different key signatures to key C major to maintain the consistency of the training data. The training chord sequences are divided into fragments, each of which containing 10 chords.

Inferring. With the trained model, we infer the upcoming chord progression based on the played bars of M. Specifically, for each

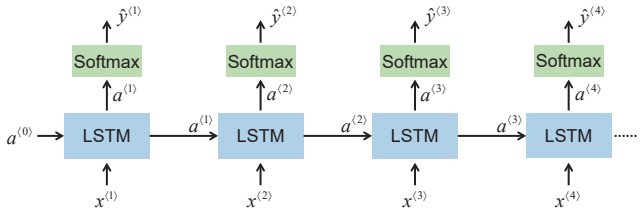

Figure 4: The LSTM structure of the training process.

inferred chord in the progression, we chose the one with the highest probability as the inferred result. We design to infer the longest chord progression without repeating sequences.

Based on our LSTM model, we elaborate the cost $D_c(\cdot)$ defined as the similarity between inferred chord progression and M^*, i.e. the ratio's reciprocal of the same chord sequence length. If the length is 0, we set $D_c(\cdot)$ as 1. The chords are estimated by [50].

The weights w_t, w_p, w_k, and w_c allow us to control the acceptable range of pitch, tempo, and key distances, and the harmony of chord progression in music transitions. These transition cost terms ensure that the seamless music transition progression has minimal transition cost (i.e. $C_T(\cdot) \rightarrow 0$).

4.5 Background Music Optimization

Since our optimization problem is combinatorial and the number of combination items can vary (e.g., retrieved music can be played in any bars), it is difficult to define a closed-form solution. To deal with this complication, we adopt the Reversible Jump MCMC (RJMCMC) framework to explore the space of possible music tracks extensively.

To efficiently explore the solution space, we apply the simulated annealing process in the optimization process. We define a Boltzmann-like objective function:

$$f(M^*) = \exp(-\frac{1}{t} C_{\text{total}}(M, I, M^*)), \quad (4)$$

where t is the temperature of the simulated annealing process, which decreases gradually throughout the optimization. There are two types of moves that can be selected by the optimizer:

(1) *Swap Music*: randomly change to another music from *BGM Dataset*;
(2) *Retrieve Bars*: randomly change to another bar within the music.

The selection probabilities of the moves to swap music and retrieve bars are p_m and p_b. By default, we set the selection probabilities as $p_m = p_b = 0.5$.

To decide whether to accept the proposed music M^*, our approach compares the total cost value $C_{\text{total}}(M, I, M^*)$ of the proposed M^* with the total cost value of $C_{\text{total}}(M, I, M_0)$ of the original state with a randomly selected bar of a randomly selected music M_0. To maintain the detailed balance condition of the RJMCMC method, the acceptance probabilities $Pr(M^*|M_0)$ are equivalent for the two move types: $Pr(M^*|M_0) = \min(1, f(M^*)/f(M_0))$.

At the beginning of the optimization, the temperature t is set to be high (1.0) to prompt the optimizer to aggressively explore possible solutions. The value of t decreases by 0.2 every 100 iterations until it reaches zero. We terminate the optimization if the

Figure 5: Example virtual scenes used in both quantitative and qualitative experiments. Here we display images captured from the main view. Please refer to our supplementary material for the corresponding panorama images and the synthesized background music for different navigations.

absolute change in the total cost value is less than 5% over the past 30 iterations.

4.6 Playback Configuration

When the upcoming music M^* is determined through the optimization process, after the current bar of M is played, our approach automatically plays the first two to four bars of M^* with the following configurations: gradually speed up or slow down the tempo, increase or decrease the volume. In practice, we find that this leads to a more comfortable auditory experience during music transition.

5 EXPERIMENTS

In this section, we discuss several quantitative and qualitative experiments conducted to evaluate the effectiveness of our scene-aware background music synthesis approach. Our approach was implemented on an Intel Core i7-5930K machine running in an NVIDIA TITAN GPU with 12GB graphics card memory.

5.1 Methods for Comparison

Different Approaches. We compared three approaches of background music synthesis:

- Randomly synthesized background music;
- Our synthesized scene-aware background music;
- Professionally composed background music. We recruited three professionals who have been studying music theory and music composition for 5 years.

We compared results of these approaches in quantitative and qualitative experiments.

Validation Dataset. We created a virtual city to conduct experiments (shown in Fig. 5), which consists of 9 different scenes, namely, *City Center*, *School*, *Burger Store*, *Supermarket*, *Infirmary*, *Abandoned Street*, *Boxing Stadium*, *Gun Shop*, and *Haunted House*. The scenes are associated with different emotion states. We navigated in the virtual world and generated 15 navigations of different scenes. Each navigation contains four randomly selected scenes with random navigation time (at least 20 seconds at each scene). The background music for each navigation is synthesized by different approaches aforementioned, i.e. each navigation has 3 background music generated by the 3 different approaches. Please refer to supplementary materials for the validation dataset.

Quantitative Experiment. All three approaches used our collected *BGM Dataset*. We measured the performance of the results in emotion expression and music transition. Specifically,

- for navigation in each scene, we used scene and music emotion recognition models to recognize the emotional state (through salient objects and color tones) of the scene and the corresponding music. The recognition results are represented by a 5D one-hot vector, each element of which presents an emotion state (i.e. calm, happy, sad, angry, and fearful). We computed the Euclidean distance as the emotion expression error;
- for each transition error of all navigations, we averaged the tempo, pitch, and key distances. We recruited 10 professional musicians and asked them to score the harmony of the chord progression synthesized by different approaches;
- for each navigation, we recorded the synthesis time of the BGM synthesis process of three different approaches.

Qualitative Experiment. We recruited 20 participants with reported normal or corrected-to-normal vision, no color-blindness, and normal hearing. Half of the participants are professional musicians. Before each study, the participants were given a task description and encouraged to ask any question. The participants sat 35cm in front of a screen (with 1440×900 resolution). Auditory input was provided by a pair of Logitech G430 headphones with 7.1 channel surround sound output.

The goal of this experiment is to evaluate how well the music we synthesized matches the corresponding navigation. We asked the participants to rate the emotion expression, transition seamlessness, and overall experience of navigation with different background music, using a 1-5 Likert scale, with 1 meaning rough music transition and bad visual correspondence performance and 5 meaning the opposite. The music pieces were randomly selected from the results of different approaches so as to avoid bias.

5.2 Results and Analysis

Quantitative Experiment. We conducted quantitative experiment with each approach discussed in Sec. 5.1, synthesizing background music to match with different scenes. The results are shown in Fig. 6. The results show the capability of each approach in synthesizing background music in accordance with the scene visual information and in performing seamless transition simultaneously.

For emotion expression, the overall average inconsistency across all virtual scenes of the results synthesized by our approach attained the lowest error ($M = 0.81, SD = 0.57$), closely following the results composed by professional musicians (*Professional Synthesis*) ($M = 0.85, SD = 0.64$). The errors of our approach are higher than those of *Professional Synthesis* on pitch distance (the former: $M = 0.27, SD =$

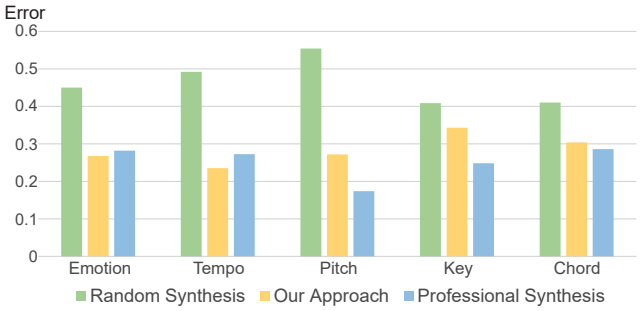

Figure 6: Quantitative errors of different approaches, including emotion expression and transition. For visualization of different errors, we normalize the results to [0,1] accordingly.

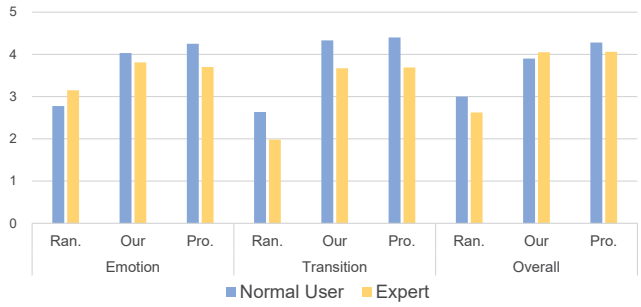

Figure 7: User ratings of emotion expression, seamless transition, and overall experience of different approaches, i.e. *Random Synthesis*, our approach, and *Professional Synthesis*.

0.24; the latter: $M = 0.17, SD = 0.26$), key distance (the former: $M = 0.34, SD = 0.23$; the latter: $M = 0.25, SD = 0.49$), and chord progression harmony (the former: $M = 0.34, SD = 0.23$; the latter: $M = 0.25, SD = 0.49$). The error bar of tempo depicts different results. Our results obtained the lower error ($M = 0.24, SD = 0.22$) than the results of *Professional Synthesis* ($M = 0.27, SD = 0.40$), indicating that our approach is strict in music tempo control. As shown in Fig.6, results on chord progression are comparable, i.e. our approach resulted in an error of $M = 0.30, SD = 0.37$, and *Professional Synthesis* resulted in an error of $M = 0.28, SD = 0.27$. The quantitative results indicate that our LSTM chord prediction model is suitable for our task.

In addition, we note that creating background music manually could be challenging as it involves music selection, cropping, and stitching to match the visual observations on different scenes, as well as performing seamless transition between different scenes. Our approach automates these tasks.

To demonstrate that our approach can perform real-time background music synthesis, we recorded the time for creating the background music for each navigation using different approaches. Results can be created much faster using our approach ($M = 0.025$ mins, $SD = 0.007$ mins) compared to *Professional Synthesis* ($M = 34.23$ mins, $SD = 13.09$ mins). The experts claimed that they need to find music that matches the emotional expression of the scene in the dataset for nearly 20 minutes, and then perform the editing between music clips. Therefore, our computational approach makes it possible for non-experts to generate background music for real-time applications.

Qualitative Experiment. Our qualitative experiment consisted of three parts: i) emotion expression evaluation (verifying the emotion consistency between virtual scenes and the corresponding background music); ii) music transition evaluation (verifying the seamlessness of music transition); iii) overall evaluation (verifying the experience enhancement of background music during navigation in the virtual environment). We conducted a semi-structured interview about the users' experience to explore other factors influencing the ratings.

Fig. 7 shows the ratings of the qualitative experiments. We display the ratings by the normal users and experts, respectively. Overall, the average emotion expression, music transition, and overall ratings are 2.96, 2.31, and 2.81 for the *Random Synthesis* results, 3.92, 4.00, and 3.97 for our results, and 3.98, 4.05, and 4.17 for the *Professional Results*. For the three kinds of ratings on the music synthesized by different approaches, the ratings of professionals are mostly lower than that of normal users. This can be explained by the fact that experts are more rigorous on music performance evaluation, especially on music transition. The detailed ratings tabulated by user types can be found in the supplementary material.

To confirm that our results are efficacious, we performed One-Way ANOVA test in qualitative experiments. In all cases, our syntheses are more preferable than random syntheses and are comparable to professional syntheses. The p-values are less than 0.05 between the results of *Random Synthesis* and our approach: emotion ($F_{[1,599]} = 9.125, p < .5$); transition ($F_{[1,449]} = 13.504, p < .05$); overall ($F_{[1,149]} = 8.370, p < .05$). On the contrary, there are no significant differences between our results and *Professional Synthesis* results: emotion ($F_{[1,599]} = 0.418, p = .68$); transition ($F_{[1,449]} = 0.126, p = .90$); overall ($F_{[1,149]} = 1.910, p = .06$).

To verify the factors considered in our optimization contributed to the overall experience, we computed Bivariate (Pearson) correlation coefficients between the ratings of the overall experience and other factors respectively. There are positive correlations between the overall experience and emotion expression ($r = .65, p < .05$); and between the overall experience and transition seamlessness ($r = .58, p < .05$). The higher the music performance with respect to emotion and transition is, the higher the overall experience.

Besides the two considered factors, some users commented that the music delay during the scene transition could affect the overall experience. This is a limitation posed by the seamless transition. The music delay happens not only in our results (caused by optimization process), but also occasionally in the professional result (caused by the complete playback of phrases). Moreover, some users commented that the absence of characters and events in scenes will affect the emotion judgement. This can be explained by the fact that human beings may have strong a priori judgement about the emotion associated with some scene types. For example, a boxing arena is often associated with the emotion of excitement. Such

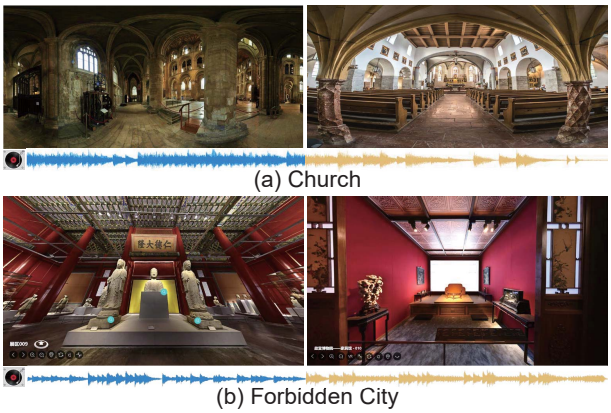

Figure 8: Music synthesis for real-world scenarios. With the (a) panorama images captured in churches or (2) 3D tours in the online museum of Forbidden City (©The Palace Museum [1]), our approach synthesized background music to play during the navigations.

feedbacks give us some interesting insights about considering prior knowledge of scene activities in synthesizing background music.

6 APPLICATIONS

We demonstrate the efficacy of our approach and present several useful applications. As our results consist of synthesized background music, we encourage readers to watch the accompanying video to view and hear the full results. The background music could be synthesized in real-time when the user navigates a virtual environment, and it could also seamlessly transition according to the scene that the user navigates to.

6.1 Game Music Design

Game designers are aware that music is not merely a detached backdrop to the action on screen - it can ebb and flow, adding emotional depth and breadth to scenes. It also helps maintain the player attention and considerably changes how the players play the game [49]. For example, music can be used to depict characters, to signal a change in the game state, to communicate an event to the player, or to build up the dramatic tension of a scene. Our approach supports game music design, such as shooting game *PUBG (PlayerUnknown's Battlegrounds)*. The real-time BGM can be automatically synthesized according to the virtual scene where the user is located. Please refer to supplementary materials for the synthesized results. Game designers can also specify the emotion type according to game storytelling, e.g., synthesizing background music for a role-playing game based on the game character's location in a virtual world.

6.2 Real World Music Design

Our approach can automatically synthesize background music for 3D virtual tours. For example, as shown in Fig. 8 (a), given the panorama images taken in churches as input, our approach synthesizes matching background music that the user can listen to while viewing the panorama images on a screen or in virtual reality. Our approach can be similarly applied for synthesizing background music for other virtual tours such as for real estate showcases.

6.3 Preferences-based Music Design

To match the synthesized music with the music preferences of different users, our approach could constrain the solution space of the music optimization on two conditions analyzed from the user's preferred music playlist: (a) we can use the method proposed by [17] to detect the music genre preference of users and constrain the solution space with the corresponding music genres; (b) we can constrain the tempo range using the beats per minute of each music in a playlist. As shown in Fig. 8 (b), our approach can be applied to synthesize background music during the navigation in the palace museum of Forbidden City[1]. With the user input playlist, the detected music genre (i.e. Chinese-style music) and tempo (i.e. between 60 bpm to 100 bpm) were used during the background music synthesis. Please refer to the supplementary material for the results.

7 CONCLUSION

We propose a computational approach to automatically synthesize real-time background music considering the virtual scene that a user is navigating. Guided by a visual sentiment analysis, our approach synthesizes music that matches the emotion states conveyed by the scene, as well as allowing seamless transition. Our approach can enable interesting applications such as music design for game and real world scenarios, and can also incorporate a user's music preference.

Limitations and Future Work. Scenes may be associated with different intended uses and styles featuring different lights, backgrounds, layouts, etc. Due to the difficulty of learning effective and general emotion representations of such factors by computer vision techniques, we only tested our approach on salient objects. Based on our approach, the user might circumvent such issues by explicitly specifying emotions deemed appropriate for the scenes when synthesizing background music.

To perform visual analysis, we use static panorama images as input for recognizing the corresponding emotion expression. The advancement of real-time visual information input, such as video recording based on the user's head pose or gaze [47], will reduce the impact of the information that users are not interested in and targeted analyze the visual information, thus enhancing the background music performance.

Our current approach considers emotion and transition parameters for synthesizing background music for scenes. It would be interesting to consider additional factors such as the continuation of emotional state during the transition between different scenes, as well as storytelling during the navigation. Future work may also consider additional musical instruments to yield more diverse background music.

ACKNOWLEDGMENTS

The authors are grateful for Adobe's support in this research. We thank Kelian Li et al. for the help of music generation for different virtual scene navigations.

[1] https://en.dpm.org.cn/

REFERENCES

[1] 2018. Inside the booming business of background music. https://www.theguardian.com/news/2018/nov/06/inside-the-booming-business-of-background-music.

[2] Sami Abboud, Shlomi Hanassy, Shelly Levy-Tzedek, Shachar Maidenbaum, and Amir Amedi. 2014. EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative neurology and neuroscience* 32, 2 (2014), 247–257.

[3] Mohammed Habibullah Baig, Jibin Rajan Varghese, and Zhangyang Wang. 2018. MusicMapp: A Deep Learning Based Solution for Music Exploration and Visual Interaction. In *ACM Multimedia*. 1253–1255.

[4] Laura-Lee Balkwill and William Forde Thompson. 1999. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music perception: an interdisciplinary journal* 17, 1 (1999), 43–64.

[5] Jared S Bauer, Alex Jansen, and Jesse Cirimele. 2011. MoodMusic: a method for cooperative, generative music playlist creation. In *UIST*. ACM, 85–86.

[6] Thomas Baumgartner, Michaela Esslen, and Lutz Jäncke. 2006. From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International journal of psychophysiology* 60, 1 (2006), 34–43.

[7] Axel Berndt, Knut Hartmann, Niklas Röber, and Maic Masuch. 2006. Composition and arrangement techniques for music in interactive immersive environments. *Audio Mostly* 2006 (2006), 53–59.

[8] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. 2011. An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis.. In *ISMIR*, Vol. 11. 633–638.

[9] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing* 65 (2017), 15–22.

[10] Sofia Cavaco, J Tomás Henriques, Michele Mengucci, Nuno Correia, and Francisco Medeiros. 2013. Color sonification for the visually impaired. *Procedia Technology* 9 (2013), 1048–1057.

[11] Fu-Yin Cherng, Yi-Chen Lee, Jung-Tai King, and Wen-Chieh Lin. 2019. Measuring the Influences of Musical Parameters on Cognitive and Behavioral Responses to Audio Notifications Using EEG and Large-scale Online Studies. In *ACM SIGCHI*. ACM, 409.

[12] John Clough and Gerald Myerson. 1986. Musical scales and the generalized circle of fifths. *The american mathematical monthly* 93, 9 (1986), 695–701.

[13] Daniel PW Ellis and Graham E Poliner. 2007. Identifyingcover songs' with chroma features and dynamic programming beat tracking. In *ICASSP*, Vol. 4. IEEE, IV–1429.

[14] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. 2018. Emotional attention: A study of image sentiment and visual attention. In *CVPR*. 7521–7531.

[15] Eric Fassbender, Deborah Richards, Ayse Bilgin, William Forde Thompson, and Wolfgang Heiden. 2012. VirSchool: The effect of background music and immersive display systems on memory for facts learned in an educational virtual environment. *Computers & Education* 58, 1 (2012), 490–500.

[16] JG Fox. 1971. Background music and industrial efficiency - a review. *Applied ergonomics* 2, 2 (1971), 70–73.

[17] Heitor Guimaraes. 2018. Music Genre classification using Convolutional Neural Networks. Github.

[18] Johann David Heinichen. 1969. Der General-Baß in der Composition [1728]. *Hildesheim: Olms* (1969).

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Christine Hosey, Lara Vujović, Brian St Thomas, Jean Garcia-Gathright, and Jennifer Thom. 2019. Just Give Me What I Want: How People Use and Evaluate Music Search. In *ACM SIGCHI*. ACM, 299.

[21] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. 2017. Deeply supervised salient object detection with short connections. In *CVPR*. 3203–3212.

[22] Haikun Huang, Michael Solah, Dingzeyu Li, and Lap-Fai Yu. 2019. Audible Panorama: Automatic Spatial Audio Generation for Panorama Imagery. In *ACM SIGCHI*. ACM, 621.

[23] Danyal Imran. 2016. Music Emotion Recognition. Github. https://github.com/danz1ka19/Music-Emotion-Recognition.

[24] Junki Kikuchi, Hidekatsu Yanagi, and Yoshiaki Mima. 2016. Music composition with recommendation. In *UIST*. ACM, 137–138.

[25] Peter J Lang. 1979. A bio-informational theory of emotional imagery. *Psychophysiology* 16, 6 (1979), 495–512.

[26] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 1998. Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological psychiatry* 44, 12 (1998), 1248–1263.

[27] Jen-Chun Lin, Wen-Li Wei, James Yang, Hsin-Min Wang, and Hong-Yuan Mark Liao. 2017. Automatic music video generation based on simultaneous soundtrack recommendation and video editing. In *ACM Multimedia*. 519–527.

[28] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. 2010. Learning to detect a salient object. *TPAMI* 33, 2 (2010), 353–367.

[29] Xi Lu, Xiaohang Liu, and Erik Stolterman Bergqvist. 2019. It sounds like she is sad: Introducing a Biosensing Prototype that Transforms Emotions into Real-time Music and Facilitates Social Interaction. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, LBW2219.

[30] Vincent P Magnini and Emily E Parker. 2009. The psychological effects of music: Implications for hotel firms. *Journal of Vacation Marketing* 15, 1 (2009), 53–62.

[31] Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. 2019. In a Silent Way: Communication Between AI and Improvising Musicians Beyond Sound. In *ACM SIGCHI*. ACM, 38.

[32] Leonard B Meyer. 2008. *Emotion and meaning in music*. University of Chicago Press.

[33] Meinard Müller and Jonathan Driedger. 2012. Data-driven sound track generation. In *Dagstuhl Follow-Ups*, Vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[34] Adrian C North, David J Hargreaves, and Jon J Hargreaves. 2004. Uses of music in everyday life. *Music Perception: An Interdisciplinary Journal* 22, 1 (2004), 41–77.

[35] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *ECCV*. Springer, 801–816.

[36] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2013. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI* 36, 3 (2013), 521–535.

[37] Debbie Richards, Eric Fassbender, Ayse Bilgin, and William Forde Thompson. 2008. An investigation of the role of background music in IVWs for learning. *ALT-J* 16, 3 (2008), 231–244.

[38] Judy Robertson, Andrew de Quincey, Tom Stapleford, and Geraint Wiggins. 1998. Real-time music generation for a virtual environment. In *Proceedings of ECAI-98 Workshop on AI/Alife and Entertainment*. Citeseer.

[39] Steve Rubin and Maneesh Agrawala. 2014. Generating emotionally relevant musical scores for audio stories. In *UIST*. ACM, 439–448.

[40] Zhengshan Shi and Gautham J Mysore. 2018. LoopMaker: Automatic Creation of Music Loops from Pre-recorded Music. In *ACM SIGCHI*. ACM, 454.

[41] Jianchao Tan, Jyh-Ming Lien, and Yotam Gingold. 2016. Decomposing images into layers via RGB-space geometry. *TOG* 36, 1 (2016), 1–14.

[42] Zhenyu Tang, Nicolas Morales, and Dinesh Manocha. 2018. Dynamic Sound Field Synthesis for Speech and Music Optimization. In *ACM Multimedia*. 1901–1909.

[43] Quoc-Tuan Truong and Hady W Lauw. 2017. Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In *ACM Multimedia*. 1274–1282.

[44] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of experimental psychology: General* 123, 4 (1994), 394.

[45] Patrik Vuilleumier. 2005. How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences* 9, 12 (2005), 585–594.

[46] Ju-Chiang Wang, Hsin-Min Wang, and Shyh-Kang Jeng. 2012. Playing with tagging: A real-time tagging music player. In *ICASSP*. IEEE, 77–80.

[47] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. 2019. A deep Coarse-to-Fine network for head pose estimation from synthetic data. *Pattern Recognition* 94 (2019), 196–206.

[48] Yujia Wang, Wenguan Wang, Wei Liang, and Lap-Fai Yu. 2019. Comic-guided speech synthesis. *TOG* 38, 6 (2019), 1–14.

[49] Guy Whitmore. 2003. Design with music in mind: A guide to adaptive audio for game designers. *Gamasutra, May* 29 (2003).

[50] Yiming Wu and Wei Li. 2019. Automatic audio chord recognition with midi-trained deep feature and BLSTM-CRF sequence decoding model. *Transactions on Audio, Speech and Language Processing* 27, 2 (2019), 355–366.

[51] Richard Yalch and Eric Spangenberg. 1990. Effects of store music on shopping behavior. *Journal of Consumer Marketing* 7, 2 (1990), 55–63.

[52] Yi-Hsuan Yang and Homer H Chen. 2010. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2010), 762–774.

[53] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*. 381–388.

[54] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*. 308–314.

[55] 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 *ACM Multimedia*. 47–56.

[56] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*. 3550–3558.