AI Color Organ: Piano Music Visualization using Onset Detection and HistoGAN

Shu Xu[1, *]

[1]Shanghai United International School, Shanghai, China

*Corresponding author: shu.jack.xu@wy.suis.com.cn

Abstract. The music visualization algorithm described in this study allows users to construct piano audio files using imported image files. This paper contributes to previous studies and designs of sonification by highlighting the effectiveness of utilizing onset detection in creating intuitive sonic changes. The audio-visual correspondences employed in this study could be expanded to many other syntheses and sample manipulation techniques. Translating visual information into sonic changes could yield many creative applications in music production, as it offers musicians a simultaneously optical and auditory production experience. This approach to audio manipulation also increases the unpredictability of the sound output, which could be appealing to experimental musicians seeking to control sounds with the visual structure of artworks that they enjoy, as opposed to precise parameters. It is looking forward to seeing creative implementations of the techniques in audio-visual artworks, music production tools, and interactive multimedia systems. These results shed light on guiding further exploration of AI composing.

Keywords: Onset Detection; ReHistoGAN; Music Visualization; Color Organ

1. Introduction

Coined by Roger Fry in 1912, "visual music" was a term that originally describes the pictorial translation of music [1]. Since then, more than a century of technological advancement and artistic exploration has bestowed new meanings to the concept of visual music. Nowadays, visual music presents itself in films, computer graphics, and countless other media systems. In addition, musical information is not only transformed into visuals but also vice-versa [2].

The approaches explored in the research of visual music are very expansive, one prominent interest was in the correlation between colors and sound characteristics. For instance, Isaac Newton, Louis Bertrand Castel, and many others maintained that there is a real analogy between elementary colors and the notes of the musical scale. Newton, e.g., named seven supposedly primary colors of the spectrum–red, orange, yellow, green, blue, indigo, and violet–one to parallel each note on the musical scale. While Castel incorporated his own scheme into his color organ: blue for do, green for re, yellow for mi, red for sol, etc., seen a sketch in Fig. 1 [3]. The color organ that Castel has pioneered is one of the earliest instruments designed to visualize music. Musicians compose with organs that have colors mapped to sound according to their pitch, amplitude, and timbre. As the performer plays the keyboard, corresponding color values will be displayed on a screen [4].

As one of the earliest forms of music visualization, the color organ has slowly embodied itself to become a tradition in representing music in the visual medium. The concept has been implemented through various mediums such as the ocular harpsichord of Johann Gottlob Krüger in 1743; the pipe organs of Bainbridge Bishop in 1877; the Lumigraph of Oskar Fischinger in 1940s, and the Virtual Reality Color Organ of Jack Ox in 2000.

With the current proliferation of generative AI Artworks, it is intuitive to marry Artificial Intelligence with the concept of color organs to create new forms of music visualizations. Past attempts at making audiovisual content using deep learning include "Deep Music Visualizer" in 2019 and "Lucid Sonic Dreams" in 2021. The former algorithm was developed by Matt Siegelman, which syncs pitch, volume, and tempo features of the audio input with the generated class and noise vector fed into the BigGAN model. The latter package, created by Mikael Alafriz, took a similar approach. The algorithm allows users to generate GAN-based music visualizations by extracting amplitude and pitch information from the input audio and manipulating the input vector according to amplitude

changes and feeding StyleGAN2 a vector containing 512 numbers, which determines the output image [5]. Visually "Lucid Sonic Dreams" is more effective at establishing audio to visual correspondences. This is primarily due to the separation of percussive and harmonic elements of the input audio. By mapping the amplitude of percussive audio elements to a "pulse" parameter, the generated GAN images could pulsate according to the rhythm of the music. One commonality between the two projects is their usage of audio features as drivers of noise and class vectors. This approach shall serve as the foundation for the AI Color Organ.

Infusing the Color Organ concept would allow us to explore the correlations between sonic features and GAN-based visuals in meaningful ways. As mentioned earlier, previous deep music visualizers focused on using amplitude and pitch as driving data. However, there are many more feature extraction techniques that remained unexplored in these implementations. The algorithm discussed in this paper explored three more techniques, onset detection, tempo detection, and RMS. Moreover, the usage of pitch information in previous implementations only have a loose connection with the generated visuals. This is because in both "De chromagram ep Music Visualizer" and "Lucid Sonic Dreams", pitch information is translated into chromagrams, with the 12 tones mapped to 12 random classes of the GAN. The resulting effect is seemingly random changes in class output from a set style. In this paper, the chromagram would be bestowed more meaningful correlations with the generated visuals by mapping each note to a corresponding color output according to Newton's color scale. The rest part of the paper is organized as follows. Section 2 will describe the audio analysis and video generation components of the algorithm. Subsequently, Section 3 will demonstrate the results and discussion based on the algorithms. Afterwards, the Sec. 4 will present the limitations of the current study as well as future prospects accordingly. Eventually, a brief summary will be given in Sec. 5.

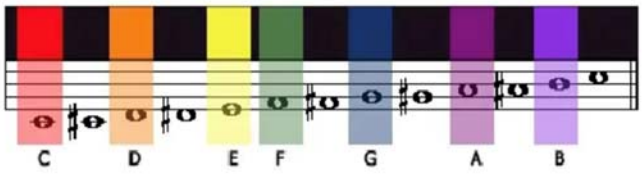

Fig. 1 Newton's color scale

2. Algorithms

To mimic the color organ created by Castel, the AI model has to be able to first, evolve according to the playing of each new note. Second, generate colored images according to newton's color scale. The full implementation of the AI color organ concept consisted of two parts, the audio feature extraction, and the video generation process.

2.1 Audio Feature Extraction

The primary tool for audio analysis that was utilized by the algorithm is the librosa python package. In this case, three types of audio information were retrieved: Note onset, chromagram, and RMS. Note onset is detected using the librosa.onset.onset_detect function. By specifying the audio time series, sampling rate, and hop length, which returns an array detailing the occurrence of each new note in milliseconds. The program utilizes librosa's implementation of the superflux onset algorithm which effectively filters out false positives induced by vibratos. The resulting array obtained through this process is then converted from milliseconds into corresponding frame counts. At frames where onsets are detected, noise vectors would be manipulated, creating a morphing effect that transforms the GAN-generated images in latent space [7]. An example of the onset detection is given in Fig. 2.

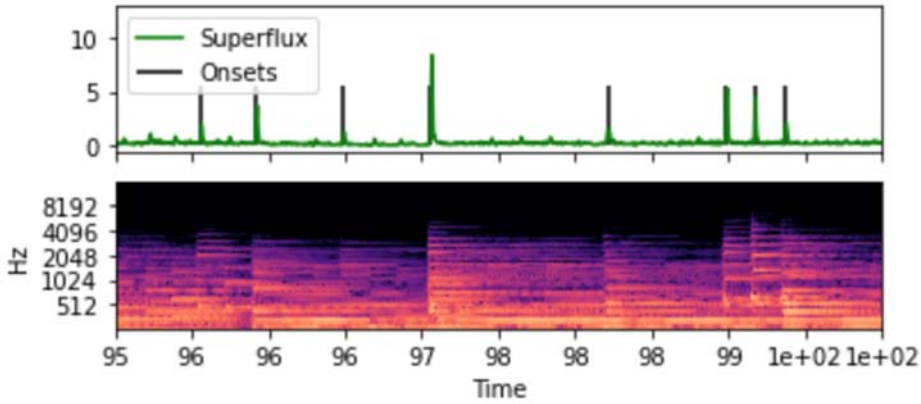

Fig. 2 Onset detection on the piano composition.

As for the color of the images generated, the primary goal as stated at the beginning is for the color of the generated imagery to correspond to the pitch of the note input. This meant that the algorithm would need to first identify the pitch of each new note. This task is achieved by feeding audio into the librosa.feature.chroma_cqt function [7]. By passing in the audio time series and sample rate, a 2D array representing the chromagram is thus created as depicted in Fig. 3.

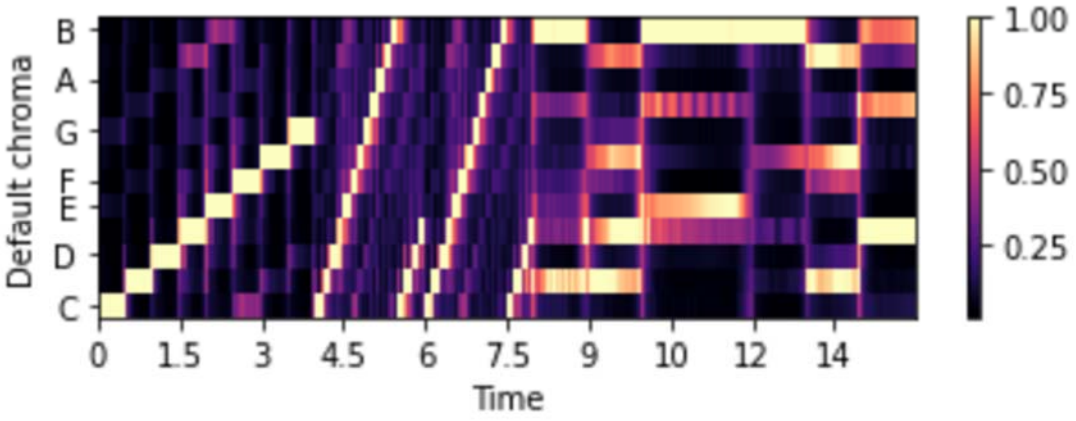

Fig. 3 Visualization of the resulting chromagram

One issue that comes with this approach is the polyphonic nature of piano pieces. The presence of chords and melodies raises the issue of prioritizing which color to display. The approach that this study took was to first apply a low-pass filter to diminish the chord voices (generally lower pitched) and enhance the melody (generally higher pitched). The high-passed audio signal is then converted to a chromagram according to the aforementioned process. For each frame, the most maximum pitch value is selected, thus determining the color of the final image.

The RMS values serve as a representation of the "energy" of the input audio as illsutratetd in Fig. 4. Since a piano piece is usually playing in expressive velocities, RMS data gives a useful representation of the sound amplitude. The value is extracted using librosa.feature.rms with long frame size. In this way, the change in amplitude is "smooth out", which is useful for creating gradual visual changes [7].

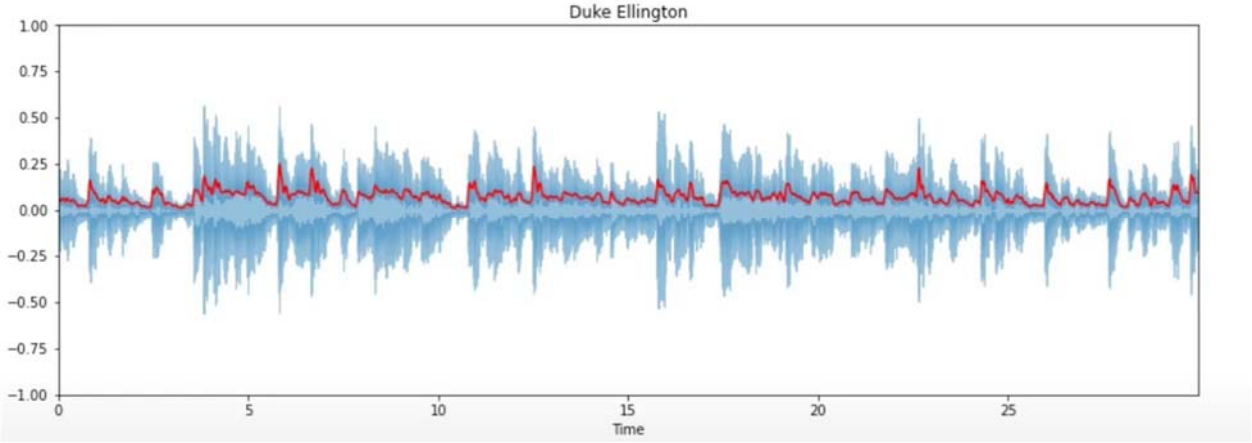

Fig. 4 RMS data for the song “Duke Ellington”.

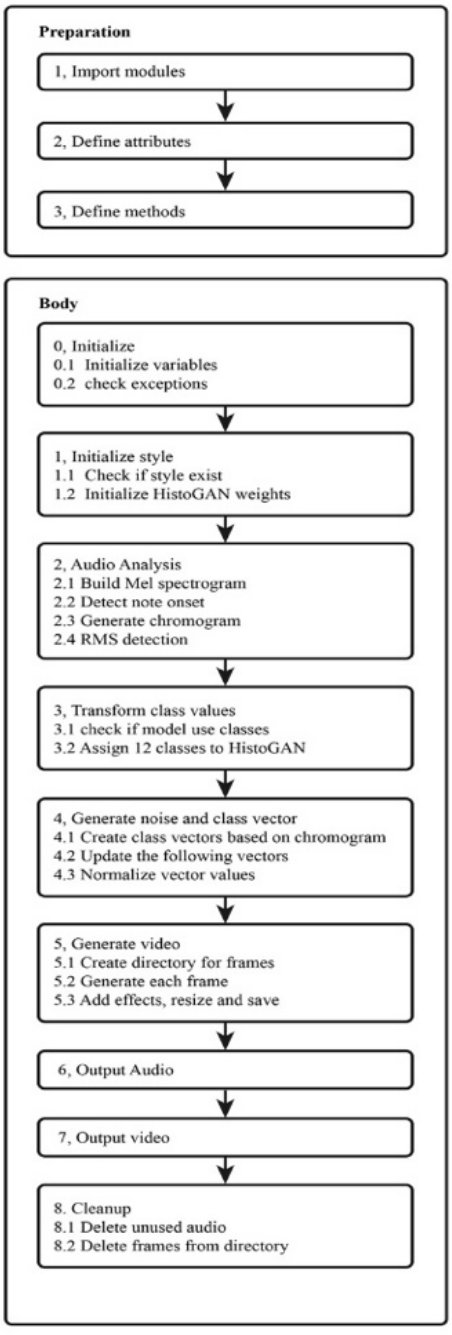

Fig. 5 The AI Color Organ algorithm.

2.2 Video-Generation & Algorithm Logic

First, input vectors are initialized and interpolated. This serves as the video's "base motion". The parameter speed_fpm controls how fast this motion goes, where "FPM" stands for "Frames Per Minute". Essentially, the number of vectors initialized per minute. For each succeeding frame, the onset variable controls the interpolation of each successive visual, chroma determines the color histogram that is fed into HistoGAN, and RMS parameters control the color contrast of the video output.

HistoGAN offers a histogram-based method for controlling the color of GAN-generated images. In comparison to traditional color transfer methods [8]. Note data interpreted by the chromagram is translated to Newton's color scale. Color histograms correspond to each of the colors within the scale. The ReHistoGAN is added to the program as a part of post-processing. Holistically, the logic of the algorithm is presented in Fig. 5.

3. Results & Discussion

The resulting algorithm learns to map color information, represented by the target color histogram, to an output image's colors with a realism consideration in the recolored image. Maintaining realistic results is achieved by learning proper matching between the target colors and the input image's semantic objects.

In combination with audio analysis, the program allows real-time composition and performance of audio-visual parameters from a single interface. The interface also enables the user to redefine the macro-level audio-visual effect associations described in this paper in a modular fashion. It allows the user to define their own audiovisual relationships that share either similar or dissimilar characteristics. I have suggested ways in which these mappings may be related in a way that make artistic sense as the qualities of an auditory or visual process may share similar 'perceptual qualities'. The instrument allows for the manipulation of visual grains via granular synthesis, a technique previously found exclusively in audio synthesis software.

4. Limitations & Prospect

Nevertheless, it should be noted that this study has some shortcomings and drawbacks. One limitation of the artworks generated using an onset detection approach is in processing fast-tempo music. Since each note onset occurs within a short interval, StyleGAN 2 has to interpolate noise vectors between two images rapidly, thus causing visual clutter. Thus, it is advisable to use piano pieces that are under 150bpm. Secondly, the algorithm employed an emphasis filter as a means of isolating the melody from the chords of a piano piece. This assumes that chord progressions tend to be of lower frequencies. As such, piano pieces that plays at higher octaves the convolute the onset detection.

In the future, it would be fruitful to explore the correspondence between higher-level features. As Dannenberg pointed out, visual music programs have the tendency to "draw connections between music and image using superficial parameters such as instantaneous amplitude or pitch" [9]. Techniques such as emotion detection and segmentation by contrast could be explored [10, 11]. These methods could help extract deep, emotional, structural, and hidden information from audio.

5. Conclusion

In conclusion, an implementation of the GAN-based Color Organ is proposed. Specifically, the audio-visual correspondences between note onset and GAN-generated image is explored. In addition, the application, the application HistoGAN allows the algorithm to establish connections between pitch and color. By mapping the chromagram to Newton's color scale, this project successfully

intertwined the tradition of Color Organs within a GAN-based music visualizer. Overall, these results offer a guideline for constructing music according to visualization based on AI techniques.

References

[1] Rekveld, Joost. The Origin Of The Term Visual Music – Light Matters. 16 Mar. 2013, Retrieved from: http://www.joostrekveld.net/?p=1105.

[2] Reykjavik Center for Visual Music. About RCVM. Retrieved from: http://www.rcvm.is/. Accessed 7 Aug. 2022.

[3] Marks Lawrence. The Unity of the Senses: Interrelations Among the Modalities. 1978.

[4] McDonnell Maura. Visual Music. 2007.

[5] Alafriz, Mikael, editor. "Introducing 'Lucid Sonic Dreams': Sync GAN Art To Music With A Few Lines Of Python Code!" *Medium* , 14 Mar. 2021, Retrieved from: https://towardsdatascience.com/introducing-lucid-sonic-dreams-sync-gan-art-to-music-with-a-few-lines-of-python-code-b04f88722de1.

[6] Batty J, Horn K, Greuter S. Audiovisual granular synthesis: micro relationships between sound and image. Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death. 2013: 1-7.

[7] McFee B, Raffel C, Liang D, et al. librosa: Audio and music signal analysis in python. Proceedings of the 14th python in science conference. 2015, 8: 18-25.

[8] Afifi M, Brubaker M A, Brown M S. Histogan: Controlling colors of gan-generated and real images via color histograms. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7941-7950.

[9] Dannenberg R B. Interactive visual music: a personal perspective. Computer Music Journal, 2005, 29(4): 25-35.

[10] Thaut M. Rhythm, music, and the brain: Scientific foundations and clinical applications. Routledge, 2013.

[11] Weihs C, Jannach D, Vatolkin I, et al. Music data analysis: Foundations and applications. Chapman and Hall/CRC, 2016.