Video background music generation: Dataset, method and evaluation

Abstract

Generating background music manually for videos is a difficult task. To ease the burden, a method to automatically generate video background music is proposed. In the paper, the dataset SymMV, the method V-MusProd, and the evaluation metric VMCP are proposed.

Keywords: Video background music generation

1 Introduction

With the rapid growth of social platforms, the needs to find suitable music for videos extend from professional fields, e.g., soundtrack production in the film industry, to amateur usages like video blogs and TikTok short videos. However, finding proper music for videos and making alignments are difficult. Thus, automatically generating background music for videos is of great value to a wide range of creators.

In the paper, a novel video and symbolic music dataset named Symbolic Music Videos (SymMV) is introduced. It contains piano covers of popular music with their official music videos carefully collected from the Internet.

The video background music generation model proposed in the paper is V-MusProd. V-MusProd decouples music generation into three progressive transformer stages: chord, melody, and accompaniment. It first predicts a chord sequence, then generates melody conditioned on chords and finally generates accompaniment conditioned on chords and melody.

Finally, a new evaluation metric, named Video-Music CLIP Precision (VMCP) is proposed, which extends the vision-language CLIP model to video and music domain to measure the video-music correspondence. However, it will not be addressed in this report.

2 Related works

A previous work, which won the best paper award of ACM MM 2021, is worthy to be mentioned:

Video background music generation with controllable music transformer

The paper [2] is the first work to address the problem of video background music generation. The CMT method considers the the music-video rhythmic consistency. However, it ignores the symantic-level correspondence between the video and the music, which can sometimes lead to conflicting styles.

The method V-MusProd is an improvement of CMT by taking semantic information into consideration. However, the model of V-MusProd is not based on the model of CMT; the two models, in fact, almost shares nothing similar.

3 Prerequisites

Music generation is a branch of AI-Generated Content (AIGC). As the representation of music is different from images and 3D models, the field is not familiar to those with computer vision or computer graphics background. Therefore, this section is dedicated to introduce some fundamental knowledge related to music generation, otherwise those concepts would not be easily understood.

3.1 Audio vs Symbolic Music

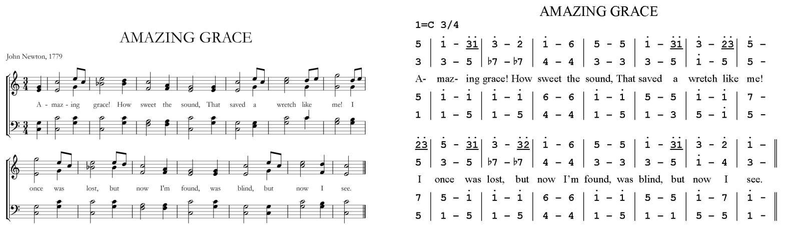

Figure 1. An example of symbolic music. Amazing Grace, written using staff notation and numbered notation. Images via Wikipedia.

In physics, sound is a vibration that propagates as an acoustic wave through a transmission medium. If the wave is sampled at a particular rate, then the sound wave can be stored and processed by computer. Many AI models have been proposed to generate audio contents.

Symbolic music, on the contrary, stores the music as a sequence of symbols. An example of this is sheet music, including staff notation and numbered notation (see Figure 1), which uses symbols to represent pitches, rhythms and chords. The symbols accurately represent the music and guide the performers to perform the music. In computer, MIDI files can be regared as a kind of symbolic music, which will be introduced in the following parts of the report. In terms of AIGC, if we regard the symbols as tokens, then the task of symbolic music generation is pretty similar to the task of text generation. The method introduced in the paper follows this path.

3.2 Music Theory Fundamentals

In this section, the concept of note, bar and chord will be introduced, which are fundamental concepts in understanding the structure of music.

A note is an isolated sound used as an atomic building block for creating music, which has the following properties:

- Note value: the note's duration relative to the musical meter. Notes are subdivided in halves, so there are whole notes, half notes, quarter notes, eighth notes, etc. A quarter note is also referred

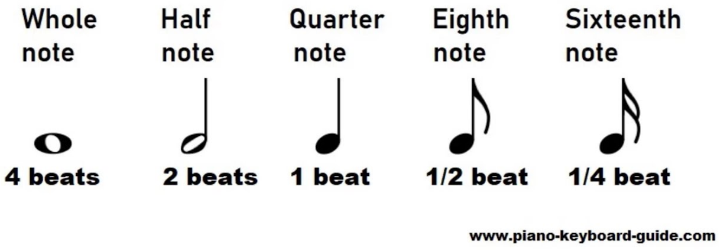

Figure 2. Note values in staff notation. Image via Piano-Keyboard-Guide.

to as a beat in this report. Note that the duration is not measured is absolute time, e.g., in milliseconds, but measured according to the tempo of the music, i.e., the number of beats per mimute. The staff notation of note values is illustrated in Figure 2.

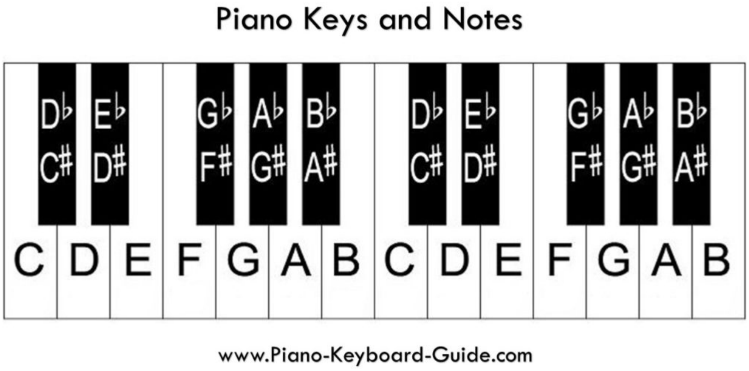

Figure 3. Piano keys and their correspoinding notes. Image via Piano-Keyboard-Guide.

- Note scale: the note's pitch, which can be quantified as a frequency, so that a note can be percieved by the listener as either "higher" or "lower". An octave-repeating scale can be represented as a circular arrangement of pitch classes. For instance, the increasing C major scale is C–D–E–F–G–A–B–[C]. In terms of piano, the correspondence between scales and keys is illustrated in Figure 3.

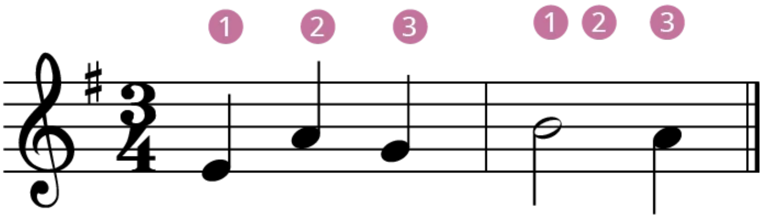

Figure 4. Time signature 3/4. There are 3 quarter notes in the first bar, and 1 half note and 1 quater note in the second bar. Image via Jade Bultitude.

A bar, or a measure, is a segment of music bounded by vertical lines, indicating one or more recurring beats. A time signature specifies how many note values of a particular type are contained in each bar. For instance, the time signature 3/4 indicates there are three quarter notes in each bar, as illustrated in Figure 4.

A chord is a harmonic set of pitches consisting of multiple notes that are sounded simultaneously. The most simple chord is a triad, which consists of three notes that can be stacked vertically in thirds. For instance,

- C major, the triad consisting of notes C, E and G;
- A minor, the triad consisting of notes A, C and E.

The chords can have significant impact on emotion in music. It is generally perceived that major chords are positive (e.g., happy), while minor chords are negative (e.g., sad). This feature of chords is exploited by the model introduced in this report, where the emotional information is encoded in chords.

3.3 MIDI File Format

The MIDI file format is the most famous file format that represents symbolic music, which is the most popular among musicians who creates digital music.

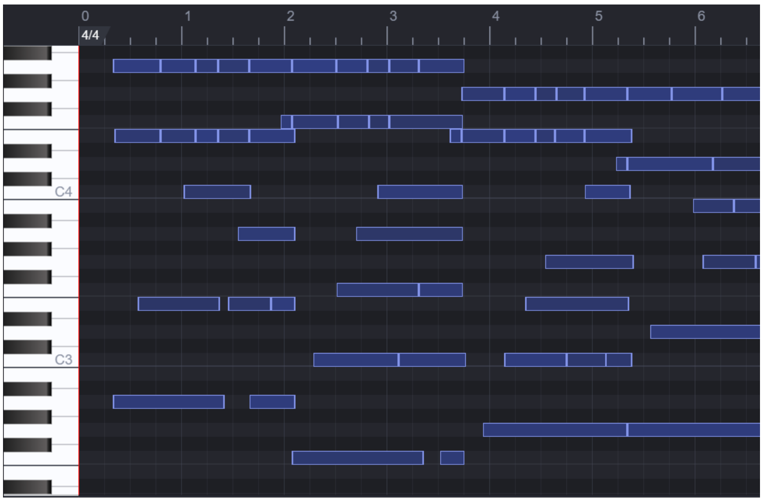

Figure 5. Visualization of a MIDI file. Each horizontal bar marks the beginning and the end of a note. Image generated by signal, an online MIDI editor.

Symbolic music is encoded in a MIDI file as a sequence of MIDI messages. Imagine we press a key on a keyboard, and, after some time duration, release the key. These two events are captured by the musical equipment and recorded in the MIDI file. A visualization of a MIDI file is recorded in Figure 5.

In this report, we only care about the following two kinds of MIDI messages:

- Note On: indicates a particular note should be played. There are two values carried by a Note On message:
  - note number, the note to play (i.e., which key to press);

  - velocity, how much force the note should be played.

  Note that a Note On message with zero velocity is effectively the same as a Note Off message.

- Note Off: indicates a particular note should be released. There are two values carried by a Note Off message:
  - note number, the note to release (i.e., which key to release);
  - velocity, how much force the note should be released.

The time information is not a constituent part of Note On/Off messages, but given in other parts of MIDI files.

4 Dataset and Method

4.1 SymMV: The Dataset

The collected SymMV dataset contains 1140 pop piano music in both MIDI and audio format with the corresponding official music video with a total duration of 76.5 hours. SymMV is split into the training set (1000 pairs), validation set (70 pairs), and test set (70 pairs). Our dataset also includes various annotated metadata, such as chord progression, tonality, and rhythm.

However, currently the complete dataset is no longer available, since

- some videos are removed from the website;
- some videos are not accessible due to copyright protection restrictions.

Therefore, only 1070 out of 1140 video clips are successfully downloaded.

4.2 V-MusProd: The Model

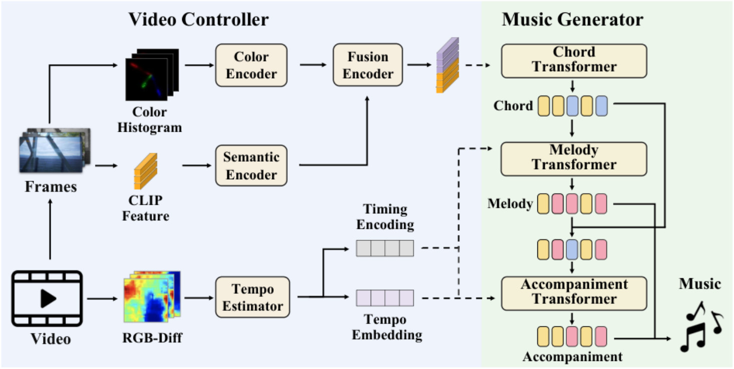

Figure 6. The pipeline of V-MusProd. Image via [1].

The following features are extracted from video clips.

- Semantic features: The pretrained CLIP2Video [3] is used as the extractor to encode raw video frames into semantic feature tokens without finetuning.

- Color Features: The color histogram of each video frame, i.e., a 2D feature map proposed in [4] is extracted, to represent the color distribution in a non-linear manifold.

  The semantic features, as well as color features, are used to generate the chord, which reflects the emotion of the generated background music.

- Motion Features: The RGB difference with intervals of 5 frames (0.2 seconds) is extracted and map the mean RGB difference of a video to the music tempo. A linear projection is used from the minimum and maximum RGB difference to the tempo range of $[90, 130]$.

The music generation process is decoupled into three steps:

- Chord Transformer: We adopt a transformer decoder architecture for Chord Transformer to learn the long-term dependency of input video feature sequences. The event-based token sequence is added with positional encoding and fed into the Chord Transformer as a query. Meanwhile, the style features from video controller are fed as keys and values.

- Melody Transformer: We employ an encoder-decoder transformer architecture for melody generation. The encoder receives a chord sequence as input, and then the decoder generates a note sequence as the output melody. Considering the relatively short-range dependency between melody and chords, we adopt a bar-level cross-attention mask so that each decoder token can only attend to the contextual encoder output within the previous current or next bar.

- Accompaniment Transformer: Similarly, we also adopt an encoder-decoder transformer to generate the accompaniment sequence. Since accompaniment closely correlates with chords and melody, we merge the generated chord sequence with the melody and then pass the merged sequence to Accompaniment Transformer as conditional input. We also apply the same bar-level cross-attention mask as in Melody Transformer. Eventually, the generated accompaniment is directly merged with the melody to form the final music.

Why decoupling?

When writing a piece of music, a musician first choose a chord progression, which is a series of chords played one after another. The chord progression conveys the emotion of the music. Then the musician tries to write a melody over the chords, and then the accompaniment. By decoupling into three steps in the model, we are imitating the way a musician writes a piece of music, and thus achieve better performance.

5 Implementation details

So far the source code of V-MusProd is not yet available.

The impementation is incomplete. Only the chord transformer is implemented, thus only producing a sequence of chords, instead music.

First, a script is written to download the videos listed in the dataset from Youtube. As mentioned before, only 1070 out of 1140 video clips are successfully downloaded.

Then the source code of [3] is used to generate the video semantic embedding of video frames. The algorithm to generate log-chroma space image, using the inverse-quadratic kernel, as in [4], is written completely from scratch.

The package “fast transformers” is employed to implement the model.

The hyperparameters are the same as those listed in the appendix.

6 Results and analysis

As the implementation is incomplete, the generated music cannot be shown here.

Below are the first 10 chords of the chord sequence of Avicii’s Heaven (https://www.youtube.com/watch?v=fqzhtvLWefA), which is generated by the chord generator:

```
880 2200 F#m7
2200 2640 AM
2640 3300 F#m7
3300 3520 DM
3520 4400 DM7
4400 4840 F#m7
4840 5060 AM
5060 5280 DM
5280 6600 DM7
6600 7040 AM
```

7 Conclusion and future work

In the report, the video background music generation model V-MusProd is introduced. Only the chord generator is implemented; the melody and accompany generator is not yet implemented. In the future, the above two generators will be implemented.

References

[1] Zhuo, L., Wang, Z., Wang, B., Liao, Y., Bao, C., Peng, S., ...& Liu, S. (2023). Video background music generation: Dataset, method and evaluation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15637-15647).

[2] Di, S., Jiang, Z., Liu, S., Wang, Z., Zhu, L., He, Z., ...& Yan, S. (2021, October). Video background music generation with controllable music transformer. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 2037-2045).

[3] Fang, H., Xiong, P., Xu, L., & Chen, Y. (2021). Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097.

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

[5] Uitdenbogerd, A., & Zobel, J. (1999, October). Melodic matching techniques for large music databases. In Proceedings of the seventh ACM international conference on Multimedia (Part 1) (pp. 57-66).