Definition & Method

2019年8月8日 9:52

What? BGM automatic generation based on …

- … MusicBERT: A joint model for video & music representation learning
    - -> Task 1: next sentence prediction
    - -> Task 2: masked bar prediction
        - -> Pattern analysis of music make it possible to feed participles to BERT
        - -> Some predictable attributes are just like those in language domain
        - -> Evaluation for predictable parts?
- … hirearchical latent vector model
    - -> trained a set of encoder-decoder based on symbolic latent vector
    - -> is the mapping from video to latent code available?
- … Multi-task model
    - -> an encoder of dense captioning & a decoder of style transfer

Input

Short Video & Homemade Video (Different from movies BGM: sound effect)
Trained with music videos dataset, trying to be general

Output:

Polyphonic Music Generated
(To simplify our task, output SHOULD be on symbolic domain,
Symbolic output could be tranferred into synthesized music afterwards)

Related Work

2019年8月8日 10:04

Separate Music Generation:

Based on VAE or GAN, and concerning with only symbolic domain:

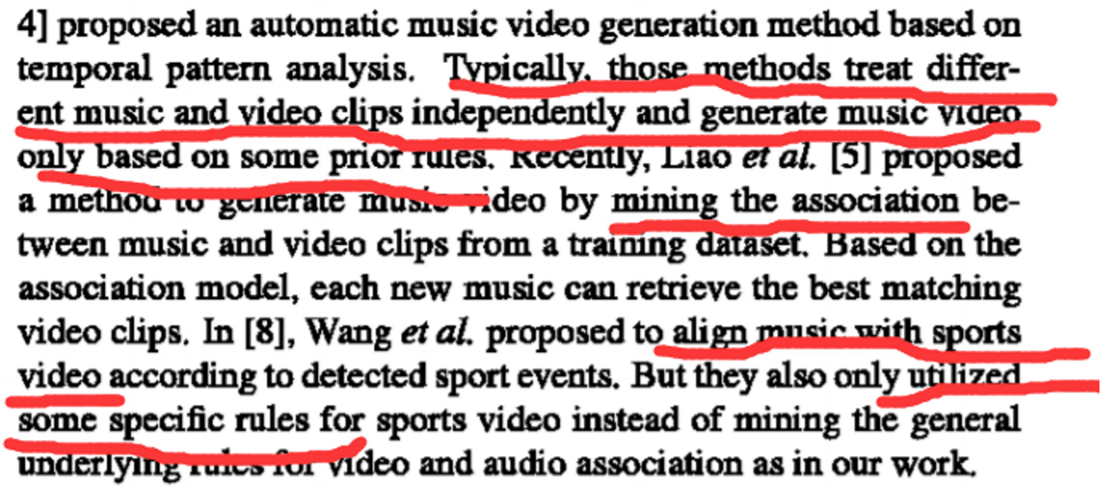
- MelodyRNN (Waite et al. 2016)
- DeepBach (Hadjeres, Pachet, and Nielsen 2017)
- Symbolic domain -> audio performance: like PerformanceRNN (Simon and Oore 2017)

Association:

- Music Composition: based on specific rules
- Music Video Generation

Retrieve & Alignment/Synchronization & maybe some trasition stuff

1. <Mining Association Patterns between Music and Video Clips in Professional MTV>
   U: color and structure tensorhistogram of video clips
   V: features extrcted from time domain or spectral domain
   Using U -> V to learn H space, a latent semantic aspects
2. <Auto-generation of Professional Background Music for Home-made Videos>



4] proposed an automatic music video generation method based on temporal pattern analysis. Typically, those methods treat different music and video clips independently and generate music video only based on some prior rules. Recently, Liao *et al.* [5] proposed a method to generate music video by mining the association between music and video clips from a training dataset. Based on the association model, each new music can retrieve the best matching video clips. In [8], Wang *et al.* proposed to align music with sports video according to detected sport events. But they also only utilized some specific rules for sports video instead of mining the general underlying rules for video and audio association as in our work.

Segmentation:

according to shot detections in video (two levels segmentation in music, self-similarity which computing structure patterns for the second level)

Video Representation:

color, motion features

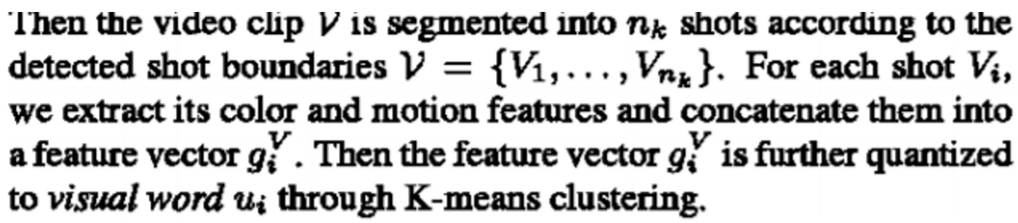
Music Representation:

the second-level segmentation could be modeled with Markov chain

Association Model: GMM

Generation: selection & transition

LOW level & HIGH interpretability



Then the video clip V is segmented into n_k shots according to the detected shot boundaries $\mathcal{V} = \{V_1, \ldots, V_{n_k}\}$. For each shot V_i, we extract its color and motion features and concatenate them into a feature vector g_i^V. Then the feature vector g_i^V is further quantized to *visual word* u_i through K-means clustering.

More learning-based music generation philosophy, and mine the underlying professional knowledge & rules.

High level semantic and scene analysis for automatic generation is needed, also we need to investigate underlying knowledge and rules on higher level.

3. <Automatic Realistic Music Video Generation from Segments of Youtube Videos>
   A reverse task.
4. <PerformanceNet: Score-to-Audio Music Generation with Multi-Band Convolutional Residual Network>
   Score-to-audio works could help us simplify our task into a symbolic level.

Others:
<MorpheuS: generating structured music with constrained patterns and tension>
Generation based on learning music structure patterns & tension model