Generating Audio for Muted Piano-Playing Video Using Deep Learning

Project Proposal

XIA Junzhe
20493411
jxiaaf@connect.ust.hk

YANG Baichen
20493198
byangak@connect.ust.hk

HUANG Zeyu
20493631
zhuangbi@connect.ust.hk

March 9, 2019

It is noticed that human playing the piano effectively generates a series of well-patterned actions, i.e. the position of hands and the keys and the depth of keys being pressed down. The notes that the piano generates strongly obey to this visual pattern. Hence, it becomes reasonable to recognize the visual patterns from piano-playing videos and reproduce the instrumental sounds using machine learning.

Some related work has been done by scholars in this field [1] [2] [3]. Akbari et al. [4] proposed a system of automatically annotating piano-playing video by tackling issues such as recognizing the keyboard and detecting the pressed keys according to some artificially formulated rules related to the visual nature of the piano keyboard. Although a high accuracy has been achieved, it is hard to prove those hard-coded rules can apply to most cases.

Our objective is to adapt some deep learning techniques related to object detection into this task and build a more robust model to detect what keys are being pressed in the video as well as the velocity of every single key press, which will enable us to retrieve the notes which are sounded. After the note sequence is obtained, we will use some software instrument to reproduce the piano sound and remix it with the original video. Apart from the basic task just mentioned, we also hope to make this model support real-time audio generating, which may require us using some faster network architectures like Faster R-CNN and cope with the synchronization between the detected notes and the audio to be generated.

The dataset we are going to use partially comes from the previous work [2], and we also plan to label some videos manually which may enrich the materials related to this research topic. The original videos will be fetched from Youtube 8m video dataset [4].

We will evaluate our result in two aspects. During the training, we will mainly focus on mathematical accuracy, namely, comparing the notes the model detected with the original ones. Specifically, the performance metric is going to be a customized loss function that describes the difference between the original note sequence and the one that the model generated. The expected result should be a small variance, which represents an ideal similarity between the audios. After the prototype of the model is obtained, we will also introduce psychological evaluation to assess the quality of generated audio, and then fine-tune the model.

References

[1] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., & Freeman, W. T. (2016).Visually indicated sounds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2405-2413).

[2] Akbari, M., Liang, J. & Cheng, H. Multimed Tools Appl (2018) 77: 25513. https://doi.org/10.1007/s11042-018-5803-1

[3] Akbari, Mohammad & Cheng, Howard. (2015). Real-Time Piano Music Transcription Based on Computer Vision. IEEE Transactions on Multimedia. 17. 1-1.10.1109/TMM.2015.2473702.

[4] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675.