Generating Audio for Muted Piano-Playing Video Using Deep Learning

Project Milestone

XIA Junzhe
20493411
jxiaaf@connect.ust.hk

YANG Baichen
20493198
byangak@connect.ust.hk

HUANG Zeyu
20493631
zhuangbi@connect.ust.hk

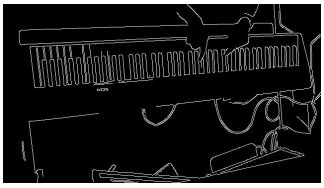
1. Introduction

1.1. Improved Keyboard Detection

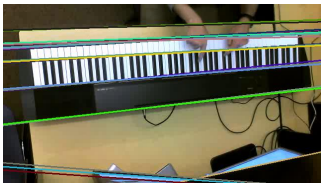
Among all of the previous work that we have studied, no deep learning based method is used to localize the keyboard in video frame. Instead, all of them used traditional CV method such as Hough Line Transformation, brightness comparison to determine the coordinates of four corners of the piano keyboard. We have first partially reproduced this algorithm.

(a) Original Image in Grayscale

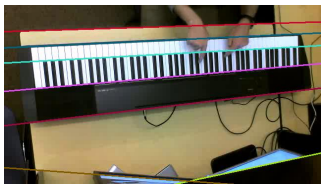
(b) Gaussian Blurred Image



(c) Canny Edge Detection



(d) Hough Line Transformation



(e) Prune Repeated Lines

Figure 1: Pipeline of Finding Boundaries of Keyboard

After pruning repeated lines, some hard-coded method will be used to finally determine the positions of four corners. However, after applying this method to multiple images with different light condition and hand positioning, we found following drawbacks:

- To achieve a satisfying performance, several hyperparameters in Hough Line Transformation and Canny Edge Detection needs to be adjusted dramatically in different images.
- To accurately determine the corners' coordinates, it is required there is no hands posing above the keyboard. It is hard to guarantee since most of piano videos do not contain such a "empty" frame. In some cases, the camera angle will even change slightly, where the traditional method will fail.

Due to the weak robustness of the traditional CV method, we decided to use a convolutional neural network to localize coordinates of four courners of the keyboard. However, we do not have any existing dataset for this task. So we labelled a number of images captured from different piano playing videos, which come from the dataset of a previous work, as well as some videos on the Internet.

Figure 2: Labelling Tool

To augment this dataset, we will also consider to artificially put keyboard image into some random backgrounds and perform some basic transformation such as flipping and stretching on images present.

Parallel to augmentation of the dataset, we have also started to evaluate the performance of several different CNNs on this task.

2. Future Work

2.1. Finalize Keyboard Localization Network

After experiments and evaluations on different networks, we will be able to decide which CNN that we will use for keyboard localization.

2.2. Keyboard Separation Algorithm

After the keyboard can be correctly recognized, we will use perspective transformation to normalize the keyboard in video frame into a preset rectangle structure. Then we will work on an algorithm to extract boundaries of each key, namely, further separate the keyboard into individual keys on this.

2.3. Train Pressed Key Detection Network

When individual keys can be localized and normalized, it is feasible for us to perform deep learning on each key to determine their conditions (pressed or not pressed). The quality of the eventual generated music heavily depends on the performace of this network.

2.4. Velocity Detection Network

A music piece without emotional variation is rigid and undesired. To detect the velocity information of a key pressing event, we proposed to use a recurrent network to analysis relative frames in order to derive an approximate velocity value of this pressing.

2.5. Audio Generation

Upon we can derive all notes information, we need to reconstruct them as a MIDI file. During this process, we may need to smooth out some noisy notes such as those with exceptionally short duration. Besides, it may be necessary to normalize the velocity information we retrieved to avoid unexpected dynamic change.

References