

# 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

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.

### 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 computer vision methods 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.

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 hyper-parameters 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 corners of the keyboard.

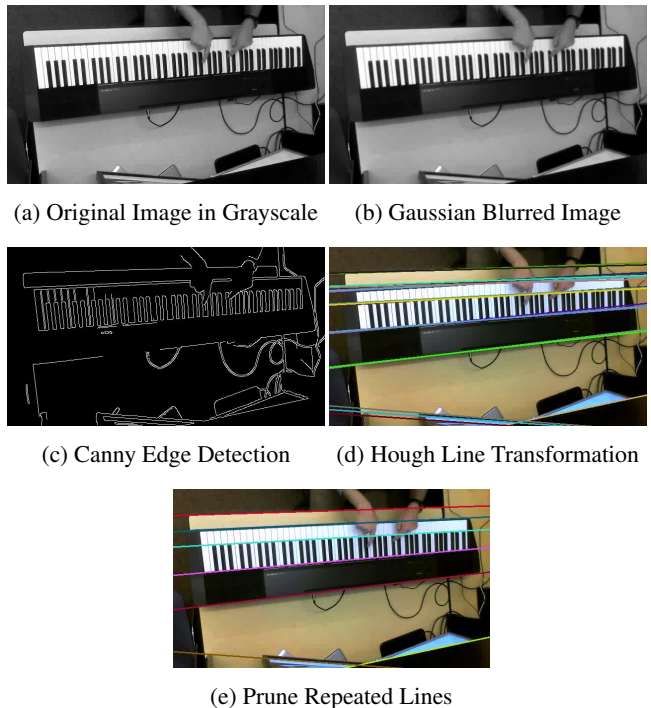


Figure 1: Pipeline of Finding Boundaries of Keyboard

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 [1], as well as some videos on the Internet.

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.

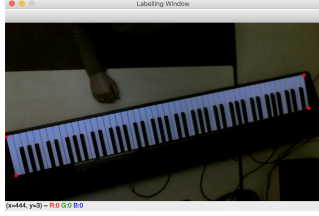


Figure 2: Labelling Tool

## 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 performance 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.

In this project, we aims to achieve hand gesture detection in piano playing and generate the correponding audio afterwards. It is also expected that we can thereby generate piano sound by our model without a real sound-producing piano, but only with a “fake” one.

## 3. Problem Statement

In this section we will mainly discuss our problem formulation, dataset sources, the evaluation pattern and expected results.

### 3.1. Objective Formulation

### 3.2. Dataset

The current dataset we have is a external dataset from previous peoples research [1]. It consists of several muted videos and their corresponding midi files. The major problem is that the audio and the video do not start and end at the same time, and the offset is not provided. Hence, we have to manually label the time offset from the beginning of the video to the beginning of the audio. This task is accomplished.

While we are labeling the data, we have found several traits of the dataset:

- There are several pianists, each with their own playing style.
- All keys on the keyboard are covered.
- They have included some fake gestures.
- The playing times vary.

But there are several shortcomings as well:

- The fake gestures are almost identical, which is moving the palms above the keyboard horizontally back and forth after finishing playing. We doubt the ability of eliminate all false positives including those more subtle ones occurring while playing.
- A large portion of the dataset is pressing the keys from left to right one by one using only one finger. This playing pattern sheds less shadow and cause less interference than playing in the real world. The low speed also reduces the difficulty to learn this specific pattern. In short, we worry about the occurrence of an overfit on simple playing patterns.
- Several audio files are corrupted. Either they are empty, or they have midi messages with messed up time informations. Whats more, the leftmost key and the rightmost two keys are broken and are not recorded. One particular key somewhere in the middle is presumably broken as well.
- Only one or two from the 14 players plays meaningful pieces instead of random fiddling. We think more meaningful pieces should be included, since playing some common chords may generate common and critical graphic information that is uncommon among piano newbies keyboard-scrubbing.

Therefore, we believe that it is helpful to add our own dataset, where, beside solving these problems, we want to provide the following enhancements

- Provide the keyboard localization data. We want to localize the keyboard using neural network instead of traditional CV method.
- Provide the strength information while pressing the key. Then we can generate more emotional music.

We plan to augment the dataset after the milestone, in parallel with the training process.

### **3.3. Evaluation and Expected Result**

## **4. Deep Learning Approach**

In this section we will propose a deep learning approach towards the above problems.

## **5. Preliminary Results**

Currently we've pre-processed the dataset and generated some preliminary labelling result by our neural network.

## **6. Future Work**

In the future, there are plenty of work to be done.  
Firstly, enlarging and augmenting our dataset.  
Secondly, refine our network structure.  
Thirdly, conduct evaluation on the result.

## **References**

- [1] M. Akbari, J. Liang, and H. Cheng. A real-time system for on-line learning-based visual transcription of piano music. *Multimedia Tools Appl.*, 77(19):25513–25535, Oct. 2018.