# Optical Music Recognition for MIDI-LilyPond File Generation

Evan Matthews
evanmm3@illinois.edu
University of Illinois Urbana-Champaign

## ABSTRACT

TODO: Abstract

**ACM Reference Format:**
Evan Matthews. 2024. Optical Music Recognition for MIDI-LilyPond File Generation. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

TODO: Introduction

In the world of music or audio transcription/arrangement, the complexity of sound and score data has allowed human performance to remain the current state-of-the-art. Several factors contribute to this circumstance: the number of potential audio and score file types, ambiguity on how performance qualities are notated, and overall inconsistencies between recordings and their respective scores. Machine learning models, in turn, have been a crucial step towards reducing these inconsistencies. Their ability to learn the nonlinearities and artistic qualities that otherwise plague audio computation have allowed for noteworthy advancements such as the WaveNet[**?** ]. The remaining issue in the process of sound generation is the amount of data available to train with. Worthwhile audio data remains difficult to collect due to its large size and potential copyright issues. In particular, trying to condition off another medium is incredibly difficult as current datasets lack the correlations necessary to streamline the audio transcription process. Datasets such as MAESTRO are close to ideal results, but some intermediate steps are needed to learn correlations between audio files and musical scores.

With the following conditions in mind, I propose an Image-to-MIDI model to serve a few purposes. First, this model will serve to convert image data to relative musical data to a high degree of accuracy. The process of converting MIDI to an image is trivial, but the opposite is a known Optical Music Recognition (OMR) problem in that the correlation is nonexistent. Second, a highly reliable model of this type will be capable of serving as an intermediate step in future audio computation endeavors, as the nontrivial nature of OMR is a barrier for reliable model training between scores and other types of audio data. Finally, a conversion of this type allows for more flexible datasets because MIDI can be converted into a number of other audio file types.

From a technical standpoint, my project will consist of: the construction of a trained model to relate images to MIDI data, a paper describing the process of constructing the model, and any notable results and accuracies, and a few demos from which the accuracy of the model can also be visually confirmed. The model will generally reference existing OMR research and will serve as an attempt to construct a completely new CNN model, (not be an implementation of an existing paper). Existing MIDI datasets have been narrowed down to strictly two-stave piano works by Johann Sebastian Bach to maintain rhythmic and compositional uniformity, although a generalized model for various instruments and number of staves can theoretically be produced given a large enough dataset.

Along with serving as my four-credit-hour project for the course, this model will serve as a supplementary step in my MS thesis- a separately-proposed model for generating audio conditioned on sheet music, advised by Paris Smaragdis. I can confirm that neither project will trivialize the other, and the flexibility of this project will open the door for other interesting ML-audio research in the future. For further results, I may pursue auto-transcription with this model to further differentiate from my planned thesis.

## 2 BACKGROUND

### 2.1 Optical Music Recognition

TODO: OMR background

### 2.2 Lilypond, MIDI

## 3 EXPERIMENT

### 3.1 Key resources

The Image-to-MIDI convolutional neural network will be built from scratch while referencing promising results from OMR and image-to-text research, (should the CNN need fine-tuning or other enhancement). Currently, the model dataset is the complete Bach Midi Index and will be narrowed down to fit additional constraints: two staves, single instrument, rhythmic quantization at the 16th or 32nd- note level. Model testing will focus primarily on self-trained work, but I am interested in comparing against pre-trained models if I have extra time. For hardware, training will take place on a Linux GPU cluster, courtesy of Paris Smaragdis and the UIUC-CS Audio Computing Lab. Programming work is using a combination of Python with Pytorch for the CNN, and Bash scripts with Lilypond to prepare and normalize the dataset.


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

# 4 RESULTS

# 5 DISCUSSION

TODO: discussion

# 6 CONCLUSIONS

TODO: conclusions