A HYBRID HMM-RNN MODEL FOR OPTICAL MUSIC RECOGNITION

Liang Chen
Indiana University Bloomington
chen348@indiana.edu

Rong Jin
Indiana University Bloomington
rongjin@indiana.edu

ABSTRACT

Optical music recognition (OMR) serves as one of the key technologies in Music Information Retrieval by mining symbolic knowledge straightforwardly from score image. A full-fledged OMR system encompasses both image recognition and music interpretation parts to convert image data to symbolic representation. However, this process proved to be remarkably challenging so the state-of-the-art OMR systems still leave much to be desired. The significant development of deep learning in recent years has brought great success to different domains, *e.g.* object recognition and scene understanding, and aroused researcher's interest to address unsolved problems with this new weapon. Shi et al. [2] introduced the first attempt of using deep learning approach to solve OMR but the model applies more appropriately to text than music scores. In this paper we propose a hybrid model that combines the power of Hidden Markov Model (HMM) and Recurrent Neural Network (RNN) for an end-to-end score recognition scheme.

1. INTRODUCTION

OMR is hard in several aspects. The symbol primitives often have simple shapes, yet there could be a diversity of configurations for them to represent different musical meanings. Further, symbols are usually correlated to each other and contribute to music semantics together, *e.g.* duration is jointly decided by note head, aug dots, and beams/flags; pitch is influenced by clef, keys, accidentals and position of note head; rhythm is established along with voice decisions. Some of these correlations are local, which can presumably be captured by convolutional operators. Some are long-term dependencies, requiring the model to have the power of detecting connection between symbols which are spatially apart.

2. HYBRID HMM-RNN MODEL

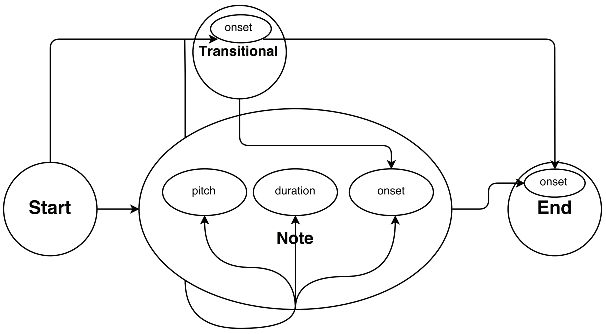


Figure 1: State transitions for music interpretation

We want to leverage the power of deep neural networks to build a holistic model that learns both local and global correlations of symbols from data. Like [1], we use Convolutional Neural Network (CNN) to extract local image features and Long-Short Term Memory (LSTM) network to capture the dependencies in the context of a whole measure. To enable sequence labeling, we slice the measure into continuous frames and use LSTM to make continuous predictions for the frames. This CNN-RNN model projects image frames to the space that accounts for music semantics such as pitch and rhythm; after projection, we interpret the result with rhythm constraints using HMM. The overall architecture is shown in Fig. 2.

Our model differs from [2] in two major aspects. First, [2] sliced the input for LSTM in feature space and implicitly used locational information from extracted features, while we split image into frames at the first step and then extract features to feed into their corresponding LSTM cells. The explicit positional information contained in ground truth has restricted the convolutional layers to directly look at frames of interest to extract effective features. Second, we jointly predict rhythm and pitch labels using a combined loss layer, but not just predict a sequence of pitches, essentially touching the multi-dimensional nature of the problem.

We define the semantic space $\mathcal{S}$ as a set of label vectors that consist of various pitches, rhythms, and voices. More specifically, we have three different categories of $s \in \mathcal{S}$ – *monophonic*: single voice and single pitch, *homophonic*: single voice and multiple pitches, *polyphonic*: multiple voices and multiple pitches. Our current exploration was primarily focused on the simplest monophonic case, but the paradigm can be naturally extended to the other two scenarios. In monophonic case, we predict duration label r and one pitch label p (both are categorical) for each frame. We assume these two labels are conditionally independent given the feature F extracted by CNN+RNN, so we connect the output of LSTM to two separate fully

Attribution: Liang Chen, Rong Jin. "A hybrid HMM-RNN model for optical music recognition", Extended abstracts for the Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference, 2016.

connected layers with weights W_r and W_p, bias b_r and b_p respectively associated to duration and pitch predictors. The probability of duration label r and pitch label p at frame t is computed using softmax: $P(r_t = r_i|F_t) = \frac{exp(W_{r_i}F_t+b_{r_i})}{\sum_{r\prime\in R} exp(W_{r\prime}F_t+b_{r\prime})}$, $P(p_t = p_i|F_t) = \frac{exp(W_{p_i}F_t+b_{p_i})}{\sum_{p\prime\in P} exp(W_{p\prime}F_t+b_{p\prime})}$.

The neural network was trained end-to-end by minimizing the negative log likelihood:

$$\begin{aligned} L_W &= -\sum_{n=1}^{N}\sum_{t=1}^{T} \log P_W(s_{n,t}|X_{n,1:t}, s_{n,1:t-1}) \\ &= -\sum_{n=1}^{N}\sum_{t=1}^{T}(\log P_W(r_{n,t}|X_{n,1:t}, r_{n,1:t-1}, p_{n,1:t-1})+ \\ &\log P_W(p_{n,t}|X_{n,1:t}, r_{n,1:t-1}, p_{n,1:t-1})) \end{aligned} \tag{1}$$

where N is the total number of measures for training and T is the number of frames in one measure.

We stacked an HMM layer over the LSTMs for music interpretation. HMM is well-suited to parse sequence of multi-dimensional states with knowledge-based constraints. We model the interpretation as searching for the best path of frame states in semantic space according to their legitimate transitions as depicted in Figure 1.

3. EXPERIMENT

3.1 Dataset

In our tentative experiment, we generate independent score measures using music21 and convert the symbolic data to images with MuseScore. To extract locations of notes and their corresponding labels we analyze the vector graph output from MuseScore and automatically annotate the horizontal range of each note as well as their pitches and durations. We then split the generated data into two separate sets: 7000 images for training and 3000 images for testing.

For each measure we apply the same $\frac{4}{4}$ time and restrict pitch from F#3 to B#5 using the same G clef. Each note was accidentaled by at most one sharp or flat.

3.2 Experimental Setting

We slice each image (one measure) into multiple frames using a frame length of 10 and hop size of 5. If the intersection between one frame and any annotated region is larger than 50% of the region, we apply the label for the note inside that region to the frame, otherwise the frame is labeled as background. We use a batch of 10 measures during training, with the learning rate set to be 0.001. The model trained after 100,000 iterations was used in test phase. We apply one convolutional layer with ReLU and max-pooling for the CNN part, and one layer of LSTM for the RNN part.

3.3 Evaluation

Given limited time, we only performed evaluations with our current net configuration and monophonic dataset. Check Tab. 1 and Fig. 3 for some initial results. Our next step will be extending the experiments to polyphonic scores and different net configurations.

Pitch	Rhythm	Pitch (w/o bkg)	Rhythm (w/o bkg)
94.82%	88.29%	75.47%	43.17%

Table 1: Per-frame accuracy of pitch and rhythm predictions by CNN-RNN, bkg means background frames.

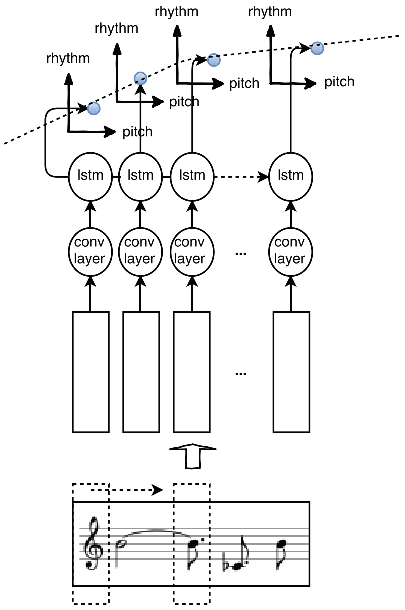
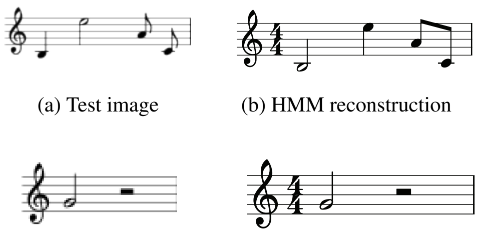


Figure 2: Overview of HMM-RNN model

4. REFERENCES

[1] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[2] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *CoRR, abs/1507.05717*, 2015.



(a) Test image

(b) HMM reconstruction

(c) Test image

(d) HMM reconstruction

Figure 3: Reconstruct original measure with HMM