A HYBRID HMM-RNN MODEL FOR OPTICAL MUSIC RECOGNITION

Liang Chen, Rong Jin, Simo Zhang, Stefan Lee, Zhenhua Chen, David Crandall
Indiana University Bloomington

ABSTRACT

Optical music recognition (OMR) serves as one of the key technologies in Music Information Retrieval by mining symbolic knowledge directly from images of scores. A full-fledged OMR system encompasses both image recognition and music interpretation to convert image data to symbolic representations. This process has proved to be remarkably challenging, so the state-of-the-art OMR systems still leave much to be desired. The development of deep learning in recent years has brought great success in different domains, *e.g.* object recognition and scene understanding, and aroused researcher's interest to address unsolved problems with this new tool. Other work has introduced the first attempts of using deep learning to address OMR, but the models apply more appropriately to text than music scores. In this paper we propose a hybrid model that combines the power of Hidden Markov Models (HMM) and Recurrent Neural Networks (RNN) for end-to-end score recognition.

1. INTRODUCTION

OMR is hard in several aspects. Symbol primitives often have simple shapes, but there are many possible configurations of them to express the same musical meaning. Further, symbols are usually correlated with 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 positions of note heads; rhythm is established along with voice decisions. Some of these correlations are local, and can presumably be captured by convolutional operators. Some are long-term dependencies, requiring the model to have the power of detecting connections between symbols which are spatially separate.

2. HYBRID HMM-RNN MODEL

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 Donahue *et al.* [3], we use Convolutional Neural Networks (CNN) to extract local image features and Long-Short Term Memory (LSTM) networks to capture the dependencies in the context of a whole measure. To enable sequence labeling, we slice each measure into sequential frames and use LSTM to make continuous predictions for the frames. This CNN-RNN model projects image frames to a space that accounts for music semantics such as pitch and rhythm; after projection, we interpret the result with rhythm constraints using HMMs. The overall architecture is shown in Fig. 1.

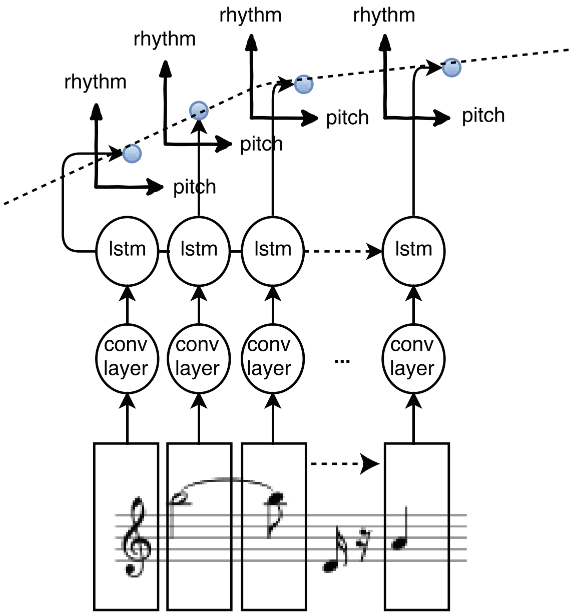


Figure 1: Overview of HMM-RNN model

Our model differs from [5] in two major aspects. First, [5] sliced the input for LSTM in feature space and implicitly used location information from extracted features, while we split the image into frames in the first step and then extract features to be fed into their corresponding LSTMs. The explicit location information contained in ground truth restricts 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, not just predicting a sequence of pitches, to directly model the multi-dimensional nature of the problem.

We define the semantic space $\mathcal{S}$ as a set of label vector that consists 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

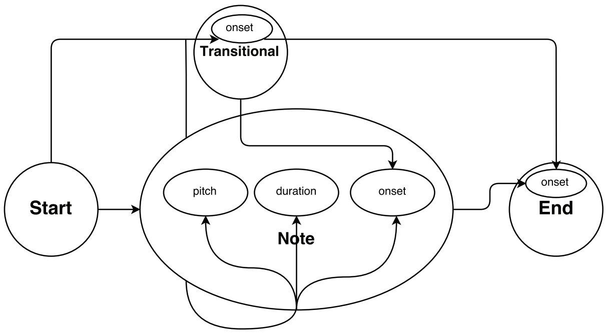
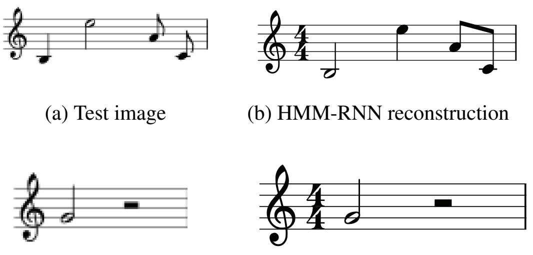


Figure 2: State transitions for music interpretation


(a) Test image

(b) HMM-RNN reconstruction

(c) Test image

(d) HMM-RNN reconstruction

Figure 3: Reconstructed measures with HMM-RNN

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 the monophonic case, we predict a duration label r and pitch label p (both are categorical) for each frame. We assume these two labels are conditionally independent given the feature F extracted by the CNN+RNN, so we connect the output of the LSTM to two separate fully connected layers with weights W_r and W_p, and biases 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:

$$\begin{aligned} P(r_t = r_i|F_t) &= \frac{exp(W_{r_i}F_t + b_{r_i})}{\sum\limits_{r\prime \in R} exp(W_{r\prime}F_t + b_{r\prime})} \\ P(p_t = p_i|F_t) &= \frac{exp(W_{p_i}F_t + b_{p_i})}{\sum\limits_{p\prime \in P} exp(W_{p\prime}F_t + b_{p\prime})} \end{aligned} \tag{1}$$

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

$$\begin{aligned} L_W &= -\sum_{n=1}^{N}\sum_{t=1}^{T} \log P_W(s_{n,t}|X_{n,1:t}, s_{n,1:t-1}) \\ &= -\sum_{n=1}^{N}\sum_{t=1}^{T} (\log P_W(r_{n,t}|X_{n,1:t}, r_{n,1:t-1}, p_{n,1:t-1}) + \\ &\qquad \log P_W(p_{n,t}|X_{n,1:t}, r_{n,1:t-1}, p_{n,1:t-1})) \end{aligned} \tag{2}$$

where N is the total number of measures for training and T is the number of frames in a given 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 2.

3. EXPERIMENT

3.1 Dataset

In our preliminary experiments, we create training data by generating independent score measures using music21 [1] and converting the symbolic data to images with MuseScore [2]. 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 pitches from F#3 to B#5 using the same G clef. Each note was associated with at most one accidental.

3.2 Experimental Setting

We sliced each measure into multiple frames using a frame length of 10 pixels and hop size of 5 pixels. If one frame intersects with over 50% of any annotated 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 initial learning rate set to be 0.001. The model trained after 100,000 iterations was used in the test phase. We used Caffe [4] for neural net implementation and applied one convolutional layer with ReLU and max-pooling for the CNN part, and one layer of LSTMs for the RNN part.

3.3 Evaluation

Given limited time, we performed preliminary evaluations with only a single net configuration and monophonic dataset. With CNN-RNN, we have achieved 94.82% per-frame accuracy for pitch prediction and 88.29% per-frame accuracy for rhythm prediction, but after leaving out background frames, the pitch and rhythm accuracies dropped to 75.47% and 43.17%. Fig. 3 shows us some qualitative results reconstructed by the HMM-RNN model. Our next step will be to extend the experiments to polyphonic scores and improve results with different net configurations.

4. REFERENCES

[1] `http://web.mit.edu/music21`.

[2] `https://musescore.org`.

[3] 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.

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[5] 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.