A PARALLEL FUSION APPROACH TO PIANO MUSIC TRANSCRIPTION BASED ON CONVOLUTIONAL NEURAL NETWORK

Fu'ze Cong[*], *Shuchang Liu*[*], *Li Guo*[*] and *Geraint A. Wiggins*[#,†]

*Beijing University of Posts and Telecommunications
Key Lab of Universal Wireless Communications, Ministry of Education, Beijing, China
School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
† AI Lab, Department of Computer Science, Free University of Brussels, Belgium

ABSTRACT

In this paper, a supervised approach based on Convolutional Neural Networks (CNN) for polyphonic piano transcription is presented. The system consists of pitch detection model, onset/offset detection model, and note search model. The pitch detection model is a single-channel CNN predicting the probabilities of pitches contained in one frame of the audio. The onset/offset model based on dual-channel CNN is used for estimating the probabilities of each pitch's onset or offset in a frame. The note search model is rule-based; it integrates the outputs of the pitch model and onset/offset model to determine the final onset, offset and pitch of notes in audio. Two experiments with different dataset conditions are accomplished to compare with state-of-the-art approaches on the same datasets. Experimental results reveal that the proposed approach preforms better in both frame- and note-based metrics.

Index Terms— Automatic music transcription, deep learning, convolutional neural network, note onset/offset detection.

1. INTRODUCTION

Automatic music transcription (AMT) is the process of generating some form of notation-like musical score from a given acoustic musical signal. Music transcription is considered to be a hard problem, in which human experts outperform current computational systems. However, polyphonic AMT is even more difficult because the combinatorially large output domain makes the modeling more complex. The AMT problem consists of several subtasks such as multi-pitch detection, note onset/offset detection, instrument recognition, extraction of rhythmic information, and so on [1]. In this paper, we concentrate on the multi-pitch detection and note onset/offset detection of polyphonic piano audio and try to extract the pitch, onset and offset of piano notes.

Many frame-based AMT system attempt to recognize pitches in each time frame, and search the onset and offset of per pitch according to the pitch estimation results. In pitch recognition, the most popular methods are based on spectrogram decomposition such as non-negative matrix factorization (NMF) [2-4], sparse decomposition [5] and probabilistic latent component analysis [6-7]. As an alternative, machine learning approaches which classify features extracted from frames to the output pitches are attracting increasing interest. Many classification models have been used to identify pitches in each independent frame, such as support vector machines (SVM) [8-9], deep belief networks (DBN) [10], recurrent neural networks (RNN) [11] and convolutional neural networks (CNN) [12]. The classification models achieve a higher frame-based F-measure but perform worse than spectrogram decomposition in onset detection. To generate note-level transcription results, there are several post-processing methods applied to determine note onsets and offsets. Hidden Markov models (HMM) [7] and hybrid RNNs [13] are used to model the relationship between frame-based outputs and provide a prior probability which can help generate note-level result.

The structure of the classical approach in the onset detection task can be divided into an onset detection function (ODF) and peak picking function of the ODF [14]. Stasiak et al. [14] used deep neural network (DNN) to combine the outputs of several ODFs to high-level probability. Schlüter and Böck [15] applied CNNs to detect the edges of the spectrogram, which can be viewed as onsets of the notes and visualized each layer of the CNN.

In order to improve the performance of a frame-based AMT system in note-based output, we attempt to integrate onset/offset detection model into the current AMT system. However, the current onset/offset detection method can only give the probability of an onset or offset. Therefore, it is hard to apply current onset/offset detect models to polyphonic AMT because of lack of pitch information. In order to solve this problem, we combine the convolution kernels of both pitch detection model [12] and onset detection model [15], and propose a dual-channel CNN model to estimate not only the onset/offset probability, but also to predict in which pitches there is an onset or offset. With the onset/offset probability of each pitch, the AMT system pro-

posed performs better in both frame and note evaluation [12, 16]. Additionally, in order to simulate the practical application case, there are two patterns [7] to split the dataset.

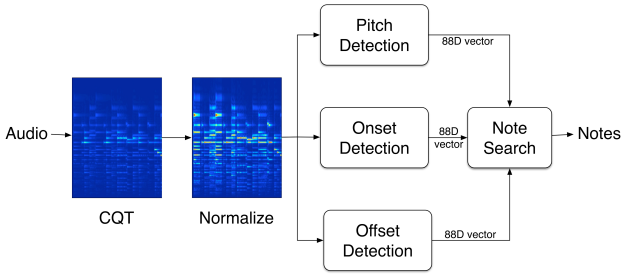

Fig. 1. Overview of the proposed system. The extracted features are fed to pitch, onset and offset detection models. The output about pitch, onset and offset are integrated by note search model to determine the final note events which are the results of transcription.

2. METHOD

The proposed AMT system can be divided into four parts: feature extraction; pitch detection; onset/offset detection; and note search. Fig. 1 shows an overview of the system. Firstly, frame-based spectrum features of the input piano audio are computed. Then several adjacent feature vectors are fed into pitch detection model and onset/offset model. The pitch detection model estimates the probability of each pitch contained in the frame. The onset/offset detector outputs the onset/offset probability of 88 pitches. The note search model combines the results of the models above, and determines the onset, offset time and pitch of each piano note which yields the completion of transcription.

2.1 Feature Extraction

The input piano audio files with 16-bit resolution at 44.1kHz sampling rate are transformed into a time-frequency representation. Instead of short-time Fourier transform (STFT), we use constant Q transform (CQT) [17] to extract frequency features. This is because the result of CQT is in the log frequency axis and log-frequency which is linear in pitch domain is preferred over linear frequency.

The audio data is down-sampled to 16kHz from 44.1kHz and CQTs are calculated over 7 octaves with 36 bins per octave. Every 1024 samples with Hamming window of audio are computed and the hop size is 512. The result of CQT is a 252-dimensional real vector with a frame rate of 31.25Hz. Then we combine 11 frames into one feature image as the input of CNN-based pitch model and onset/offset model.

2.2 Pitch Detection Model

CNNs are neural networks that use convolution instead of matrix multiplication in at least one of their layers [18]. CNNs are a specialized kind of neural network for processing data that has a grid-like topology, such as image data with RGB channels, which can be thought of as three-dimensional grid-like data. In convolutional layers, a set of weights called the *convolution kernel* are multiplied by part of input data with same shape. Results of the convolution kernel make up a feature map according to the position of the convolution data in the input tensor. There are also pooling layers that simplify the feature map, and fully connected layers which handle with inputs by vector multiplication.

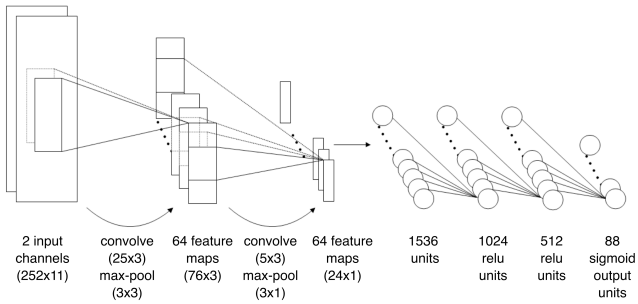

Fig. 2. Architecture of CNN used in pitch detection model. Starting from a stack of two channels, convolutional layers and max-pooling layers in turn compute a set of 64 feature maps classified with fully connected layers.

The structure of the CNN-based pitch detection model is showed in Fig. 2. The input of the model is (252, 11, 2) tensors standing for 11 frames of 2 channel piano audio. The output of the model is an 88-dimensional vector corresponding to the probabilities of pitch in notes A0-C8 on a piano. The labels of the input features are 88-dimensional binary vectors standing for the pitches contained in the center frame (which is the 6^{th}) of the input tensor. We choose the kernels with shapes 25x5 and 5x3, which have been shown perform better than others [12]. The motivation of the kernel structure design is that the pitch information depends on the CQTs which are the first dimension of input tensors (responding to 252); therefore, the convolution kernels are designed to be tensors whose first dimension is bigger than the second in order to get more frequency features in one frame.

2.3 Onset/offset Detection Model

Compared with other deep learning models, CNNs are good at edge detection, because convolution is an effective way of describing changes by applying the same linear transformation of a small local region across the entire input [18]. Onset or offset events can be regarded as the vertical edges in the frequency spectrum, which CNNs can readily detect. Schlüter and Böck [15] applied CNNs to detect onset events in monophonic and polyphonic music. The result indicates that CNNs can achieve better recognition (higher F-measure) than the previous state-of-the-art.

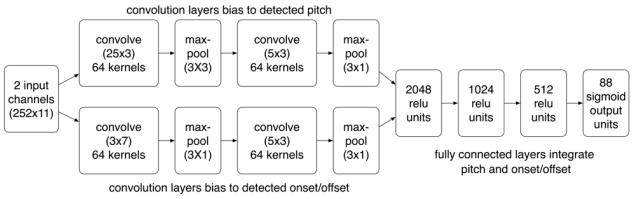

Fig. 3. Structure of CNN-based onset/offset detection model.

Compared with current onset/offset detection models [14, 15, 19] which only generate the onset time of notes, the model proposed in this paper detects both onset/offset and pitch information of notes. The architecture of the CNN-based onset/offset detection model is showed in Fig. 3. The input tensors are the same as those of pitch detection model, that are the images composed by CQTs of 11 frames. There are two convolutional layers in parallel with different kernels. The convolutional layers and max pooling layers in the upper part of Fig. 3 use the same kernels as those of pitch detection model, which extract features about pitches. However, the convolutional layers of the lower part have the wider kernels in contrast to those used for extracting pitch information. Since the second dimension of the input tensor stands for temporal features, wider kernels can detect different levels of energy bursts in long or short-window spectrograms [15] which can then be used to estimate onset and offset probabilities. After the parallel convolution layers, fully connected layers synthesize the feature maps of pitch and onset/offset to obtain the final output which is an 88-dimensional vector corresponding to the onset or offset probabilities of piano pitches from A0 to C8.

2.4 Note Search Model

Our proposed note search model is composed of two processes: onset/offset searching and note searching. First, the most likely frames of onset/offset events are searched by algorithm. After that, the note search model uses the pitch probabilities and their onset offset point per pitch to determine the final onset and offset events of each note. The detail of these two processes is as follows.

2.4.1 Onset/offset Searching

The results of onset/offset detection are fed to the onset/offset searching algorithm per pitch. For every pitch, there is the same approach to search onset and offset. Taking the onset search as an example, the probabilities of continuous temporal frames (which are the outputs of onset detection model) are filtered by a threshold (δ) to make binary results in each pitch. Then the centers of each successive positive frame are viewed as the time points of onset events, which are the outputs of onset searching. The result of note search model with different thresholds (δ) is showed in Fig. 4.

2.4.2 Note Searching

After determining the onset and offset in every pitch, we can ascertain the duration of the notes by using the probabilities of pitches in each frame. The detailed note searching algorithm is as follows. In each pitch, the first offset event between two adjacent onset events will be searched and labeled as final offset event. If there is no offset event detected, we will find three continuous frames whose pitch probabilities are less than 0.1 as the final offset event.

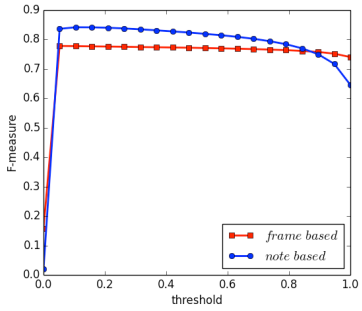

Fig. 4. Frame- and note-based F-measures of note search model with different thresholds (δ) of onset/offset searching. Best results were achieved at thresholds between 0.05 and 0.15.

3. EVALUATION

We use two F-measures to evaluate our proposed AMT system, they are frame and note based [16]. The frame-based metrics sum the true positives, false positives and false negatives of every frame. The note-based metrics are calculated similarly. Specifically, a note event is supposed to be correct when its pitch is right and the onset predicted is with in a 32ms range of ground truth onset.

3.1 Dataset

The models proposed are trained on the MAPS dataset, which consists of about 60 hours of audio recordings. There are 270 pieces of piano music with the corresponding ground truth MIDI transcriptions. There are nine categories of recordings corresponding to different piano types and recording conditions. Among the dataset, 60 pieces of piano music are played by real piano (Disklavier) and other 210 are synthesized by software. To assess the proposed system in different usage scenarios, we designed two distinct experimental conditions. In Condition I, all the audio files are randomly divided into 5 equivalent parts. 80 percent of the pieces are used in training, and 20 percent remained are used for testing. In the training dataset 26 tracks are selected for validation for determining the hyper-parameters. Both real and synthesis audio files are used in training and testing.

In Condition II, models are trained with 210 pieces of synthesis audio, and real piano recordings are used to evaluate the performance of systems. The hyper-parameters are the same in both Conditions. We think that Condition II is more useful in practice, for, in real applications, the sound

of instruments is usually unknown and can not be used for training. As described in section 2.1, the CQTs are computed with 64ms window and 32ms hop. The size of the test data is about 480,000 frames.

F-measure [%] for methods on datasets I and II

Method	*Condition I*		*Condition II*	
	Frame	Note	Frame	Note
Vincent et al. [4]	59.78	69.00	59.60	59.12
Sigtia et al. [12]	74.45	67.05	64.14	54.89
Our Method	**77.76**	**84.16**	**65.02**	**68.23**

Table 1. Top two rows show the results of NMF-based system [4] and CRNN-based system [12]. System proposed achieves the best F-measure in both frame and note on datasets of *Condition I* and *Condition II*.

3.2 Training

The CNN-based pitch detection model and onset/offset detection model are trained by mini-batch gradient descent. The output layers of CNNs use a sigmoid activation function, and the cost function is sigmoid cross entropy. The labels for training are 88-dimensional vectors. For pitch detection models, if the frame is in the duration of one pitch, the relevant position in the vector will be labeled as *true*. Similarly, for onset/offset detection model, we label the 5 closest spectrogram frames to one onset/offset event as *true* in the corresponding position of label vectors. As optimizer, Adam [20] is used because of its rapid convergence speed and less hyper-parameters. Number of convolution layers $L_c \in \{1, 2, 3\}$, shape of kernels $s \in \{(3, 5), (3, 7), (25, 3), (35, 5)\}$, number of fully connected layers $L_{fc} \in \{2, 3\}$ and number of hidden units in fully connected layers $H \in \{512, 1024, 2048\}$ are tested for pitch onset and offset detection. Relu [21] is used as activation function of convolution layers and fully connected layers, and the dropout rate [22] of fully connected layers is 0.7.

3.3 Results

Table 1 summarizes the F-measures of our system and state-of-the-art systems on the datasets of Conditions I and II. The new method proposed achieves higher frame and note based F-measures than the previous approaches. The frame-based F-measures of our system and the CNN-based system in [12] are similar, because the systems both use CNNs to detect pitch information, and the frame-based F-measure depends mostly on the performance of pitch detection model. However, the frame-based result of the system proposed is slightly better, which indicates our onset/offset detection model can make better amendments to the pitch results than the RNN-based music language model of Sigtia et al. [12].

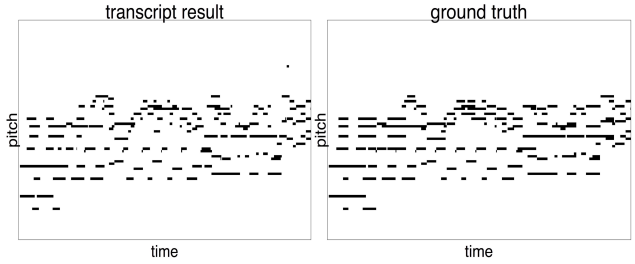

Fig. 5. Transcription result and ground truth for the first 30 seconds of track *MAPS_MUS-alb_esp2_AkPnCGdD*.

Additionally, it can be seen that the note-based F-measure of the system proposed is higher than Vincent et al.'s NMF-based method [4], indicating that the CNN onset detection model can detect the onset events of every pitch with high accuracy, and performs better than method based on spectrogram decomposition which is usually regarded as more suitable in onset detection [12]. The proposed system trained in Condition II, also generates better transcription results than other systems. As expected, the performance of systems train on Condition II dataset are worse than that of Condition Fig. 5 is a graphical representation of the results of our AMT system trained in dataset mode I. The onset events of ground truth are mostly detected, while the duration of each note are usually shorter than the ground truth, because the energy of later parts of notes is relatively low.

4. CONCLUSION

In this paper, a novel AMT system for polyphonic piano music is introduced. The proposed system uses CNN-based onset/offset detection model which has two parallel convolution layers with different kernels to detect the onset/offset events in each pitch, and rule-based note search model to combine the onset/offset events and pitch information. System proposed increases the transcription performance in both frame- and note-based evaluation. In contrast to a CRNN-based system [12] and an NMF-based system [4], our system improves performance of note-based evaluation, while maintaining the advantage of CNN-based model in frame-based evaluation, which we attribute to the high-accuracy onset/offset detection model.

In the future, more suitable machine learning methods will be tested in our note search model, to overcome the inflexibility of the current rule-based model. Additionally, other feature extracting methods will be used to increase the transcript accuracy in practice.

5. ACKNOWLEDGEMENTS

This work is supported by the 111 Project of China (B16006) and Li Guo is the corresponding author.

6. REFERENCES

[1] E. Benetos, S. Dixon, and D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future Directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, Jul. 2013.

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing To Audio and Acoustics*, 2003.

[3] S. Abdallah and M. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. of the 5th International Symposium Music Information Retrieval Conf. (ISMIR)*, Barcelona, Spain, Oct. 2004.

[4] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[5] A. Rizzi, M. Antonelli, and M. Luzi, "Instrument learning and sparse NMD for automatic polyphonic music transcription," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1405-1415, 2017.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Proc. Workshop Advances in Models for Acoustic Processing at Nips*, 2006.

[7] E. Benetos and S. Dixon, "A shift-invariant latent variable model for automatic music transcription," *Computer Music Journal*, vol. 36, no. 4, pp. 81–94, 2012.

[8] G.E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing,* vol. 2007, no. 1, pp. 1–9, 2007.

[9] K. Fathurahman, and D. P. Lestari, "Support vector machine-based automatic music transcription for transcribing polyphonic music into MusicXML," in *Proc. International Conference on Electrical Engineering and Informatics (ICEEI)*, Bali, Indonesia, Jun. 2015.

[10] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A classification-based polyphonic piano transcription approach using learned feature representations," in *Proc. 12th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Miami, FL, USA, Oct. 2011.

[11] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012.

[12] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Trans. Audio, Speech, Language Process.,* vol. 24, no. 5, pp. 927–939, May 2016.

[13] S. Sigtia, E. Benetos, N. Boulanger-Lewandowski, T. Weyde, and Simon Dixon, "A hybrid recurrent neural network for music transcription," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Queensland, Australia, Apr. 2015.

[14] B. Stasiak, J. Mońko, and A. Niewiadomski. "Note onset detection in musical signals via neural-network-based multi-ODF fusion," *International Journal of Applied Mathematics and Computer Science,* vol. 26, no. 1, pp. 203–213, 2016.

[15] J. Schlüter and S. Böck, "Improved musical onset detection with Convolutional Neural Networks." in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP),* Florence, Italy, May 2014.

[16] M. Bay, A.F. Ehmann, and J.S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. of the 13th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, Kobe, Japan, Oct. 2009.

[17] J.C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[18] Goodfellow, I., Y. Bengio, and A. Courvillem, *Deep Learning*, MIT Press, Cambridge, 2016.

[19] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks," in *Proc. of the 11th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*. Utrecht, Netherlands, Aug. 2010.

[20] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science,* 2014.

[21] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. International Conf. on International Conf. on Machine Learning (ICML)*, Haifa, Israel, June 2010.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, June 2014.