International Journal of Engineering & Technology, 7 (3.34) (2018) 590-593

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Single Predominant Instrument Recognition of Polyphonic Music Using CNN-based Timbre Analysis

Daeyeol Kim[1], Tegg Taekyong Sung[2], SooYoung Cho[3], Gyunghak Lee[4], Chae-Bong Sohn[5*]

[1,2,3,5]*Dept. of Electronics and Communications Engineering, Kwangwoon University, Seoul, Korea*
[4]*ICAF, Kwangwoon University, Seoul, Korea*
[*]*Corresponding author E-mail: cbsohn@kw.ac.kr*

Abstract

Classifying musical instrument from polyphonic music is a challenging but important task in music information retrieval. This work enables to automatically tag music information, such as genre classification. In previous, almost every work of spectrogram analysis has been used Short Time Fourier Transform (STFT) and Mel Frequency Cepstral Coefficient (MFCC). Recently, sparkgram is researched and used in audio source analysis. Moreover, for deep learning approach, modified convolutional neural networks (CNN) widely have been researched, but many results have not been improved drastically. Instead of improving backbone networks, we have researched on preprocessing process.
In this paper, we use CNN and Hilbert Spectral Analysis (HSA) to solve the polyphonic music problem. The HSA is performed at the fixed length of polyphonic music, and a predominant instrument is labeled at its result. As result, we have achieved the state-of-the-art result in IRMAS dataset and 3% performance improvement in individual instruments

Keywords: *Instrument recognition, Convolution neural network, Timbre analysis, Hilbert spectrum analysis, Intrinsic mode functions*

1. Introduction

The music is composed of various instruments that play the role of melody, harmony, and bass. The instruments have their own timbres, and human can easily determine which tone is composed of polyphonic music and which timbre is playing the melody line. However, it is not simple to detect these characteristics using a computer. In the real world, music is usually played with several different instruments. Moreover, the style, skill, and tone of players are all different, and extracting the information of those is very difficult.
In the field of Music Information Retrieval (MIR), tagging information related to the playing instrument works as a key role. For example, the relevant information can be additionally informed to the digital music sources and used in the music recommendation system reflecting the preferences of users [1]. Also, it is necessary for people to search musical information using instrument information. The types of instruments are also available for genre analysis of music. Music recommendation system using musical instrument information and genre information enables for uses to increase satisfaction.
Traditionally spectral analysis method has been used to analyze the audio signal process. It divides input audio signal into short time unit. Then using STFT or MFCC to extract audio features to transform spectral, exploiting entire features at once. STFT easily shows the intensity of the frequency over time but requiring domain transformation causes the losses in the original signal. MFCC uses filters on the speech signal to extract features, but filters also lose contents of an input signal. In this paper, we use electroencephalogram (EEG) Hilbert Spectrum which is based on Hilbert transform and EMD to classify musical instrument information. With additional Analysis-Intrinsic Functions Mode (HSA-IMF) method, we have achieved the state-of-the-art result [2].

2. Materials and Methods

Previously, machine learning based methods require diverse domain knowledge of input data in pre- and post-processing. This leads to significant limitations in achieving the desired result. Nonetheless, recent approach, deep learning can extract necessary features from input data without explicit domain knowledge by exploiting multitude feature weights and their nonlinear functions. These techniques have been widely applied to several speech recognition problems and have shown better results than previous methods.
For the deep learning approach in analyzing an audio signal problem, time series signal and image transformation are researched. Forward one uses Recurrent Neural Network (RNN) and summing entire signal sizes to create time-dependent voice signal, which is difficult to interpret. Later one is transforming the audio signal to a 2D image to analyze intensity and shape of the signal frequency. Researchers have been carried conversion of speech signals to images out to focus on the analysis.

2.1. Convolution Neural Network

Convolution neural network is a network that is designed to recognize the transfer of visual patterns themselves, unlike the image feature extraction methods such as SIFT (Scale Invariant Feature Transform) or HOG (Histogram of Oriented Gradient) [2,3].

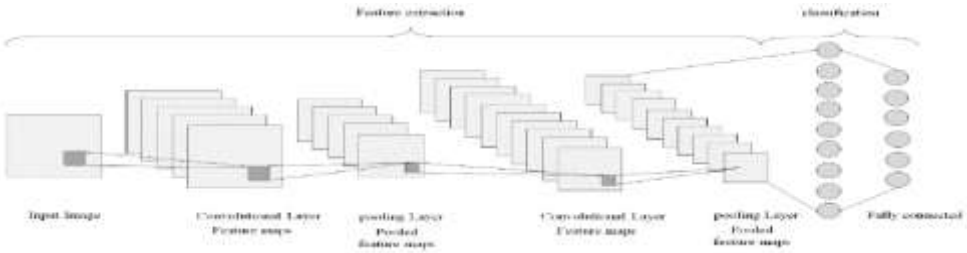

Figure 1: The basic structure of CNN

Figure 1 shows the basic structure of CNN called LeNet. CNN is divided into feature extraction and classification. In the feature extraction, convolutional layer extracts input features using convolution filters and follow activation function, which process filter outputs to non-linear value. Following that, sub-sampling leaves the necessary features in the pooling layer. This reduces the size of data and computing resources, preventing overfitting. In the recognizing handwritten recognition task, the accuracy was close to 97% by using LeNet structure. However, in order to recognize a large-scale image, more convolutional layers and high computation power are required [4].

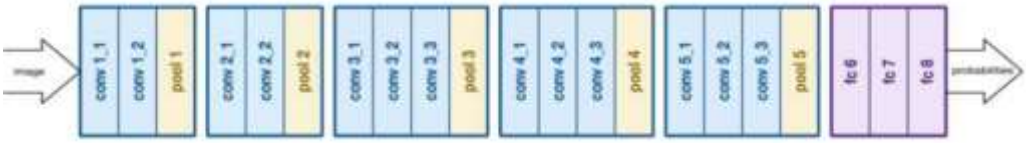

Figure 2: The structure of VGG-16

Figure 2 shows the structure of VGG-16. It only uses small 3x3 convolution filters, unlike the previous models, using large filters in the feature extraction. A stack of small filters will have the same performance as a large filter. Also, this way reduces the number of training weights and makes the decision function more discriminative by employing several non-linearity activation functions [5].

2.2. Audio Signal Processing Methods

Many researchers have been used machine learning to analyze audio signals in the MIR field. Previously, Hidden Markov Model (HMM), Conditional Random Field (CRF) and Mel-cepstral are applied in the preprocessing phase and achieved only 67-70% accuracy [6,7,8]. For the approach using deep learning, RNN is directly used in raw data to extract features, but only size information in the time domain is acquired, causing an unsatisfactory result. As a solution, the researchers have transformed input signal to image and applied CNN.

2.2.1. Audio Feature Extraction

STFT is a method of slicing data to fit a predetermined window at each time, Fast Fourier Transform(FFT), moving to next time, and FFT. (equation 1.) As a result, the frequency spectrum of each time is obtained and becomes two-dimensional data [9].

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}$$

Equation 1. Discrete-time Short-time Fourier transform

MFCCs represent the convergence of Mel-frequency cepstrum (MFC), which expresses the power spectrum of the short-term signal. While general cepstrum divides the frequency band uniformly, for the MFCC, the band is evenly divided on the Mel-scale, acquiring better sound representation. For this reason, it has been widely used in the MIR task. MFCCs can be obtained by processing a Fourier transform on short-term audio, a log of the power spectrum using a Mel-scale filter bank, and a discrete cosine transform (DCT) [10,11].

2.2.2. Hilbert Spectrum Analysis – Intrinsic Mode Functions (Hsa-Imf)

A commonly used in EEG analysis, the Hilbert transform can measure both instantaneous amplitude and instantaneous phase. Attained the intensity and phase of the signal at every moment, it is easy to analyze the signal.

$$H(u)(t) = \frac{-1}{n} \lim_{\varepsilon \to 0} \int_{\varepsilon}^{\infty} \frac{u(t+\tau) - u(t-\tau)}{\tau} d\tau$$

Equation 2. Hilbert transform

Equation 2 shown Hilbert transform, which is useful to analyze a real signal by converting the positive and negative frequency portion to a respected size and zero.

EMD is a method of breaking a signal down without leaving the time domain. Comparing to wavelet decomposition, this method is decomposed by consists of signals. EMD decomposes signal to Intrinsic Mode Functions (IMFs), which is sufficient to make an original signal. With EMD, it is relatively easy to analyze signals without any loss [12].

Assumed that the component of AM-FM decomposition is IMF, the EMD method calculating the Hilbert spectrum is HSA-IMF [13]. We proposed an algorithm that combines the improvement of the problem of Hilbert transform proposed by Rato with the algorithm [14].

2.3. Proposed Method

In order to recognize the predominant instrument, we have approached to timbre aspect. We propose preprocessing method using the HSA-IMF function. This shows waveforms and intensities as an image. The detailed implementation of the proposed method proceeds as follows: audio signal preprocessing, image preprocessing, network learning.

2.3.1. Audio Signal Preprocessing

CNN is a deep learning method that automatically analyzes the image and automatically finds the representative features, and additional pre-processing can increase performance greatly. Consequently, HSA-IMF is applied according to the designated time window to generate the image. Input audio signal is a stereo signal with left and right channels, and each channel is converted to mono signal using root-mean-square (RMS).

2.3.2. Image Preprocessing

CNN receives a fixed size of image as input, however the images generated by the audio pre-processing are too large to be cropped

to fitted size. Moreover, it is difficult to learn the feature depending on the difference in the color of the image. To resolve this problem, we apply histogram equalization to pre-processed image and achieved better performance in image classification [15,16,17].

2.3.3. Network Architecture

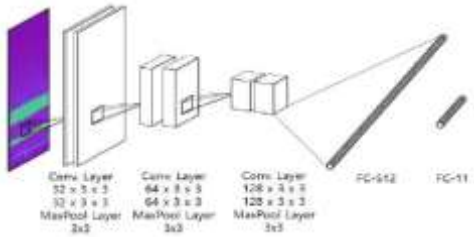

Figure 3: Proposed network structure

We have used a VGG-16 model with the modification that behaves like a normal sorter in large-scale images. Figure 3 shows the proposed network structure. It is specifically designed according to our dataset image. Also, we have changed the size of filter to 3x3 and used 3x3 max-pooling layer [18].

3. Experiment

3.1. IRMAS Dataset and Testing Configuration

The IRMAS dataset has the major instruments commonly used in the orchestra. It consists of 44,100 Hz sampling rate and 16-bit stereo wave files, each with a single labeling of the predominant instrument. It also provides 2874 test datasets and 6705 test datasets with varying lengths from 5s to 20s [16].
In order to see the change in the pre-processing result of the network, we have divided input data into several hours and followed STFT, MFCC, and HSA-IMF. Since the fixed input size of CNN, resizing according to input size is required. In addition, we have changed the dropout value, resulting in higher accuracy effect.

4. Results and Discussion

We have experimented with IRMAS dataset to identify 10 orchestral instruments and performed STFT, spectrograms of MFCCs, and HSA-IMF. Experiments were carried out by modifying the time-domain and dropout values of each transform. We fixed learning rate at 0.001 and epoch at 60.

Table 1: Accuracy of transform by interval

	0.375s	0.5s	0.625s	0.75s	0.875s	1s
STFT	69%	69%	66%	71%	72%	75%
MFCCs	71%	71%	65%	70%	74%	75%
HSA-IMF	74%	72%	69%	66%	41%	29%

Table 1 shows the accuracy of the transmission form according to the interval. Experiments were carried out by changing the dropout value for the highest accuracy.

Table 2: Accuracy of transform by dropout

	0	0.1	0.2	0.3	0.4	0.5
STFT	75%	73%	70%	72%	71%	77%
MFCCs	75%	71%	65%	70%	75%	77%
HSA-IMF	74%	75%	76%	67%	67%	80%

Table 2 shows the accuracy according to dropout figures. As a result, the proposed method using HSA-IMF as pre-processing improves accuracy by 3%, compared to existing STFT and MFCCs. Applying the Class Activation Map (CAM) to the proposed method, we can verify the corresponding place yields a major component tone of the instrument.

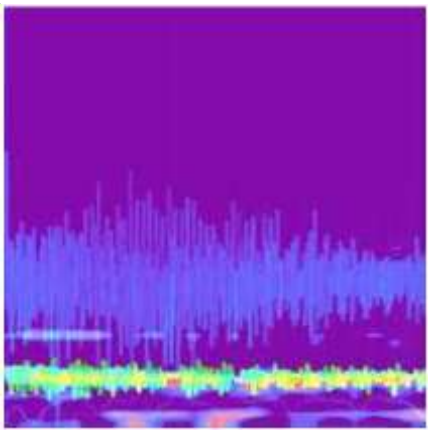

Figure 4: Result of CAM

Figure 4 shows the result of verifying the basis of CNN input data by using Class Activation Map (CAM).

5. Conclusion

In this paper, we have used HSA-IMF method instead of STFT method in pre-processing phase. We have experimented IRMAS dataset and performed at several fixed time intervals. As a result, the HSA-IMF method classified the predominant instrument at the short time. In conclusion, the specification in audio signal processing is difficult to learn because it uses the input spectrum, which has overlapped time and frequency information. Therefore, in order to analyze the frequency easily, it is possible to obtain higher performance by applying various signal processing methods. Using the HSA-IMF method, the performance was 3% higher than that using the previous STFT and MFCCs methods, and it is expected to contribute to genome analysis as well as predominant instrument analysis.

Acknowledgment

This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under ATC Program. No.10052464, 'Multi Dimensional Visualization and Analytic for IoT Data'

References

[1] Downie, J. S. (2003). Music information retrieval. Annual review of information science and technology, 37(1), 295-340
[2] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13), 3812-3814.
[3] Rakotomamonjy, A., & Gasso, G. (2015). Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23(1), 142-153.
[4] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324. doi: 10.1109/5.726791
[5] CNNs Architevtures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more … (2018). Retrived from httpsL//medium.com
[6] Xu, M., Duan, L. Y., Cai, J., Chia, L. T., Xu, C., & Tian, Q. (2004, November). HMM-based audio keyword generation. In Pacific-Rim Conference on Multimedia (pp. 566-574). Springer, Berlin,
[7] Joder, C., Essid, S., & Richard, G. (2011). A conditional random field framework for robust and scalable audio-to-score matching. IEEE Transactions on Audio, Speech, and Language Processing, 19(8), 2385-2397.
[8] Boreczky, J. S., & Wilcox, L. D. (1998, May). A hidden Markov model framework for video segmentation using audio and image

features. In Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on (Vol. 6, pp. 3741-3744). IEEE.

[9] Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 25(3), 235-238.

[10] Eronen, A., & Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on (Vol. 2, pp. II753-II756). IEEE.

[11] Mel-frequency cepstral coefficient analysis in speech recognition. (2006). 2006 International Conference on Computing & Informatics. doi: 10.1109/ICOCI.2006.5276486

[12] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. (1998, March). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences (Vol. 454, No. 1971, pp. 903-995). The Royal Society.

[13] Sandoval, S., De Leon, P. and Liss, J. (2015). Hilbert spectral analysis of vowels using intrinsic mode functions. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). doi: 10.1109/ASRU.2015.7404846

[14] Rato, R. T., Ortigueira, M. D., & Batista, A. G. (2008). On the HHT, its problems, and some solutions. Mechanical Systems and Signal Processing, 22(6), 1374-1394.New child vaccine gets funding boost. (2001)

[15] Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014, May). Deep learning for monaural speech separation. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on (pp. 1562-1566). IEEE.

[16] Han, Y., Kim, J. and Lee, K. (2017). Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1), pp.208-221. doi: 10.1109/TASLP.2016.2632307

[17] Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In Advances in neural information processing systems (pp. 1096-1104)

[18] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.