Fusion of Hilbert-Huang Transform and Deep Convolutional Neural Network for Predominant Musical Instruments Recognition

Xiaoquan Li, Kaiqi Wang, John Soraghan, and Jinchang Ren(✉)

Department of Electronic and Electrical Engineering, University of Strathclyde, Royal College Building, 204 George Street, Glasgow G1 1XW, UK
jinchang.ren@strath.ac.uk

Abstract. As a subset of music information retrieval (MIR), predominant musical instruments recognition (PMIR) has attracted substantial interest in recent years due to its uniqueness and high commercial value in key areas of music analysis such as music retrieval and automatic music transcription. With the attention paid to deep learning and artificial intelligence, they have been more and more widely applied in the field of MIR, thus making breakthroughs in some sub-fields that have been stuck in the bottleneck. In this paper, the Hilbert-Huang Transform (HHT) is employed to map one-dimensional audio data into two-dimensional matrix format, followed by a deep convolutional neural network developed to learn affluent and effective features for PMIR. In total 6705 audio pieces including 11 musical instruments are used to validate the efficacy of our proposed approach. The results are compared to four benchmarking methods and show significant improvements in terms of precision, recall and F1 measures.

Keywords: Predominant musical instrument recognition (PMIR) · Convolutional neural network (CNN) · Hilbert-Huang Transform (HHT)

1 Introduction

Music information retrieval (MIR) has drawn significant research attention in the last decade and has been successfully applied in many applications, such as music retrieval and automatic music transcription [3]. Instead of manually identifying the rhythm, genre and timbre by our ears, MIR techniques can automatically label audio data based on their time and frequency information. As a sub-task of MIR, predominant musical instrument recognition (PMIR) enables customers to search music by instruments, as well as making music transcription easier and more accurate [6]. However, PMIR is a very challenging topic and current PMIR approaches have yet to be commercialised due mainly to the lack of robust performance. However, it is quite useful in some applications such as assisting automatic music transcription (AMT) detection, crude instrument classification, and instrument characterisation.

The identification of a musical instrument depends primarily on its timbre [7]. For the physics point of view, the timbre produced from an object is determined from its

J. Romero et al. (Eds.): EvoMUSART 2020, LNCS 12103, pp. 80–89, 2020.
https://doi.org/10.1007/978-3-030-43859-3_6

vibrational state, which characterises the object's waveform and harmonic properties. For the viewpoint of cognitive psychology [8], timbre contains complex musical features such as the attack time, the temporal envelope, the spectral envelope. Both spectral and temporal features of sounds contribute to timbre perception. For a specific musical instrument, its spectrum change is very complicated. Due to different playing techniques, the same type of musical instrument will have apparent changes in timbre. It may also be affected by mood, humidity, temperature during performance. Therefore, identifying instruments in music pieces is an exciting and challenging problem.

Generally, there are two kinds of musical data: monophonic and polyphonic music. In monophonic music, the instrument is played independently. Most of the work on instrument recognition is done under the assumption of independent performance, which simplifies the recognition task. In the case of separate recordings, musical instrument digital interface (MIDI) can store each instrument in a channel, making it easier for a single instrument to be detected. Bhalke et al. [9] proposed a musical instruments classification method based on Mel Frequency Cepstral Coefficient (MFCC) features and a Counter Propagation Neural Network. Their method obtained an accuracy of 91.84% for recognising 19 instruments. Banerjee et al. [10] proposed an approach for string instrument recognition that gave an accuracy of 89.85% using a Support Vector Machine (SVM) and 100% with a Random Forest classifier on the IRCAM dataset which contained only 4 string-family instruments.

However, music is more often polyphonic than monophonic, such as in a symphony orchestra or recording live scenes. Recognition of a single instrument in a polyphonic music recording is therefore much more difficult, and several attempts have been made for automatic recognition. Slizovskaia et al. [11] extracted instrument features through a standard bag-of-features pipeline and achieved a 67% classification accuracy on IRMAS database [12], which includes 11 different instruments in the recordings. Han et al. [6] integrated MFCC and CNN together to get a classification accuracy of 63.3% on IRMAS dataset as in [12]. Li et al. [13] achieved 82.74% accuracy on the MedleyDB dataset [14] by applying a CNN model on the raw audio data. In 2018, Hung et al. [15] achieved an 81.7% accuracy by using the constant Q-transform (CQT) and skip connection methods on MedleyDB and other datasets. In 2019, Gururani et al. [16] proposed an approach for handling weakly labeled data in an attention enable models in the OpenMIC dataset, which contains 20 instruments [17]. Their method achieved an average F1-score of approx. 81.03%.

Existing approaches for identifying instruments in monophonic music have achieved relatively satisfied performance. However, there is still a large gap for improving the accuracy of instrument recognition in polyphonic music pieces. In this paper, we propose a new framework for PMIR which uses the Hilbert-Huang Transform (HHT) to generate a feature matrix from music recordings and input these features into a new deep convolutional neural network for automatically learning instrument features from polyphonic music recordings [12]. In the experimental results, it shows that the proposed method outperforms other benchmarking methods where useful analysis and findings are carried out.

2 The Proposed Method

In the proposed framework (Fig. 1), we use the HHT [18] to generate the Hilbert spectrum for each instrument in the polyphonic music pieces. Then we build a deep convolutional neural network (DCNN) to take the Hilbert spectrum as input and produce the classification label as the output.

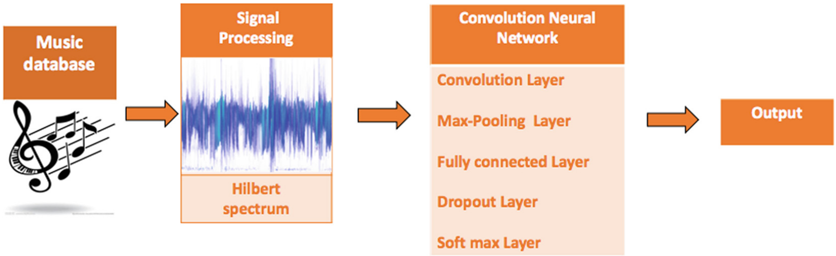

Fig. 1. The flowchart of the proposed PMIR system.

2.1 Hilbert-Huang Transform

Hilbert-Huang Transform (HHT) was proposed by Huang from NASA in 1998 [19]. This method is suitable for nonlinear and non-stationary signal analysis. At present, HHT technology has been applied in many fields, such as geophysics and biomedicine, and achieved excellent results [20]. HHT mainly includes two parts: Empirical Mode Decomposition (EMD) and Hilbert Spectral Analysis (HSA), where EMD is used to first decompose the given signals into several Intrinsic Mode Functions (IMFs). The Hilbert transform is then applied for each IMF to obtain the corresponding Hilbert Spectrum. Finally, the Hilbert spectrum of the original signal can be obtained by summing all the Hilbert spectra of the IMF. Samples of the Hilbert spectra for each instrument can be seen in Fig. 2, where they were played by cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and voice, respectively. It shows the performance of HHT varies according to different instruments. HHT was first used in PMIR in 2018 by Kim et al. [5], where HHT was proposed to replace the short-time Fourier Transform (STFT) and MFCC and achieved improved results. HHT and Fourier transform based calculation is given as follows for comparison:

$$x(t) = Re\sum_{i=1}^{n} a_i(t)e^{j\int w_i(t)dt} \tag{1}$$

$$x(t) = \sum_{i=1}^{\infty} a_k e^{jk(2\pi/T)t} \tag{2}$$

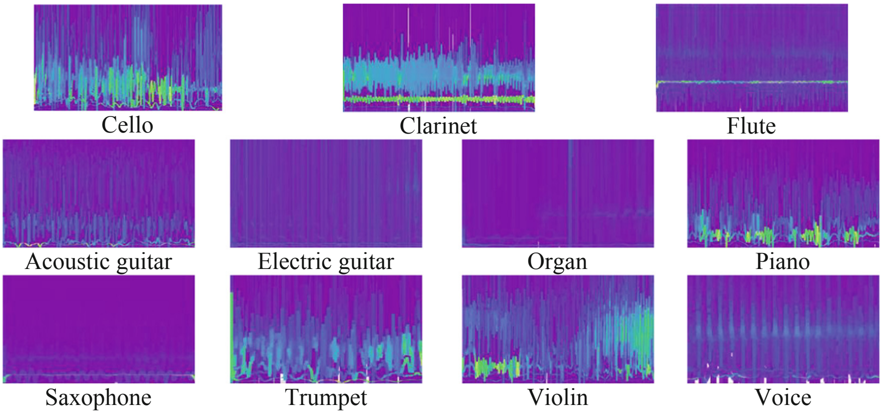

Fig. 2. Example of Hilbert spectra for each of the 11 instruments.

The main difference is that HHT can be considered as a phase shift converter whilst the traditional Fourier analysis uses a series of trigonometric basis functions for orthogonal operations on signals [21]. However, the resulting Fourier spectrum is only the weighted mean of the frequencies over a certain period and cannot be accurately described by time-frequency changes. To this end, HHT is adopted in our paper to define the instantaneous frequency for coping with more complex signals such as polyphonic music pieces.

2.2 Architecture of DCNN

Our DCNN was inspired by the VGG-16 model [22]. VGG-16 contains 16 hidden layers (13 convolutional and 3 fully connected). The pooling size is always set as 2×2 and the filter size is set as 3×3, and it shows that deepening the network layers can improve performance [22]. In our proposed DCNN model, we have 6 convolutional layers followed by a dropout layer, a fully connected layer, another dropout layer and a softmax layer. Each convolutional layer is followed by a max pooling layer. Therefore, our model has much less parameters but higher performance.

In Fig. 3, we illustrate the architecture of the proposed DCNN. The specific details of the architecture are provided in Table 1. There are six convolutional layers in the proposed model, where each convolutional layer is followed by a batch normalization layer, an activation layer, and a max pooling layer. The feature matrix produced by the Hilbert spectrum has size $135 \times 240 \times 3$, where 3 means RGB channels, and the size of each channel is 135×240. RGB channels give a colorful spectrogram. It has three channels (Red, green, blue). They can provide more information, such as frequency energy rather than grayscale. A rectified linear unit (ReLU) is selected as the activation function for each activation layer due to its popularity and ability to increase the learning speed. In the first convolutional layer the stride size is 2×2, which is changed to 1×1 for the rest of the convolutional layers. Both pool size and stride size of each max pooling layer are set to 2×2. In addition, the input for each convolutional layer is

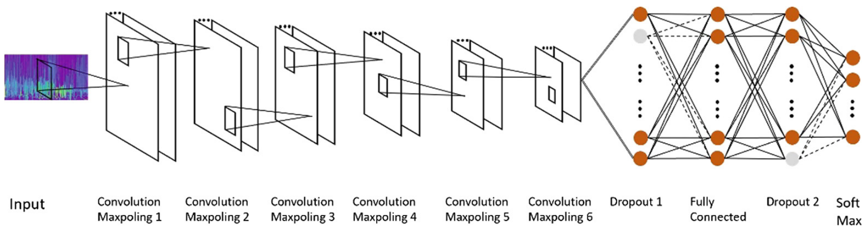

Fig. 3. Flowchart of the proposed DCNN

Table 1. Proposed DCNN structure.

Layers	Output size	Description
HHT (Input)	135 × 240 × 3	Feature matrix from Hilbert spectrum
Convolution 1	68 × 120 × 32	Filter size: 7 × 7; Stride size: 2 × 2
Max pooling 1	34 × 60 × 32	Pool size: 2 × 2; Stride size: 2 × 2
Convolution 2	34 × 60 × 64	Filter size: 5 × 5; Stride size: 1 × 1
Max pooling 2	17 × 30 × 64	Pool size: 2 × 2; Stride size: 2 × 2
Convolution 3	17 × 30 × 128	Filter size: 3 × 3; Stride size: 1 × 1
Max pooling 3	8 × 15 × 128	Pool size: 2 × 2; Stride size: 2 × 2
Convolution 4	8 × 15 × 256	Filter size: 3 × 3; Stride size: 1 × 1
Max pooling 4	4 × 7 × 256	Pool size: 2 × 2; Stride size: 2 × 2
Convolution 5	4 × 7 × 512	Filter size: 3 × 3; Stride size: 1 × 1
Max pooling 5	2 × 3 × 512	Pool size: 2 × 2; Stride size: 2 × 2
Convolution 6	2 × 3 × 1024	Filter size: 3 × 3; Stride size: 1 × 1
Max pooling 6	1 × 1 × 1024	Pool size: 2 × 2; Stride size:2 × 2
Dropout 1	1024	Dropout fact: 0.25
Fully connected	1024	Output size: 11
Dropout 2	1024	Dropout fact: 0.25
Softmax	11	Softmax function

same-padded in order to preserve the spatial resolution. The number of filters for each convolution layer is twice that of the previous layer, increasing from 32 up to 1024 at the last layer. After the final max pooling layer, a dropout layer is added before and after the fully connected layer to avoid overfitting. In the end, the Softmax function is used for 11 instruments classification.

3 Experiments

In this experiment, 6705 audio files were extracted from the IRMAS dataset [12] and used to investigate the performance of the proposed method. The audio files are in 16-bit stereo .wav format, sampled at 44.1 kHz, and have a length of 3 s. There are 11 different instruments in the music (see Fig. 2). The percentage of training, validation, and testing data was set to 55%, 15% and 30% respectively.

3.1 Key Parameters

Adjusting hyper-parameters to improve network performance is a vital step in deep learning. To get the optimal performance, we only vary the minibatch size from 50 to 300 in intervals of 50. From Table 2, the best testing accuracy can achieve 85% when the minibatch size is 250. As a result, the batch size in our experiments was set to 250. With minibatch size equaling 250, the classification accuracy of proposed model with varying dropout factor from 0 to 0.5 at a step of 0.05 is shown in Fig. 4. Based on the results, the dropout rate is set as 0.25 in our work. For the other parameters, Adam is used as the optimiser with a learning rate of 0.01.

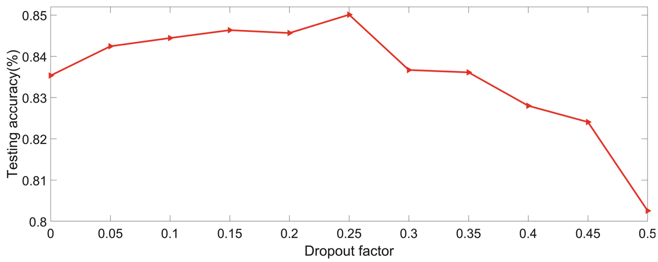

Fig. 4. Classification performance with different dropout factors.

Table 2. Training performance with different minibatch size

Batch size	50	100	150	200	250	300
Testing accuracy (%)	84.30	84.96	84.09	84.54	**85.01**	84.67

3.2 Comparison with Other Methods

To further evaluate the efficacy of the proposed PMIR framework, three conventional approaches are used for benchmarking in terms of precision, recall and F1-measurement [23]. These conventional frameworks are based on Audio Content Analysis (ACA) system [20, 24] and three machine learning model, i.e. random forest (RF) [4], SVM [1] and shallow neural network (SNN) [2]. These conventional methods are representative in the field of machine learning, and the ACA features include most of the features used in music analysis. The tree number of RF is 300, the LIBSVM toolbox [1] is selected as SVM learner, and the neuron number in the hidden layer in SNN is set as 70. For each music piece, the ACA system is used to extract 19 features from both time and frequency domains such as Peak envelop, autocorrelation coefficients, MFCCs, and pitch chroma. Then these features will concentrate into a feature vector with the length of 105 and be entered into the machine learning model for instruments' classification. What's more, we compared it with a new HAS-IMF CNN algorithm proposed in 2018 [5], using the same dataset and computing environment.

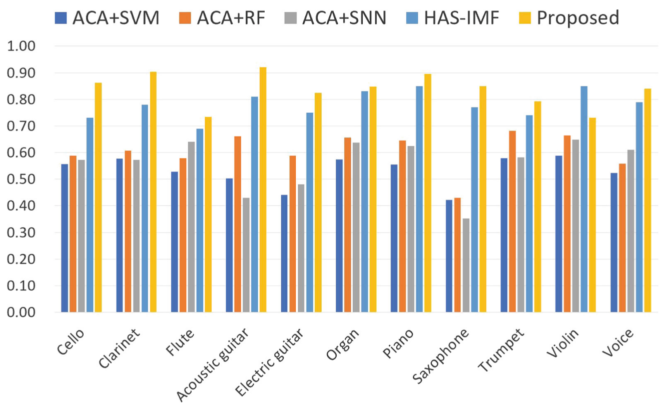

Fig. 5. F1-measurement of each instrument for five methods.

Table 3. Overall precision, recall and F1-measurement of five methods.

Methods	Precision	Recall	F1
ACA+SVM [1]	0.53	0.54	0.53
ACA+SNN [2]	0.57	0.56	0.56
ACA+RF [4]	0.61	0.62	0.61
HAS-IMF [5]	0.77	0.80	0.78
Proposed	**0.82**	**0.85**	**0.84**

Experimental results including the overall precision, recall and F1-measurement are presented in Table 3, and the F1-measurement of each instrument is shown in Fig. 5. As can be seen, the proposed PMIR framework generates the best overall performance and outperforms conventional frameworks in classifying individual instruments. In Table 3, the ACA features are combined in the first three classifiers, i.e. SVM, SNN, and RF. Among them, the RF based method seems the best, yet it is significantly poorer than the HAS-IMF, due mainly to the integration of the HHT spectrogram and CNN model. However, thanks to our improved CNN structure, the proposed model has significantly outperformed all others including HAS-IMF, where the precision, recall, and F1 measures are improved by 5%, 5% and 6%, respectively.

In our model, the batch normalization and Max pooling are followed by each convolutional layer, and the dropout layers are put before and after the fully connected layer. However, in HAS-IMF, a dropout layer is followed by every two convolutional layers. Although the dropout layer can reduce the training time, too many dropout layers may lead to the network not being fully trained. Furthermore, it does not include a batch normalization layer which may lead to the data becoming unbalanced. Therefore, our proposed CNN model outperforms HAS-IMF by 6%.

In Figs. 5 and 6, the classification performance of individual instruments is presented. As can be seen HAS-IMF and the proposed method are significantly better than the fusion of traditional features (ACA) and other machine learning techniques. For the

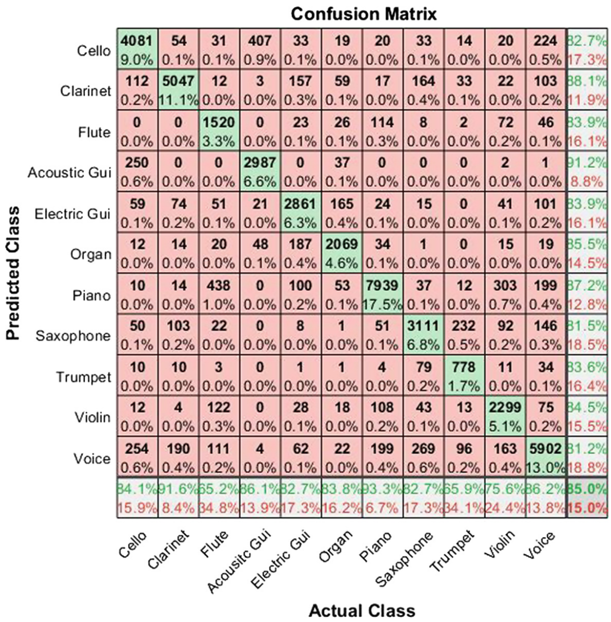

Fig. 6. Confusion matrix of proposed methods in 11 instruments.

violin, HAS-IMF gives the best result. But with a better structure of deep learning network, our proposed method produces better performance on the rest of the individual instruments. Another finding is that the performance of individual instruments is potentially related to the types of instruments. For example, as can be seen in Fig. 6, the saxophone sometimes is misclassified to a clarinet and trumpet, because both saxophone and clarinet belong to the woodwind family, and both saxophone and trumpet belong to the wind family. In addition, piano is mostly misclassified into violin and flute. The main reason is that this dataset has many variables, it cannot only discuss about timbre. The pitch of violin and flute is very high, and the piano sometimes composes the main melody by a high pitch. Therefore, there is the confusion of the piano with violin and flute. Also, the human voice is often misclassified into cello, clarinet, piano and saxophone since it is very complicated, and its pitch or timbre may be close to some of the instruments.

4 Conclusions

In this paper, a new framework for predominant musical instrument recognition (PMIR) in polyphonic music was proposed. First the Hilbert-Huang Transform (HHT) is used to generate the Hilbert spectrum of the audio data, which is then used as input to a deep convolutional neural network (DCNN). The optimal DCNN model is trained on the IRMAS dataset, and objective evaluation has shown that the proposed method can give a classification accuracy of 85%. The proposed method also outperforms three conventional approaches, which shows that the image-based deep learning method has good potential for music instrument recognition. In the future, the MedleyDB dataset [14] and OpenMIC-2018 dataset [17] will be used to evaluate the proposed method and more comprehensive experiments will be done to further validate the usefulness and effectiveness of the proposed method. Key parameters' selection and network structure will also be optimised. In addition, the performance of state-of-art deep learning models such as stacked autoencoders [25], and residual neural networks [26] on PMIR will be investigated. Optimized model like gravitational search algorithm [27] can be applied to improve the feature selection.

Acknowledgements. The authors would like to thank, Dr Yijun Yan for useful discussions and Calum MacLellan for kindly proofreading of the paper.

References

1. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011)
2. Battiti, R.: First-and second-order methods for learning: between steepest descent and Newton's method. Neural Comput. **4**, 141–166 (1992)
3. Downie, J.S., Ehmann, A.F., Bay, M., Jones, M.C.: The music information retrieval evaluation eXchange: some observations and insights. Stud. Comput. Intell. **274**, 93–115 (2010)
4. Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**(3), 18–22 (2002)
5. Kim, D., Sung, T.T., Cho, S., Lee, G., Sohn, C.-B.: A single predominant instrument recognition of polyphonic music using CNN-based timbre analysis. Int. J. Eng. Technol. **7**, 590–593 (2018)
6. Han, Y., Kim, J., Lee, K., Han, Y., Kim, J., Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **25**, 208–221 (2017)
7. Fletcher, N.H., Rossing, T.D.: The Physics of Musical Instruments. Springer, Heidelberg (2012)
8. McAdams, S., Giordano, B.L.: The perception of musical timbre. In: The Oxford Handbook of Music Psychology, pp. 113–123 (2016)
9. Bhalke, D., Rao, C.R., Bormane, D.S.: Automatic musical instrument classification using fractional fourier transform based-MFCC features and counter propagation neural network. J. Intell. Inf. Syst. **46**, 425–446 (2016)
10. Banerjee, A., Ghosh, A., Palit, S., Ballester, M.A.F.: A novel approach to string instrument recognition. In: Mansouri, A., El Moataz, A., Nouboud, F., Mammass, D. (eds.) ICISP 2018. LNCS, vol. 10884, pp. 165–175. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94211-7_19
11. Slizovskaia, O., Gómez, E., Haro, G.: Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies. In: SMC 2016 – 13th Sound and Music Computing Conference, Proceedings, pp. 442–447 (2016)
12. Bosch, J.J., Janer, J., Fuhrmann, F., Herrera, P.: A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In: ISMIR, pp. 559–564 (2012)

13. Li, P., Qian, J., Wang, T.: Automatic instrument recognition in polyphonic music using convolutional neural networks. arXiv preprint: arXiv:1511.05520 (2015)
14. Bittner, R.M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., Bello, J.P.: MedleyDB: a multitrack dataset for annotation-intensive mir research. In: ISMIR, pp. 155–160 (2014)
15. Hung, Y.N., Chen, Y.A., Yang, Y.H.: Multitask learning for frame-level instrument recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 381–385 (2019)
16. Gururani, S., Sharma, M., Lerch, A.: An Attention Mechanism for Musical Instrument Recognition. arXiv preprint: arXiv:1907.04294 (2019)
17. Humphrey, E., Durand, S., McFee, B.: OpenMIC-2018: an open data-set for multiple instrument recognition. In: ISMIR, pp. 438–444 (2018)
18. Sandoval, S., De Leon, P.L., Liss, J.M.: Hilbert spectral analysis of vowels using intrinsic mode functions. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 569–575. IEEE (2015)
19. Müller, M.: Fundamentals of Music Processing: Audio, Analysis, Algorithms. Applications. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21945-5
20. Lerch, A.: An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. Wiley-IEEE Press, Hoboken (2012)
21. Ayenu-Prah, A., Attoh-Okine, N.: Comparative study of Hilbert-Huang transform, Fourier transform and wavelet transform in pavement profile analysis. Veh. Syst. Dyn. **47**, 437–456 (2009)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2014)
23. Yan, Y., et al.: Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cogn. Comput. **10**, 94–104 (2018)
24. Peeters, G.: A large set of audio features for sound description (similarity and classification). CUIDADO project IRCAM technical report (2004)
25. Zabalza, J., et al.: Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. Neurocomputing **185**, 1–10 (2016)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
27. Sun, G., Ma, P., Ren, J., Zhang, A., Jia, X.: A stability constrained adaptive alpha for gravitational search algorithm. Knowl.-Based Syst. **139**, 200–213 (2018)