Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network

Tom LH. Li*, Antoni B. Chan*and Andy HW. Chun*

Abstract—**Music genre classification has been a challenging yet promising task in the field of music information retrieval (MIR). Due to the highly elusive characteristics of audio musical data, retrieving informative and reliable features from audio signals is crucial to the performance of any music genre classification system. Previous work on audio music genre classification systems mainly concentrated on using timbral features, which limits the performance. To address this problem, we propose a novel approach to extract musical pattern features in audio music using convolutional neural network (CNN), a model widely adopted in image information retrieval tasks. Our experiments show that CNN has strong capacity to capture informative features from the variations of musical patterns with minimal prior knowledge provided.**

Keywords: ***music feature extractor, music information retrieval, convolutional neural network, multimedia data mining***

1 Introduction

Automatic music genre classification has grown in vast popularity in recent years as a result of the rapid development of the digital entertainment industry. As a first step of genre classification, feature extraction from musical data will significantly influence the final classification accuracy. The annual international contest Music Information Retrieval Evaluation eXchange (MIREX) holds regular competitions for audio music genre classification that attracts tens of participating groups each year. Most of the systems rely heavily on timbral, statistical spectral features. Feature sets pertaining to other musicological aspects such as rhythm and pitch are also proposed, but their performance is far less reliable compared with the timbral feature sets. Additionally, there are few feature sets aiming at the variations of musical patterns. The inadequateness of musical descriptors will certainly impose a constraint on audio music genre classification systems.

In this paper we propose a novel approach to automatically retrieve musical pattern features from audio music using convolutional neural network (CNN), a model that is adopted in image information retrieval tasks. Migrating technologies from another research field brings new opportunities to break through the current bottleneck of music genre classification. The proposed musical pattern feature extractor has advantages in several aspects. It requires minimal prior knowledge to build up. Once obtained, the process of feature extraction is highly efficient. These two advantages guarantee the scalability of our feature extractors. Moreover, our musical pattern features are complementary to other main-stream feature sets used in other classification systems. Our experiments show that musical data have very similar characteristics to image data so that the variation of musical patterns can be captured using CNN. We also show that the musical pattern features are informative for genre classification tasks.

*Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, Email: lihuali2@student.cityu.edu.hk, abchan@cityu.edu.hk, andy.chun@cityu.edu.hk

2 Related Works

By the nature of data involved in analysis, the field of music genre classification is divided to two different scopes: symbolic and audio. Symbolic music genre classification studies songs in their symbolic format, such as MIDI, MusicXML, etc. Various models (Basili et. al. [1], McKay et. al. [2], Ponce et. al. [3]) have been proposed to perform symbolic music genre classification. Feature sets representing instrumentation, musical texture, rhythm, dynamics, pitch statistics, melody, etc. are used as input for a wide variety of generic multi-class classifiers.

Identifying the music genre directly from audio signal is more difficult because of the increased difficulties in feature extraction. In symbolic musical data, information such as instrument, note onsets are readily available in the precise musicological description of the songs. For audio music however, only the recorded audio signal is readily available. Trying to apply methodologies in symbolic music analysis on auto-transcribed audio data is highly impractical since building up a reliable auto-transcription system for audio music appears to be a more challenging task than audio genre classification itself. In fact, the best candidate scored only about 70% in the 2009 MIREX medoly extraction contest, a simpler task than auto-transcription. Researchers therefore need to turn to alternative approaches to extract informative feature sets for genre classification, such as,

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

- Tzanetakis et. al. [4, 5, 6]: STFT, MFCC, Pitch Histogram, Rhythm Histogram
- Bergstra et. al. [7]: STFT, RCEPS, MFCC, Zero-crossing Rate, Spectral summary, LPC.
- Ellis et. al. [8]: MFCC, Chroma
- Lidy et. al. [9, 10]: Rhythm Pattern, Statistical Spectrum Descriptor, Rhythm Hisitogram, Symbolic Feature from auto-transcribed music.
- Meng et. al. [11]: MFCC, Mean and variance of MFCC, Filterbank Coefficients, Autoregressive model, Zero-crossing Rate, Short-time Energy Ratio.

Most of the proposed systems concentrate only on feature sets extracted from a short window of audio signals, using statistical measurements such as maximum value, average, deviation, etc. Such features are representative of the "musical texture" of the excerpt concerned, i.e. timbral description. Feature sets concerning other musicological aspects such as rhythm and pitch are also proposed, but their performance is usually far worse than their timbral counterparts. There are few feature sets which capture the musical variation patterns. Relying only on timbral descriptors would certainly limit the performance of genre classification systems; Aucouturier et. al. [12] indicates that a performance bottleneck exists if only timbral feature sets are used.

The dearth of musical pattern features can be ascribed to the elusive characteristics of musical data; it is typically difficult to handcraft musical pattern knowledge into feature extractors, as they require extra efforts to handcraft specific knowledge into their computation processes, which would limit their scalability. To overcome this problem, we propose a novel approach to automatically obtain musical pattern extractors through supervised learning, migrating a widely adopted technology in image information retrieval. We believe that introducing technology in another field brings new opportunities to break through the current bottleneck of audio genre classificaion.

3 Methodology

In this section, we briefly review the CNN and the proposed music genre classification system.

3.1 Convolutional Neural Network

The design of convolutional neural network (CNN) has its origin in the study of biological neural system. The specific method of connections discovered in cats' visual neurons is responsible for identifying the variations in the topological structure of objects seen [13]. LeCun incorporate such knowledge in his design of CNN [14] so that its first few layers serve as feature extractors that would be automatically acquired via supervised training. It is shown from extensive experiments [14] that CNN has considerable capacity to capture the topological information in visual objects.

There are few applications of CNN in audio analysis despite its successes in vision research. The core objective of this paper is to examine and evaluate the possibilies extending the application of CNN to music information retrieval. The evaluation can be further decomposed into the following hypotheses:

- The variations of musical patterns (after a certain form of transform, such as FFT, MFCC) is similar to those in images and therefore can be extracted with CNN.
- The musical pattern descriptors extracted with CNN are informative for distinguishing musical genres.

In the latter part of this paper, evidence supporting these two hypotheses will be provided.

3.2 CNN Architecture for Audio

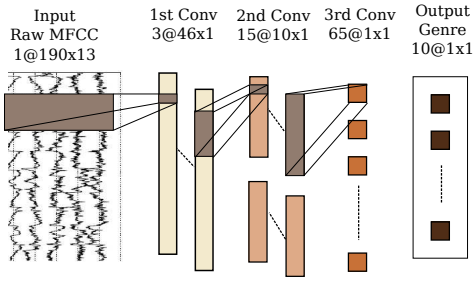

Figure 1: CNN to extract musical patterns in MFCC

Figure 1 shows the architecture of our CNN model. There are five layers in total, including the input and output layers. The first layer is a 190×13 map, which hosts the 13 MFCCs from 190 adjacent frames of one excerpt. The second layer is a convolutional layer of 3 different kernels of equal size. During convolution, the kernel surveys a fixed 10×13 region in the previous layer, multiplying the input value with its associate weight in the kernel, adding the kernel bias and passing the squashing function. The result is saved and used as the input to the next convolutional layer. After each convolution, the kernel hops 4 steps forward along the input as a process of subsampling. The 3rd and 4th layer function very similarly to the 2nd layer, with 15 and 65 feature maps respectively. Their kernel size is 10×1 and their hop size is 4. Each kernel of a convolutional layer has connections with all the feature maps in the previous layer. The last layer is an output layer with full connections with the 4th layer. The parameter selection process is described in Section 4.2.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

It can be observed from the topology of CNN that the model is a multi-layer neural network with special constraints on the connections in the convolutional layers, so that each artificial neuron only concentrates on a small region of input, just like the receptive field of one biological neuron. Because the kernel is shared across one feature map, it becomes a pattern detector that would acquire high activation when a certain pattern is shown in the input. In our experimental setting, each MFCC frame spans 23ms on the audio signal with 50% overlap with the adjacent frames. Therefore the first convolutional layer (2nd layer) detects basic musical patterns appear in 127ms. Subsequent convolutional layers therefore capture musical patterns in windows size of 541ms and 2.2s, respectively. The CNN is trained using the stochastic gradient descent algorithm [15]. After convergence, the values in the intermediate convolutional layers can be exported as the features of the corresponding musical excerpt.

The model we use is a modified CNN model presented in [16]. Compared with the traditonal CNN model, we observed that the training is easier, and the capacity loss is negligible. In return, as much as **66.8%** of computational requirement is saved.

3.3 Music Genre Classification

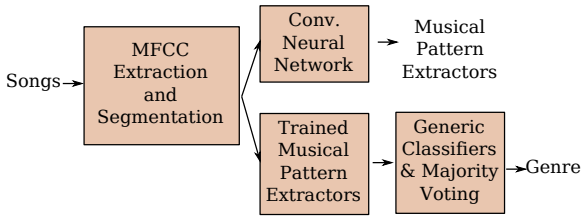

Figure 2: Overview of the classification system

Figure 2 shows the overview of our classification system. The first step of the process is MFCC extraction from audio signals. MFCC is an efficient and highly informative feature set that has been widely adopted for audio analysis since its proposal. After MFCC extraction, the input song is transformed into an MFCC map with 13 pixels wide which is then segmented to fit the input size of CNN. Provided the song label, the musical pattern extractors are automatically aquired via supervised learning. Those extractors are used to retrieve high-order, pattern-related features which will later serve as the input of generic, multi-class classifiers such as Decision Tree Classifiers, Support Vector Machine etc. After classification of each song segments, the result is aggregated in a majority voting process to produce the song-level label.

4 Results and Analysis

4.1 Dataset

The dataset of our experiment is the GTZAN dataset which has been used to evaluate various genre classification systems [4, 7, 10]. It contains 1000 song excerpts of 30 seconds, sampling rate 22050 Hz at 16 bit. Its songs are distributed evenly into 10 different genres: Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae and Rock.

4.2 CNN Pattern Extractor

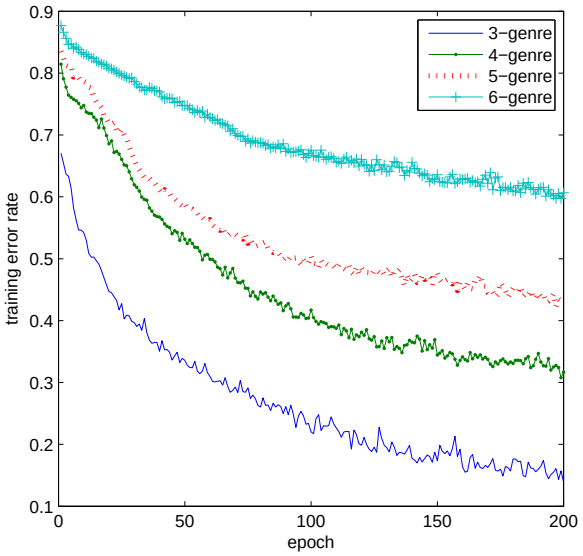

Figure 3: Convergence Curve in 200-epoch training

Figure 3 shows the convergence of the training error rate of our CNN model, on four sub-datasets extracted from the GTZAN dataset. The smallest dataset contains 3 genres: Classical, Jazz and Rock. The latter datasets increase in size as Disco, Pop and Blues genres are added. From the figure we can observe that the trend of convergence over different datasets is similar, however the training on a 3-genre dataset converges much faster than the training on a 6-genre dataset. This shows the difficulty in training CNN increases drastically when the number of genres involved in training increases. We believe this is because the CNN gets confused with the complexity of the training data and therefore never obtains suitable pattern extractors in the first few layers. Additionally we also found that the combination of genres in the 3-genre subset will not affect the training of CNN. All combinations have very similar curve of convergence.

Based on the observations above, the training of our CNN feature extractors are divided in four parallel models to cover the full 10-genre GTZAN dataset. Three models are arbitrarily selected to cover 9 non-overlapping gen-

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

res, while one model is deliberately chosen to train on the 3 most difficult-to-classify genres shown in [4], i.e. Blues, Metal and Rock. Dividing the dataset into small subsets to train the CNN feature extractors may have the side-effect that features extracted to classify songs within one subset may not be effective in intersubset classification, and therefore it may seem more reasonable to select three 4-genre models instead of four 3-genre models. We observe from our experiments that such alternative is unnecessary since features extracted from individual subsets possess a good capacity for intersubset distinction. Additionally, we also observe that the training of 4-genre subsets is far less effective and less efficient compared with training of 3-genre subsets.

Extensive experiments are also performed towards the selection of CNN network parameters. First is the network layer number. We discover that CNN with more than 3 convolutional layers is exceptionally difficult to train for the network convergence will easily get trapped in local minimas. On the other hand, CNNs with less than 3 convolutional layers do not have sufficient capacity for music classification. The convolution/subsampling size is set at 10/4 for similar criteria. Larger convolutional sizes are difficult to train, while smaller ones are subjected to capacity limitation. To determine the feature map numbers in the three convolutional layers, we first set the three parameters sufficiently large, then watch the performance of CNN as we gradually reduce the number. We discover that 3, 15 and 65 is the optimal feature map numbers for the first three convolutional layers. Reducing them further will drastically constrain the capacity of CNN feature extractors.

4.3 Evaluation

After obtaining 4 CNNs as described above, we apply the feature extractors on the full dataset to retrieve musical pattern features. We deliberately reserve 20% songs in the training of CNN as to examine the ability of our feature extractors on unseen musical data. The musical pattern features are evaluated using various models in the WEKA machine learning system [17]. We discover that the features scored very well in the 10-genre training evaluation, using a variety of tree classifiers such as J48, Attribute Selected Classifier, etc. The classification accuracy is 84% before the majority voting, and gets even higher afterwards. Additionally, musical excerpts not used in CNN training have minor difference in classification rate compared with excerpts used to train CNNs. This provides evidence to support our hypothesis in Section 3 that the variations of musical patterns in the form of MFCC is similar to those of image so that CNN can be used to automatically extract them. In addition, those patterns provide useful information to distinguish musical genres.

However, further experiments on the splitted test dataset give very poor performance compared with the training evaluation; the accuracy of below 30% is therefore too low to make any reliable judgements. It reveals that our current musical pattern extraction model has the deficiency in generalizing the musical patterns learnt to unseen musical data. We further study such phenomenon and found that the reason is two-fold: 1. Musical data is typically aboundant in its variation, and therefore it is hardly sufficient for 80 songs to represent all types of variations in one specific genre; 2. The MFCC feature is sensitive to the timbral, temple and key variation of music which further accentuates the shortage in training data.

One practical solution to these problems above is to enlarge the training dataset by adding affine transforms of songs, such as key elevation/lowering, slight tempo shift, etc. Additional data smooths the variation within one genre and boosts the overall generalizability. Similar work can be found in [16]. Alternatively, the MFCC feature input can be replaced with transforms insensitive to timbral, tempo and key variation, such as mel-frequency spectrum or chroma feature [8].

Our method on musical pattern extractor can be compared with the work in [18], which also applies an image model to audio music genre classification. It is shown that our system possesses better scalability. The texture-of-texture model used in [18] is so highly computational intensive that the authors reduce the training set to 17 songs each category. In comparison our CNN takes less than two hours to obtain feature extractors from a 3-genre, 240-song training set. The efficiency of process can be raised further with parallel computing on different combination of genres.

5 Conclusions and Future Work

In this paper we presented a methodology to automatically extract musical patterns features from audio music. Using the CNN migrated from the the image information retrieval field, our feature extractors need minimal prior knowledge to construct. Our experiments show that CNN is a viable alternative for automatic feature extraction. Such discovery lends support to our hypothesis that the intrinsic characteristics in the variation of musical data are similar to those of image data. Our CNN model is highly scalable. We also presented our discovery of the optimal parameter set and best practice using CNN on audio music genre classification.

Our experiments reveal that our current model is not robust enough to generalized the training result to unseen musical data. This can be overcome with an enlarged dataset. Furthermore, replacing the MFCCs with other feature sets such as the Chroma feature set would also improve the robustness of our model. Further application of image techniques are likely to produce fruitful results towards music classification.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

References

[1] Basili, R. and Serafini, A. and Stellato, A. Classification of musical genre: a machine learning approach *Proceedings of ISMIR* 2004

[2] McKay, C. and Fujinaga, I. Classification of musical genre: a machine learning approach *Proceedings of ISMIR* 2004

[3] de León, P.J.P. and Inesta, J.M., I. Musical style identification using self-organising maps *Web Delivering of Music, 2002. WEDELMUSIC 2002. Proceedings. Second International Conference on* p82–89 2002

[4] Tzanetakis, G. and Cook, P. Musical genre classification of audio signals, *IEEE Transactions on speech and audio processing* Volume 10, Number 5, p293–302, 2002

[5] Li, T. and Tzanetakis, G. Factors in automatic musical genre classification of audio signals *IEEE WASPAA*, p143–146, 2003

[6] Lippens, S. and Martens, J.P. and De Mulder, T. and Tzanetakis, G. A comparison of human and automatic musical genre classification *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 4, p233–236, 2004

[7] Bergstra, J. and Casagrande, N. and Erhan, D. and Eck, D. and Kégl, B. Aggregate features and AdaBoost for music classification *Machine Learning*, Volume 65, Number 2, p473–484, 2006

[8] Ellis, D.P.W. Classifying music audio with timbral and chroma features *Dins Proc. ISMIR* 2007

[9] Lidy, T. and Rauber, A. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR05)* p34–41

[10] Lidy, T. and Rauber, A. and Pertusa, A. and Inesta, J.M. Improving genre classification by combination of audio and symbolic descriptors using a transcription system *Proc. ISMIR, Vienna, Austria* 2007

[11] Anders Meng, Peter Ahrendt, Jan Larsen. Improving Music Genre Classification by Short-time Feature Integration. *IEEE International Conference on Acoustics, Speech, and Signal Processing* , 2005

[12] Pachet, F. and Aucouturier, J.J. Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 2004

[13] Movshon, JA and Thompson, ID and Tolhurst, DJ Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of Physiology* Volume 283, Number 1, p53, 1978

[14] Bengio, Y. and LeCun, Y. Scaling learning algorithms towards AI *Large-Scale Kernel Machines* 2007

[15] Spall, J.C *Introduction to stochastic search and optimization: estimation, simulation, and control* 2003, John Wiley and Sons

[16] Simard, P.Y. and Steinkraus, D. and Platt, J. Best practices for convolutional neural networks applied to visual document analysis *International Conference on Document Analysis and Recogntion (ICDAR), IEEE Computer Society, Los Alamitos* p958–962, 2003

[17] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.

[18] Deshpande, H. and Singh, R. and Nam, U. Classification of music signals in the visual domain *Proceedings of the COST-G6 Conference on Digital Audio Effects* 2001

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)