Local-feature-map Integration Using Convolutional Neural Networks for Music Genre Classification

Toru Nakashika[1], *Christophe Garcia*[2], *Tetsuya Takiguchi*[1]

[1]Department of System Informatics, Kobe University, 1-1 Rokkodai, Kobe, Japan
[2]LIRIS, CNRS, Insa de Lyon, Villeurbanne, France

nakashika@me.cs.scitec.kobe-u.ac.jp, christophe.garcia@liris.cnrs.fr, takigu@kobe-u.ac.jp

Abstract

A map-based approach, which treats 2-dimensional acoustic features using image analysis, has recently attracted attention in music genre classification. While this is successful at extracting local music-patterns compared with other frame-based methods, in most works the extracted features are not sufficient for music genre classification. In this paper, we focus on appropriate feature extraction and proper classification by integrating automatically learnt image feature. For the musical feature extraction, we build gray level co-occurrence matrix (GLCM) descriptors with different offsets from a short-term mel spectrogram. These feature maps are integratively classified using convolutional neural networks (ConvNets). In our experiments, we obtained a large improvement of more than 10 points in classification accuracy on the GTZAN database, compared with other ConvNets-based methods.

Index Terms: music genre classification, music information retrieval, music feature extraction, convolutional neural networks

1. Introduction

Recently, automatic music genre classification has become more important as digital entertainment industry developed and music contents have been widely used. In order to search proper music from enormous databases, it is necessary to assign the labels to each music beforehand. A music genre classification system assigns them automatically instead of being manual efforts.

Feature extraction from an acoustical music signal is a significant step in automatic music genre classification. Most systems in the early years mainly relied on timbre features extracted from a windowed short signal, such as MFCC, STFT, LPC, Filterbank Coefficients and Autoregressive Model [1, 2, 3]. Other methods employed statistical models of the timbre features such as histograms, means, variances, etc [4, 5, 6]. These approaches, however, extract the features frame-by-frame and do not capture the temporal information. As mentioned in [7], spectral transition in short term is considered to be an important factor for genre classification as well as timbre features of the frame.

Meanwhile, a map-based approach, which extracts 2-dimensional features from a piece of the signal and treats with them using image analysis, has been gathering more attention in recent years. Tom *et el.* [8] adopted a 2-dimensional MFCC map for musical features and classified using convolutional neural networks (ConvNets), which is widely used in image analysis tasks such as face detection [9, 10, 11, 12]. In a spoken language indentification task as well, ConvNets have been used as a classifer, whose input was a spectrogram [13]. Although the time-MFCC map or the spectrogram seem to be efficient as musical features, they do not feed efficiently with the ConvNets classifier; the ConvNets is sensitive with the position of the input map. For instance, assuming that we have 2 MFCC maps where the same music pattern (melody) appears at different time in the windowed signal, the classifer regards them as different patterns, even though they should be categorized into the same class.

Another map-based method can be found in [14], where a spectrogram from a piece of the audio is regarded as a texture image, inspired by works in image processing. The system extracts 7 statistical texture features after calculating a gray level co-occurrence matrix (GLCM) [15] from each spectrogram, and classifies the music signals using Support Vector Machine (SVM). While the GLCM feature has the advantage of being robust to small shifts in time, it cannot capture a structure of transposed music properly.

Following the above considerations, in this paper we adopt multiple GLCM maps with different parameters from a time-mel spectrogram (mel map) as musical features for genre recognition. Each GLCM map is not only robust to time shift and transpose of music, but can also capture different characteristics of musical patterns; some maps are more suited for a specific musical genre, and others for another genre. These feature maps are fused and classified using a multiple-input ConvNets. Therefore, it is expected that each map contributes to the classification accuracy.

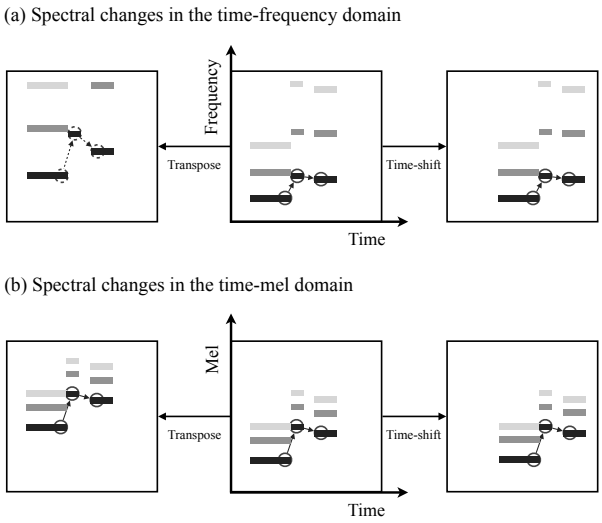

Figure 1: Comparison of spectral changes with time and pitch shifts. Each image represents a spectrogram in (a) or a mel map in (b) of a base melody (middle), a time-shifted melody (right) or a transposed melody (left). The circles in an image indicate spatial relationship of the musical tones.

2. Local feature map — GLCM

In this paper we attempt to extract musical patterns for genre classification based on gray level co-occurrence matrix (GLCM) [15], which is a one of well-known texture descriptors in image analysis, from a short-term low-resolution-in-time mel map. The GLCM encodes how often different combinations of gray levels between two pixels occur in an image. In our experiments we quantized mel maps with 16 levels (gray levels: 0∼15), and only use gray levels of 1 to 15 for the GLCM calculation in order to concentrate on note events rather than rest events. The spatial relationship between pixels is defined in terms of distance d and angle θ. In our approach, various GLCMs with different parameters cooperatively capture local music patterns: spatial relationships between musical tones in time-mel plane. We tried several values of d with θ fixed in pre-experiments, and $d = 1$ was best performed. This is understandable considering that bigrams or a left-to-right HMM, where adjacent two elements are connected, achieve great success in natural language processing or in speech recognition.

The GLCM calculated from a mel map has more efficient characteristics for genre classification than GLCM from a normal spectrogram. Suppose that we have 3 musical patterns in a windowed signal: a base melody, a time-shifted and a transposed version (Figure 1). These 3 examples are musically the same patterns and should be regarded as the same genre. The GLCM from a spectrogram, however, could misrecognize the transposed melody as a different genre because the spatial relationship between musical tones varies in the spectrogram. This is not the case with a GLCM from a mel map.

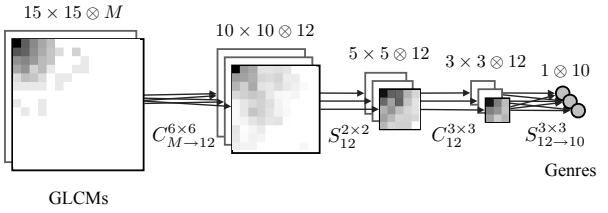

Figure 2: The proposed ConvNets architecture. $C^{p\times q}_{m\to n}$ and $C^{p\times q}_{m}$ represent convolutional operations with convolution kernels of size $p \times q$. $S^{p\times q}_{m\to n}$ and $S^{p\times q}_{m}$ are subsampling operations with $p \times q$ kernels. The layers corresponding to $C^{p\times q}_{m\to n}$ or $S^{p\times q}_{m\to n}$ are fully connected; otherwise connected 1 by 1. $i \times j \otimes k$ above each layer means that the layer has k maps of size $i \times j$.

3. Integrative classifier — ConvNets

Convolutional Neural Networks (ConvNets), proposed by LeCun et al. [9], have shown great performances in various computer vision applications, such as handwritten character recognition [9], facial analysis [10], videoOCR [11], or vision-based navigation [12]. ConvNets consist of a pipeline of convolution and pooling operations followed by a multi-layer perceptron. They tightly couples local feature extraction, global model construction and classification in a single architecture where all parameters are learnt conjointly using back-propagation.

The proposed model of ConvNet, illustrated in Figure 2, is designed for multiple inputs of M different GLCM maps, based on convolutional filters layers interspersed with non-linear activation functions and spatial feature pooling operations (sub-sampling layers). Convolutional layers $C^{p\times q}_{m\to n}$ (or $C^{p\times q}_{m}$) contain a given number of planes. Each unit in a plane receives input from a small neighborhood (local receptive field) in the planes of the previous layer. Each plane can be considered as a feature map that has a fixed feature detector, that corresponds to a convolution with a trainable mask of size $p \times q$, applied over the planes in the previous layer. A trainable bias is added to the results of each convolutional mask, and a hyperbolic tangent function, used as an activation function, is applied. Multiple planes are used in each layer so that multiple features can be detected. Once a feature has been detected, its exact location is less important. Hence, each convolutional layer $C^{p\times q}_{m\to n}$ is typically followed by a pooling layer $S^{p\times q}_{m}$ that computes the average values over a neighborhood $p \times q$ in each feature map, multiplies it by a trainable coefficient, adds a trainable bias,

Table 1: Features and GLCM parameters for validation.

Feature map	Base map	M	d	θ
i-GLCM	mel map	4	-	-
GLCM(a)	mel map	1	1	0°
GLCM(b)	mel map	1	1	45°
GLCM(c)	mel map	1	1	90°
GLCM(d)	mel map	1	1	135°
s-GLCM(a)	spectrogram	1	1	0°
MFCCM	-	1	-	-

and passes the result through an activation function.

The last layer is a classical perceptron of 10 neurons, that outputs likelihoods for the corresponding genres with values between 0 and 1. The genre with the highest likelihood is adopted as the classification result for the corresponding input block of the signal.

The proposed ConvNet is trained in a supervised way with the classical error backprop algorithm that minimizes the mean square error between obtained and desired outputs over the training set.

4. Experiments

4.1. Setup

We conducted 10-musical-genres-classification experiments using GTZAN dataset [16], which is widely used in this task. The dataset contains 100 songs for each of the following musical genres: Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae and Rock. Each song is recorded during 30 seconds with a sampling rate of 22050 Hz at 16 bit. In our experiments we use randomly-selected 90 songs from each genre for the training set (in total 900 songs) and the rest for validation (100 songs).

To evaluate the effectiveness of integrated GLCM features of a mel map ("i-GLCM"), we made a comparison with various features as shown in Table 1. The "i-GLCM" consists of 4 different GLCM maps ("GLCM(a)∼(d)" for different angles θ). For the GLCM calculation, we divided the signal into short-term pieces of 4 seconds with 2 seconds overlapping (we got 14 pieces from a signal). Then, the mel maps for each piece were calculated with a frame-length of 125 ms without overlap and filterbank-channels of 40 (the size of the map is 40×32). For "s-GLCM(a)" the spectrogram was calculated with a frame-length of 186 ms with 50% overlap.

We also compared to the accuracy of time-MFCC map ("MFCCM"). In our experimental settings, each 40-coefficients MFCC frame of length 40 ms with 50% overlap was obtained from the signal. This map was divided into 30 sub-maps, each of which spans 1 second (of size 40×50), for training and validation.

In our experiments ConvNets for "i-GLCM" and for individual GLCM map have the same architecture as described in Figure 2. For "MFCCM", the ConvNets architecture has the following layers: $C^{10\times10}_{1\to12}$, $S^{4\times4}_{12}$, $C^{3\times3}_{12\to12}$ and $S^{2\times3}_{12\to10}$.

Since each data is short term, there occur difficult-to-classify inappropriate data such as a blank map or a few-notes map. To avoid this we evaluate each method through a majority voting scheme. The final class assigned to a song is the one which was found for the majority of the blocks in the song.

Table 2: Classification accuracy of each method.

Feature map	Acc. (%)	Adapt. (%)	MSE
i-GLCM	**72.00**	**53.42**	**0.246584**
GLCM(a)	43.00	40.66	0.292312
GLCM(b)	36.00	34.17	0.313291
GLCM(c)	59.00	41.97	0.286872
GLCM(d)	38.00	34.10	0.313272
s-GLCM(a)	37.00	42.19	0.296918
MFCCM	60.20	48.51	0.277792

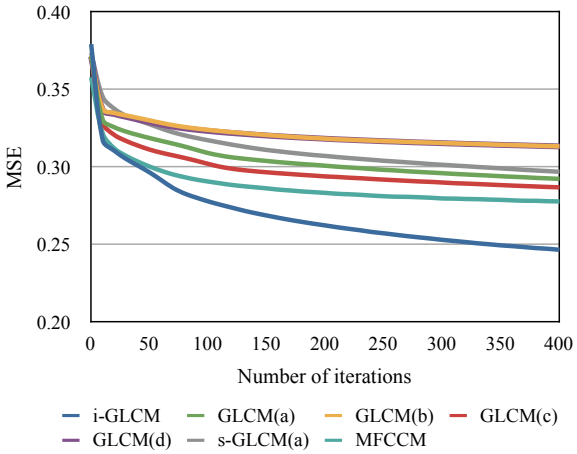

Figure 3: MSE curve over 400 iterations.

4.2. Evaluation

Experimental results are summarized in Table 2 with 3 measures: classification accuracy by majority voting on validation set ("Acc."), block-level accuracy (adaptation) on training set ("Adapt.") and mean square error for validation set ("MSE") after 400 iterations of ConvNets. Figure 3 illustrates the convergence of the mean square error on training. As shown in Table 2 and Figure 3, integrated GLCM maps achieved desirable performance on all measures not only compared to a mfcc map but also compared to any other individual GLCM maps. It is especially notable that "i-GLCM" overcomes "MFCCM" on accuracy in spite of having smaller number of feature dimensions. We can say that these interesting results are

Table 3: Adaptation for each genre (%).

Feature map	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
i-GLCM	**38.73**	**92.30**	**33.49**	28.02	**63.17**	**67.22**	85.32	32.46	**61.19**	**32.30**
GLCM(a)	30.40	87.30	22.94	30.48	37.70	42.70	79.13	11.90	45.63	18.41
GLCM(b)	15.24	35.08	15.48	35.48	24.92	43.73	85.24	14.05	42.62	29.84
GLCM(c)	23.81	83.10	16.83	**38.73**	51.27	44.84	84.29	**36.27**	24.76	15.79
GLCM(d)	13.57	45.95	11.03	34.21	25.48	45.48	**85.71**	9.05	43.41	27.14
Avg.	20.76	62.86	16.57	34.73	34.84	44.19	83.59	17.82	39.11	22.80

due to the fact that the multiple GLCM maps, which capture different musical patterns individually, are compatibly integrated with our ConvNets model that internally fuses and classifies 2-dimensional multiple features.

The accuracy of 72 % is high, being close to the accuracy of 76 %, pointed out by McKay [17] obtained when human beings correctly classify music songs.

Comparing "GLCM(a)" and "s-GLCM(a)" in Table 2 and Figure 3, we obtained better results with "GLCM(a)". As mentioned before, we believe this is because a mel map has the advantage of being robust to transposed music patterns.

To examine the results of our approach in more details, we list in Table 3 block-level accuracies of integrated GLCMs, individual GLCMs, and the average of individual GLCMs ("Avg.") for each genre. According to the table, integrated GLCMs performed better than the average of individual GLCM on all genres except for Disco. It should be noted that weak points of individual GLCMs are improved when they are fused together. For example, "GLCM(c)" map is strong in Pop but weak in Rock, while "GLCM(b)" is strong in Rock but weak in Pop. The "i-GLCM" cooparatively compensates such shortcomings with the talented-in-particular-genre individual maps.

When it comes to the relatively-low-accuracy genres, the songs in Disco, Hiphop and Pop are very close and difficult to distinguish from each others even by human beings. Hence, the system was confused; most of the songs of Disco and Pop were classified as Hiphop, the dominant genre in such musical patterns. This is the same case with the pair of Rock and the dominant Metal.

5. Conclusion

In this paper, we presented effective features and a fusion method for automatic musical genre classification. Focusing on spatial relationship between musical tones in terms of image analysis, we used GLCM maps, each of which captures different musical patterns on a mel map. These different maps are automatically fused through a convolutional neural network (ConvNet) framework. Our experiments using various inputs for the ConvNets showed that an integrated model of GLCMs best performed incorporating the benefits of individual GLCM maps. We believe that the strong point of the integrative architecture of ConvNets with appropriate multiple feature maps may be applied to speech recognition and other audio signal classification problems.

6. References

[1] G. Tzanetakis, "Musical Genre Classification of Audio Signals," IEEE Trans. Speech and Audio Pro., 10(5):293-302, 2002.

[2] Z. Fu et el., "Learning Naive Bayes Classifiers for Music Classification and Retrieval," International Conference on Pattern Recognition 2010, 4589-4592, 2010.

[3] T. Langlois and G. Marques, "A Music Classification Method Based on Timbral Features," International Society for Music Information Retrieval Conference 2009, 81-86, 2009.

[4] A. Meng et el., "Improving Music Genre Classification by Short-time Feature Integration," IEEE ICASSP 2005.

[5] S. Lippens et el., "A comparison of human and automatic musical genre classification," IEEE ICASSP 2004, 4:233-236, 2004.

[6] T. Lidy, A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," Proc. ISMIR05, 34-41, 2005.

[7] Y. Tsuji et el., "The Estimation of Music Genres Using Neural Network and its Educational Use," International Conference on Computer-Assisted Instruction 2000, 1:158-162, 2000.

[8] Tom LH. Li et el., "Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network," IMECS 2010, 1, 2010.

[9] Y. Lecun et el., "Gradient-based learning applied to document recognition," Proc. of the IEEE, 1998.

[10] C. Garcia and M. Delakis, "Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection," Pattern Analysis and Machine Intelligence, 2004.

[11] M. Delakis and C. Garcia, "Text detection with Convolutional Neural Networks," Proc. of the Int. Conf. on Computer Vision Theory and Applications, 2008.

[12] R. Hadsell et el., "Learning long-range vision for autonomous off-road driving," Journal of Field Robotics, 2009.

[13] Gregoire Montavon, "Deep learning for spoken language identification," NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

[14] Costa, Y.M.G. et el., "Music Genre Recognition Using Spectrograms," IWSSIP 2011, 151-154, 2011.

[15] B. Hua et el., "Research on Computation of GLCM of Image Texture," 2006.

[16] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Audio and Speech Processing, 10(5), 2002.

[17] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?," 7th International Conference for Music Information Retrieval, 2006.