Convolutional Neural Networksを用いた局所特徴統合による自動音楽ジャンル分類*

☆中鹿亘, Garcia Christophe (INSA Lyon), 滝口哲也, 有木康雄 (神戸大)

1 はじめに

近年のコンピュータの発展とともに音楽のデジタルコンテンツが爆発的に増大し，web 上や個人の情報端末上で音楽データを整理・検索することが困難になってきている．このような背景の中で，類似した音楽を自動的にクラスタリングする自動音楽ジャンル分類の研究が盛んに行われている．

一般に，自動音楽ジャンル分類では使用する特徴量と識別器の選択が重要となる．従来では音楽信号の各フレームから MFCC や LPC，自己回帰モデルなどの特徴量を抽出するアプローチが主流であった [1, 2]．しかしながらこういったアプローチはフレームごとに特徴量を抽出しており，フレーム間の関連性を捉えることは困難である．そこで近年，フレームごとに特徴量を抽出するのではなく，数秒程度の部分信号から時間-周波数の 2 次元特徴（特徴マップ）を抽出し，画像処理の技術を用いて音楽ジャンルを識別するアプローチが研究されている [3, 5]．

本稿では後者のアプローチに基づき，各マップから計算される画像特徴である GLCM (Gray Level Co-occurrence Matrix) [6] を特徴量とし，Convolutional Neural Networks (ConvNets) [4] を用いて複数の GLCM を統合しつつ音楽ジャンルを識別する手法を提案する．

2 局所特徴量 — GLCM

本研究では，音楽ジャンルを識別する特徴量として GLCM(Gray Level Co-occurrence Matrix) マップを用いる．GLCM とは画像処理の分野でよく知られているテクスチャ特徴であり，画像中の 2 ピクセル間の輝度 (グレーレベル) 変化の頻度を行列として表現したものである．従来研究の中には，短時間スペクトログラムから GLCM を計算するもの [5] もあったが，本稿では，隣接する数フレーム分のメルケプストラム（メルマップ）から GLCM を計算する．Fig. 1 は，ある音楽パターンのスペクトログラム (a) とメルマップ (b) を比較したものである．本アプローチではマップ中の，ある 2 ピクセル間のグレーレベル変化 (図中の丸) の頻度で音楽パターンを抽出することで，音楽ジャンルを識別することを試みる．中央の画像が元の音楽パターン，右が時間シフトしたもの，左が移調 (音高シフト) したものを示しており，いずれも音楽パターンとしては同じだと考えられるが，周波数軸が線形なスペクトログラムでは，音高シフトした場合に 2 ピクセル間の位置関係が異なってしまう (図中の点丸)．一方周波数軸が対数的なメルマップではそういった問題は生じず，正常にパターンを抽出することができると期待される．

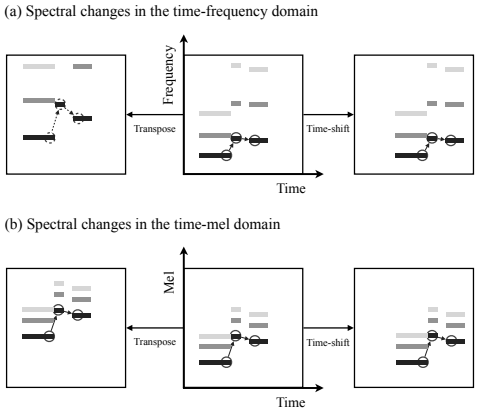

Fig. 1 Comparison of spectral changes with time and pitch shifts. Each image represents a spectrogram in (a) or a mel map in (b) of a base melody (middle), a time-shifted melody (right) or a transposed melody (left). The circles in an image indicate spatial relationship of the musical tones.

GLCM は距離を表す d，角度を表す θ の 2 つのパラメータを用いて 2 ピクセル間の位置関係を表現する．すなわち，あるパラメータを持つ GLCM は 1 つの音楽パターン (の統計量) を表現している．本研究ではパラメータの異なる複数の GLCM を，次節で述べる Convolutional Neural Networks を用いて統合し，同時に音楽ジャンルを識別する．

3 Convolutional Neural Networks

Convolutional Neural Networks (ConvNets) は，LeCun ら [4] によって提案された多層型ニューラルネットワークの 1 種であり，特に画像処理・パターン認識の分野で効果が示されてる．ConvNets は畳み込み演算を行う層 $C_{m\to n}^{p\times q}$ と，平滑化を行う層 $S_m^{p\times q}$ を交互に重ねることで，計算量を大幅に削減させると

*Local-feature-map integration using convolutional neural networks for music genre classification. by NAKASHIKA Toru, GARCIA Christophe (INSA Lyon), TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe University)

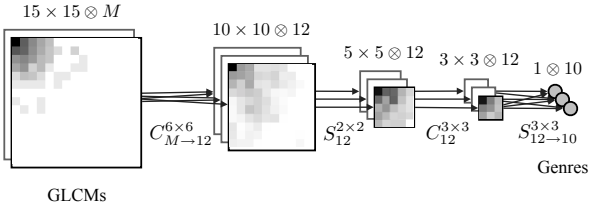

Fig. 2 The proposed ConvNets architecture. $C^{p\times q}_{m\to n}$ and $C^{p\times q}_{m}$ represent convolutional operations with convolution kernels of size $p \times q$. $S^{p\times q}_{m\to n}$ and $S^{p\times q}_{m}$ are subsampling operations with $p \times q$ kernels. The layers corresponding 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 $i \times j$.

同時に，効果的に 2 次元的な情報を集約させるといった特徴を持つ．また，異なる複数の特徴量を入力させることも可能である．本研究では，Fig. 2 のような構造を持つ ConvNets を用いて，抽出された複数の GLCM から音楽ジャンルを識別する．

4 評価実験

Pop や Jazz，Metal など，10 クラスの音楽信号を含む GTZAN 音楽データベース [7] を用いて，音楽ジャンル識別の評価実験を行った．このデータベースにはクラスごとに 30 秒の楽曲が 100 曲用意されており，本研究ではこのうち 90 曲を学習用，残り 10 曲を評価用に用いた．

提案手法による認識率の向上を確認するため，本実験では Table 1 に示す 7 つの手法により識別率を比較した．“i-GLCM” はパラメータの異なる 4 つの GLCM マップ (“GLCM(a)∼(d)”) を ConvNets により統合したものである．“i-GLCM”，“GLCM(a)∼(d)” はいずれもメルマップから GLCM を計算したのに対し，“s-GLCM(a)” はスペクトログラムから計算させている．“MFCCM” は GLCM を用いず，数フレーム分の MFCC マップを ConvNets に入力して音楽ジャンルを識別する．また，事前に予備実験として様々なパラメータ d, θ を試したが，$d = 1$ が最もよい結果を示していたため，評価実験においても $d = 1$ が用いられている．これは，自然言語処理において bigram が効果的にはたらくのと同様，隣接する要素間の関係性が重要であることを示唆していると考えられる．

各手法による音楽ジャンルの識別結果を Table 1 の右に示す．“Acc.” は 10 クラスの平均分類率，“MSE” は 400 回のイテレーション後の ConvNets の平均二乗誤差を表している．この結果から分かるように，複数の GLCM を統合した “i-GLCM” から最もよい識別結果が得られた．これは，それぞれの GLCM(a)∼(d) が Pop や Jazz など特有のジャンル識別に有効な特徴を抽出し，それらがうまく統合されたからだと考えられる．また，GLCM(a) と s-GLCM(a) を比較しても，メルマップから GLCM を抽出した GLCM(a) が高い分類率を示した．これは，メルマップから GLCM を計算すれば，音高シフトした場合でも正常にパターンを捉えることができるためであると考えられる．

Table 1 Parameters of each method and their classification accuracies.

Methods	M	d	θ	Acc.	MSE
i-GLCM	4	-	-	**72.00**	**0.246584**
GLCM(a)	1	1	0°	43.00	0.292312
GLCM(b)	1	1	45°	36.00	0.313291
GLCM(c)	1	1	90°	59.00	0.286872
GLCM(d)	1	1	135°	38.00	0.313272
s-GLCM(a)	1	1	0°	37.00	0.296918
MFCCM	1	-	-	60.20	0.277792

5 おわりに

本稿では，メルマップから GLCM を計算することで，音高シフト・時間シフトに頑健な音楽パターンを抽出し，また Convolutional Neural Networks (ConvNets) を用いて様々な GLCM（パターン）を統合させることで精度を向上させる自動音楽ジャンル分類手法を提案した．この両者の有効性を評価実験により示した．

参考文献

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

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

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

[6] B. Hua et el., “Research on Computation of GLCM of Image Texture,” Acta Electronica Sinica, 2006.

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