H-015

CNN を用いた手書き楽譜の音楽記号認識
Handwritten musical symbols recognition using CNN

榆木 徹† 浅井 紀久夫‡
Toru Niregi Kikuo Asai

1. はじめに

作曲や編曲は、専門家、アマチュアを問わず、五線紙に音符や音楽記号等を手書きですることで行われてきた。電子媒体への移行に際して、楽譜のデジタル形式への変換によるコンピュータ音楽再生が期待される。そのため、楽譜の光学楽譜認識（OMR）が行われてきた[1]。最近では印刷された活字体の楽譜だけではなく、タブレット PC で手書きした楽譜をデジタル形式に変換する市販 OMR ソフト[2]が存在する。また、オープンソースのソフト[3, 4]も提供されている。しかし、手書き楽譜にはカスレや汚れ、紙媒体を電子媒体に変換した際のゆがみ等が含まれ、認識精度が低下する。

オフライン手書き楽譜の認識では、音楽記号の認識に対して五線除去後の楽譜画像に含まれる連結成分から加重方向ヒストグラム特徴を抽出し、最近傍法により 3 つのクラス（音符、休符、その他音楽記号）に分類した[5]。また、疑似ベイズ識別関数を用いた音楽記号の分類では 90 %を超える分類成功率が示された[6]。さらに、携帯端末の普及によって、こうした手書き楽譜の音楽記号をオンラインで認識するシステムが開発されるようになった[7, 8]。手書き譜面を認識してデジタル形式に変換した後、その譜面に基づいて演奏を行うシステムも提案されている[9]。

近年、画像認識の手法として、CNN（畳み込みニューラルネットワーク）を用いた深層学習による分類手法が注目されてきた。CNN は認識に有効な特徴量を抽出することによって、画像認識や音声認識などの様々な応用に対してその有用性が認められている。手書き楽譜の音楽記号の認識に CNN を適用する試み[10, 11]も行われている。特に[11]では認識精度を改善するために、Convolutional-Recurrent Neural Network を構成したり、data augmentation や transfer learning を導入したりしている。

本研究では、オフラインで手書き楽譜の認識を行うため、畳み込みニューラルネットワーク（以下、CNN）を用いた。まず、7 層 CNN を構成し、データセット HOMUS の音楽記号 31 種およびデータセット CVC-MUSCIMA に独自収集した音楽記号を加えた 28 種に対して音楽記号の分類を行った。次に、事前学習済みネットワーク Inception-V3 を用いた転移学習を行い、データセット CVC-MUSCIMA および独自収集した音楽記号 28 種に対して音楽記号を分類した。

2. 手書き楽譜の認識

まず、データセット HOMUS [HOMUS]の音楽記号 31 種およびデータセット CVC-MUSCIMA [CVC]に独自収集した音楽記号を加えた 28 種に対して 7 層 CNN を構成し、音楽記号の分類を行った。次に、データセット CVC-MUSCIMA および独自収集した音楽記号 28 種に対して事前学習済みネットワーク Inception-V3 を用いた転移学習を適用し、音楽記号を分類した。

2.1 7 層 CNN による分類

7 層 CNN（畳み込み層、プーリング層、畳み込み層、畳み込み層、プーリング層、全結合層、全結合層）を用いて、HOMUS の音楽記号および CVC-MUSCIMA に独自収集した音楽記号を加えたデータセットを作成し、手書き楽譜の音楽記号を分類する。

HOMUS は 100 人の記譜者による手書きデータの画像 14,750 枚が 32 のクラスに分類されており、手書き楽譜認識の研究に広く使われている。本研究では HOMUS の 32 クラスのうち Dot のクラスを排除し、31 クラスを使用することとした。また、HOMUS に含まれる音楽記号の画像に加えて、CVC-MUSCIMA および独自に収集した音楽記号の画像 194 枚を含めたデータセットとし、合計 14,944 画像を用意した。音楽記号のクラスは 31 とした。CVC-MUSCIMA 自体には、記譜者 50 人による楽譜画像 1,000 枚が格納されている。記譜には同じペン、同種の楽譜が使われ、8 bit グレースケールの画像になっている。

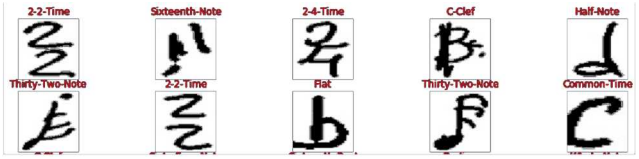


図 1 7 層 CNN による分類で使用したデータの例

表 1 に、7 層 CNN の設定を示した。Conv2D は畳み込み層、MaPooling は最大値プーリング、Dense は全結合層を表す。入力層への画像入力サイズは50 × 50画素である。学習過程で、画像の回転（左右角度 30 度）およびシフト（上下左右 20%）による水増しを行った。k-分割交差検証（k = 5）により学習と評価を行った。

表 1 7 層 CNN の設定

Conv2D	(50, 50, 16)	160
MaxPooling	(25, 25, 16)	0
Conv2D	(25, 25, 128)	18560
Conv2D	(25, 25, 256)	295168
MaxPooling	(12, 12, 256)	0
Dropout	(12, 12, 256)	0
Flatten	(36864)	0
Dense	(128)	4718720
Dropout	(128)	0
Dense	(31)	3999

† 放送大学 The Open University of Japan
‡ 放送大学 The Open University of Japan

2.2 転移学習による分類

事前学習済みネットワーク Inception-V3 を用いて、CVC-MUSCIMA および独自収集した音楽記号をデータセットとして分類する。

データセット CVC-MUSCIMA および独自収集した音楽記号の画像は全部で 194 枚であるが、画像の回転およびシフトによる水増しを行い、全データ数を 2,107 とした。この水増し処理には、音楽記号の画像を左右に 30 度回転させたり、上下左右に 20 %シフトさせたりする操作を行った。また、クラス数を 28 とした。図 2 に、7 層 CNN による分類で使用したデータの例を示す。

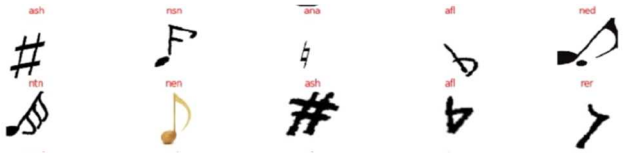


図 2　転移学習による分類で使用したデータ例

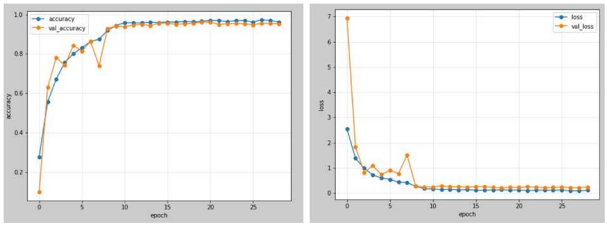
事前学習済みネットワークとして Inception-V3 [14]を使い、楽譜データセットで再学習を実施した。Inception-V3 は 1,000 クラスの画像分類を行うよう学習されたモデルで、Inception モジュールと呼ばれる小さなネットワーク（複数の畳み込み層やプーリング層から構成される）を組み込んでいるのが特徴である。CNN は一般に浅い層で汎用的な特徴を抽出し、深い層でデータに特化した特徴を抽出する傾向がある。そのため、全 314 層のうち、入力層から 250 層までを修正せず、251 層以降の層（出力層に近い層）を修正し、fine-tuning した。入力層への画像入力サイズは 139 × 139画素である。k-分割交差検証（k = 5）により学習と評価を行った。

3. 結果

7 層 CNN による分類および転移学習による分類では、用意した全データのうち、80 %を学習時に使用するデータとし、残り 20 %をテスト時に使用するデータとした。学習時に使用するデータはその 80 %を訓練データとし、20 %を検証データとした。

7 層 CNN による分類では、学習時に使用しないデータ 2,989 個のうち、正解 2,783、誤り 206 を得た。正解率は 93.1 %を得た。

図 3 に、転移学習による分類の学習曲線の一例を示す。10 epoch を過ぎた辺りで定常状態に達していることがわかる。処理には 25 epoch で約 40 分程度かかっている。



(a) 正確度　　(b) 損失関数

図 3　転移学習による分類の正確度と損失関数

転移学習による分類では、学習時に使用しないデータ 422 個のうち、正解 413、誤り 9 を得た。正解率は 97.9 %を得た。

学習に使用したデータセットの数は異なるが、7 層 CNN による分類に比べ、事前学習済みネットワーク Inception-V3 を使い、楽譜データセットで fine-tuning を行った転移学習による分類では音楽記号の認識が大きく改善しており、大規模なデータセットによる事前学習の重要性が改めて認識された。

五線譜上の位置と音の高さの関係を規定する音部記号の認識精度が高い傾向があり、その特徴が捉えやすいことが考えられる。また、7 層 CNN による分類では音符の画像サイズを50 × 50画素に設定したが、音符の細かい特徴を捉えるには不十分だった可能性がある。

4. まとめ

CNN を用いた手書き楽譜の音楽記号認識を行った。その結果、7 層 CNN による分類において正解率 93.1 %を得た。また、転移学習による分類において正解率 97.9 %を得た。転移学習では大きなデータセットで事前学習を行っており、そうした事前学習済みネットワークを使うことによって認識率の向上が期待される。今後、多様なデータセットを用いて手書き楽譜認識の改善を図りたい。また、今回はオフラインで動作する仕組みを採用したが、オンラインでの利用を想定したシステムを構築したい。

参考文献

[1] Jorge Calvo-Zaragoza, Jan Hajič Jr., and Alexander Pacha, “Understanding Optical Music Recognition”, ACM Comput. Surv., Vol.53, No.4 (2020).
[2] Neuratron, “PhotoScore”, https://www.neuratron.com/photoscore.htm (accessed in 2021).
[3] Audiveris, https://github.com/Audiveris/audiveris (accessed in 2021).
[4] Gamera toolkit, https://gamera.informatik.hsnr.de/addons/musicstaves/ (accessed in 2021)
[5] 中川 大樹, 大山 航, 若林 哲史, 木村 文隆, 三宅 康二, “手書き楽譜認識のための音楽記号の抽出・分類”, 電子情報通信学会技術報告書, Vol.115, No.24 (2015).
[6] 早川 優木, 若林 哲史, 三宅 康二, 大山 航, “オフライン手書き楽譜中の音楽記号の分類と音高認識の研究”, 電子情報通信学会技術報告書, Vol.118, No.111 (2018).
[7] H. Miyao, M. Maruyama, “An online handwritten music symbol recognition system”, International Journal of Document Analysis and Recognition, Vol.9 (2007).
[8] J. Oh, S. J. Son, S. Lee, Ji-Won, Kwon, and N. Kwak, “Online recognition of handwritten music symbols”, International Journal of Document Analysis and Recognition, Vol.20 (2017).
[9] 馬場 哲晃, 菊川 裕也, 串山 久美子, 青木 允, “簡易な手書き譜面を利用した演奏システム Gocen の設計”, 情報処理学会論文誌, Vol.54, No.4 (2013).
[10] Roberto M. Pinheiro Pereira, Caio E.F. Matos, Geraldo Braz Junior, João D.S. de Almeida, and Anselmo C. de Paiva, “A Deep Approach for Handwritten Musical Symbols Recognition”, Proc. Brazilian Symposium on Multimedia and the Web (2016).
[11] Arnau Baró, Pau Riba, Jorge Calvo-Zaragoza, Alicia Fornés, “From Optical Music Recognition to Handwritten Music Recognition: A baseline”, Pattern Recognition Letters, Vol.123 (2019).
[12] HOMUS (Handwritten Online Music Symbols) dataset, https://grfia.dlsi.ua.es/homus/ (accessed in 2021)
[13] Alicia Fornés, Anjan Dutta, Albert Gordo, Josep Lladós, “CVC-MUSCIMA: A Ground-truth of Handwritten Music Score Images for Writer Identification and Staff Removal”, International Journal on Document Analysis and Recognition, Vol.15, No.3 (2012).
[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision”, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).