H-035

YOLOv7 を用いたドラム譜の認識のための音楽記号の分解と統合に関する研究
A Study on Decomposition and composition of Music Symbols for Drum Score Recognition Using YOLOv7

小椋 隆生† Ryusei Ogura　黒木 修隆† Nobutaka Kuroki　沼 昌宏† Masahiro Numa

1. はじめに

カメラやスキャナから取り込んだテキストの自動読み取り技術が Optical Character Recognition（OCR）と呼ばれるのに対して，楽譜の自動読み取り技術[1],[2]は Optical Music Recognition（OMR）と呼ばれる．日本語や英語の文字に比べて音楽記号の種類はそれ程多くはないが，和音や連桁などの組み合わせを考慮すると画像のバリエーションが無限に広がるため，その認識は容易ではない．ドラム譜の場合は黒玉や白玉の音符に加えてシンバルを表す×印や◇印も多用される．×印や◇印についても黒玉と同様に 4 分音符や 8 分音符などの長さのバリエーションがあるため，それら全てを別のシンボルとして機械学習を行うとコストが倍増する．AI で音楽記号を認識する場合は，人が認識しやすい 4 分音符や 8 分音符のような単位よりも，さらに細かい単位である“符頭”，“付幹”，“符尾”，“連桁”に分解して認識する方が効率が上がる可能性がある．一方で，付点 4 分音符や付点 8 分音符のような場合は，“付点”を独立に認識するよりも，“符頭”とセットで新たなシンボルとして認識する方が精度が向上する可能性がある．

本研究では YOLOv7[3]を用いてドラム譜を認識するにあたり，音楽記号を要素シンボルに分解または統合する方法について検討を行ったので報告する．

2. 提案する音楽記号の分解と統合

2.1 概要

まずドラム譜における音符の構成要素を図 1 に示す．一つの音符は主に符頭，符幹，符尾の 3 つの部分から構成される．提案する手法ではこれらを“符幹・符尾”と“符頭”にわけて検出する．また，符頭の横に位置する符点に関しては符頭と合わせて検出する．

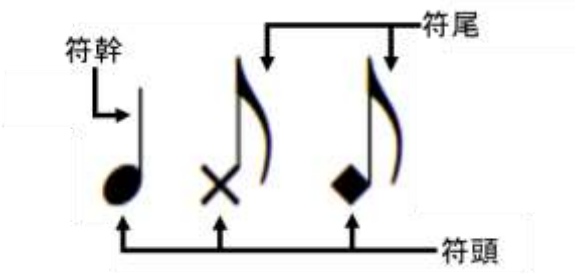


図 1 ドラム譜における音符の構成要素

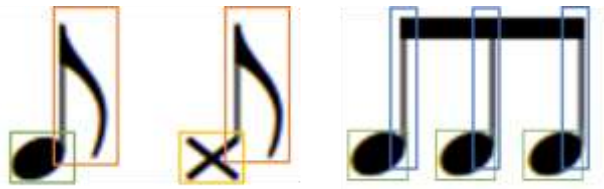
† 神戸大学大学院工学研究科，
Graduate School of Engineering, Kobe University

2.2 要素シンボルの分解と統合

8 分音符の要素シンボルへの分解の例を図 2 に示す．(a)の単独の音符の場合は，符頭と，符幹・符尾を合わせたものの二つに分けて検出する．(b)のように符尾をもつ音符が連続する場合は，連桁と呼ばれる符尾をつなげた形で表されるが，この場合も単独の音符と同様に符頭と符幹・符尾の 2 つに分けて検出する．検出した符幹・符尾については，単独と連桁のものを同一クラスで検出する手法 1 と，別クラスで検出する手法 2 を比較した．

次に符点の検出手法を図 3 に示す．手法 A では符点を単体で検出し，手法 B では符点をその直前の符頭と合わせてひとつのシンボルとして検出する．

検出した符頭と符幹・符尾のバウンディングボックスが重なった場合，それらを一つの音符として統合する．



(a) 単独の 8 分音符　(b) 連桁の 8 分音符

図 2 提案する要素シンボルへの分解

(a) 手法 A　(b) 手法 B

図 3 提案する付点の検出方法

3. 評価実験

提案手法に基づいて音楽記号を検出し，精度評価を行う．

3.1 実験内容

YOLOv7 を用いて各要素シンボルの検出を行う．付点については手法 A と B の精度を比較する．また，符幹・符尾についても単独と連桁を同一クラスとして検出する手法 1 と別クラスで検出する手法 2 の 2 種類を比較する．実験

表 1 付点の検出結果

手法	適合率	再現率	F 値
A		0	
B	0.938	1.00	0.968

表 2 符幹・符尾の検出結果

クラス		適合率		再現率		F 値	
		手法 1	手法 2	手法 1	手法 2	手法 1	手法 2
8 分	単独	0.872	0.806	0.919	0.833	0.895	0.819
	連桁		0.794		0.846		0.819
16 分	単独	0.703	0.749	0.707	1.00	0.705	0.856
	連桁		0.670		0.684		0.677
平均		0.788	0.755	0.813	0.841	0.800	0.793

表 3 要素シンボルの検出結果

要素シンボル	クラス	適合率	再現率	F 値
	ball	0.830	0.829	0.829
	cymbal	0.886	0.870	0.878
	4thflag	0.778	0.668	0.719
	8thflag_single	0.794	0.833	0.813
	8thflag	0.915	0.846	0.879
	16thflag_single	0.749	1.00	0.856
	16thflag	0.670	0.684	0.677
	4threst	0.947	1.00	0.973
	8threst	0.794	0.833	0.813
	16threst	0.887	1.00	0.940
	barline	0.933	0.874	0.903
	dot	0.938	1.00	0.968
	all	0.843	0.870	0.854
	演奏正解率	0.450		

に使用した画像はドラム譜を小節ごとに切り取ったものであり，学習用に 114 枚，評価用に 20 枚使用した．評価指標は適合率，再現率，F 値を用いた．

また，検出した音楽記号のクラスと位置情報を用いて MIDI 信号に変換する自動演奏処理を実装した．

3.2 結果と考察

まず付点と符幹・符尾の検出結果を表 1 に示す．付点については手法 A では付点が全く検出されなかったが,手法 B では F 値が 0.968 の高い精度が得られた．次に，符幹・符尾の検出結果を表 2 に示す．手法 1 と 2 で平均の F 値はどちらも 0.8 程度であり，大きく変わらなかった．

すべての要素シンボルの検出結果を表 3 および図 4，図 5 に示す．平均の検出精度については 0.8 以上の高い精度が得られたが, 個別にみると 4 分の符幹の再現率が 0.668, 16 分の連桁においては適合率と再現率ともに 0.7 未満の低い値が得られた．4 分の符幹のような形状に特徴が少ないものについては, 要素シンボルへの分解を行わない方が良い可能性がある．また，16 分の連桁に関しては 8 分と誤認識する例が多い．16 分の旗は 8 分の旗が上下に重なっているとも解釈できるため, このような誤認識が起きたと考えられる．そこで, 改善策として連桁の形状をより細かい要素シンボルへと分解することが考えられる．

なお，自動演奏の精度については 20 枚中 9 枚で完全に正しく演奏することが確認できた．

4. まとめ

本研究ではドラム譜の認識における精度向上のため, 音楽記号を要素シンボルへの分解・統合する手法を提案した．結果として F 値が平均で 0.854 の高い検出精度を得られることが確認できた．今後は 8，16，32 分音符における精度向上策として, 符尾に対するさらに細かいクラス分けや要素への分解などを行い，精度の比較を行う予定である．

参考文献

[1] Z. Huang, X. Jia and Y. Guo, “State-of-the-Art Model for Music Object Recognition with Deep Leaning,” Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China, 2019

[2] “楽譜カメラ”, https://cm.kawai.jp/products/mc/

[3]] C. Wang, A. Bochkovskiy and H. Mark Lia, “Y OLOv7: Trainable bag-of-freebies sets new state -of-the-art for real-time object detectors,” arXiv: 2207.02696, 2022

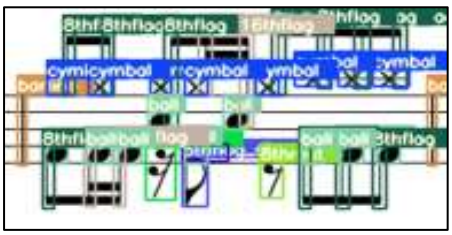
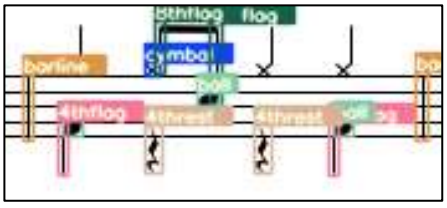
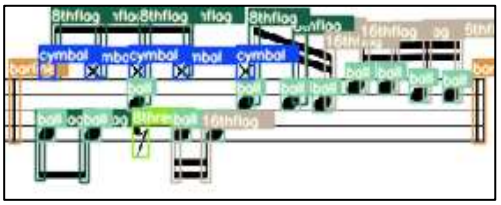


図 4 要素シンボルの正検出例



(a) 4 分の符幹の未検出



(b) 16 分の連桁の誤検出

図 5 要素シンボルの未検出・誤検出例