修 士 論 文

Mask R-CNN を用いたオフライン手書き楽譜自動認識の高精度化

令和 5 年度修了
三重大学大学院 工学研究科 情報工学専攻
ヒューマンコンピュータインタラクション研究室

桐生 拓実

はじめに

コンピュータを用いた音楽制作が一般的になった現代においても，多くの作曲家が手書きによる楽曲の制作を行っている．また手書きのまま印刷されていない楽譜が世界中に存在し，これらが印刷楽譜で流布されることは意義あることと考えられる．しかし紙媒体に記された手書き楽譜は，記譜者毎の変形が多様で激しく，紙やインクの経年劣化によって読譜が困難になり演奏や流通に適さないという問題がある．OMR（Optical Music Recognition）によって手書き楽譜を光学的に読み取ることができれば，整形した楽譜として出力したりコンピュータ上で再生可能な MIDI 形式等に変換したりすることが可能になり，電子楽譜の自動演奏，楽譜の分析，楽譜の他媒体への変換等への道が開ける．

楽譜は主に五線や音楽記号が整形され印刷されている印刷楽譜，タブレット端末等を用いて記されるオンライン手書き楽譜，コンピュータを用いずに直接手書きで記されるオフライン手書き楽譜の 3 種類に分けられる．印刷楽譜，オンライン手書き楽譜を対象とした認識技術は一部が製品化され，実用域に達しているが，オフライン手書き楽譜認識の研究が本格的に始まったのは，ICDAR2011 で実施された手書き楽譜の五線除去に関するコンペティション [1] からである．このコンペティションでは，高いレベルの手法が発表され，五線除去については実用上十分な性能が実現されている．しかし五線除去後の認識についてはまだ研究の基礎段階にある．その理由として，五線上に重ねて描かれた音楽記号の切り出し，記譜者毎に異なる音楽記号の形状や相対的位置情報の認識，複数の音楽記号が接触した場合の分離等の高度なセグメンテーションを必要とする諸課題の存在がある．従って，この分野の研究には手書き楽譜のデジタル化という実用的側面だけでなく，パターン認識研究の難問題の解決を促すという学術的側面があると言える．

これまでに発表された主要な研究成果として，Dutta らの五線間隔を用いた五線検出手法 [2]，Pacha らの深層学習を用いた音楽記号の分類手法 [3]，Baro らの連符の認識手法 [4] がある．しかし，楽譜画像の入力から全体の認識を行う OMR の実現と比べて，これらはいずれも検出，分類，認識のいずれかの要素技術の提案段階にとどまっている [5]．深層学習を導入した関連研究として，Jan らによるバウンディングボックスを用いた符頭検出の研究 [6] がある．楽譜内の音符周りのパッチ画像を学習用に複数用意し，符頭のみ

をバウンディングボックスで検出する実験を行い，Recall 値 0.96, Precision 値 0.97 の精度で検出に成功している．また，Faster R-CNN を用いた音楽記号の検出 [7] がある．この研究では，楽譜内に含まれる複数の音楽記号をバウンディングボックスによって検出することを目指しており，71 クラスの音楽記号を 93.34%の mAP (mean Average Precision) で検出している．

当研究室で発表された先行研究として，早川による音楽記号の分類と音高認識の研究 [8]，山田による高精度な音高認識を実現させる研究 [9]，梶原による音価認識に必要な符尾認識を高精度化させる研究 [10]，筆者による演奏順を考慮した音価・音高認識の研究 [11] がある．早川による研究では，音楽記号を 19 クラスに分類し，音符に距離付き細線化を施した上で，分岐点交差点除去を行って符頭の心線を分離し，その後距離値を用いて太さを復元し符頭領域を検出している．音高認識では，検出した符頭領域の重心座標を算出し，五線およびその中間に設けた音高基準線と比較して，最も近い音高基準線によりその符頭の音高を認識している．山田による研究では，音高を誤認識した符頭について高さ比率による音高認識補正と面積比率による音高認識補正，距離値を用いた符頭領域拡大を提案している．領域拡大と面積比率による音高補正を組み合わせ，検出符頭中の音高認識率を 99.47%まで向上させた．梶原による研究では，符尾の認識方法を改良するため，符頭と符尾を区別する条件や連桁に対する符尾認識手法が工夫された．その結果，符尾の認識率が 90.78%に向上した．筆者の音価・音高認識の研究では，それまでに実装されていなかった五線外の音符に対する音高認識や音符・休符の音価認識を実装し，それらの認識精度をレーベンシュタイン距離 [12] という指標を用いて評価する手法を提案した．

しかし，これら従来の音価・音高認識の研究では，和音を多く含む楽譜に対する認識精度が著しく低くなる傾向があった．和音では複数の符頭が接触して描かれているために，それらを 1 つの符頭として認識してしまうケースが多くみられた．そのため本研究では，インスタンスセグメンテーションを行う深層学習モデル Mask R-CNN [13] を用いて，連結成分単位ではない要素単位の音楽記号の検出手法を提案する．そして提案手法の評価のためにオフライン手書き楽譜のデータセットである MUSCIMA++ [14] に含まれる楽譜を Mask R-CNN に学習させ，MIDI ファイルへ変換するために最低限必要と考えられる 11 種類の音楽記号を分類させる実験を行った．さらに，物体検知モデル Segment Anything Model [15] の導入や手書き楽譜画像をランダムに生成することによるデータ拡張を行い，Mask R-CNN によるクラス分類の精度向上を図った．その結果，手書き楽譜画像のランダム生成によるデータ拡張の手法が最もよい結果となり，F 値は 0.824 であった．

クラス分類後に行う音価・音高認識は先行研究で実装済みであるが，これらは音符の符頭・符幹・符尾といったパーツを連結成分単位で認識するものであったため，符頭同士が隣接した和音や，符幹と符頭が離れて描かれた音符を正しく認識できなかった．そのため，連結成分単位ではなく，Mask R-CNN が出力したマスク情報単位の音価・音高認識を

実装した．音価・音高認識の誤読率をレーベンシュタイン距離に基づいて計算する手法によって評価したところ，完全音高認識の誤読率（音高認識と臨時記号認識を合わせたときの誤読率）が 48.4%から 23.6%へ, 完全音価認識の誤読率（音価認識と付点認識を合わせたときの誤読率）が 47.7%から 31.3%へと減少した.

目次

第 1 章

緒言

1.1　研究背景

コンピュータを用いた音楽制作が一般的になった現代においても，多くの作曲家が手書きによる楽曲の制作を行っている．また手書きのまま整形されていない楽譜が世界中に存在し，これらが整形され流布されることは意義あることと考えられる．しかし紙媒体に記された手書き楽譜は，記譜者毎の変形が多様で激しく，紙やインクの経年劣化によって読譜が困難になるため，演奏や流通に適さないという問題がある．楽譜をディジタル化することでコンピュータ上で再生可能な MIDI 形式ファイルでの出力や楽譜の分析，他媒体への変換が可能になる．これを実現するには光学的楽譜読み取り（Optical Music Recognition: OMR）と呼ばれる楽譜認識システムが必要である．

楽譜は主に五線や音楽記号が整形され印刷されている印刷楽譜，タブレット端末等を用いて記されるオンライン手書き楽譜，コンピュータを用いずに直接手書きで記されるオフライン手書き楽譜の 3 種類に分けられる．印刷楽譜，オンライン手書き楽譜を対象とした認識技術は一部が製品化され，実用域に達しているが，オフライン手書き楽譜認識の研究が本格的に始まったのは，ICDAR2011 で実施された手書き楽譜の五線除去に関するコンペティション [1] からである．このコンペティションにて，五線除去については実用上十分な性能を持つ手法が提案されている．しかし五線除去後の認識については未だ研究の基礎段階にある．その理由として，五線上に重ねて描かれた音楽記号の切り出し，記譜者毎に異なる音楽記号の形状や相対的位置情報の認識，複数の音楽記号が接触した場合の分離等の高度なセグメンテーションを必要とする課題の存在がある．従って，この分野の研究には手書き楽譜のデジタル化という実用的側面の他に，パターン認識研究の難問題の解決を促すという学術的側面があると言える．

図 1.1 に音符の符幹・符頭・符尾を示す．

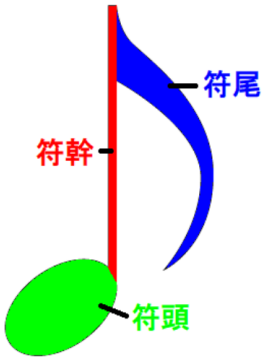

図 1.1: 音符の各パーツの名称

1.2　先行研究

五線を含む楽譜画像から音楽記号の分類と音高認識までを統一的に行う先行研究として早川らによる，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 [8] がある．この研究では，五線から分離された音楽記号を，連桁付き音符，単符，四分休符，強弱記号等 19 クラスに分類し，音符については符頭の検出と，五線に接する範囲の符頭の音高認識を行っている．

また，山田による音高認識の高精度化の研究 [9] では，音高を誤認識した符頭について高さ比率による音高認識補正と面積比率による音高認識補正，距離値を用いた符頭領域拡大を提案している．領域拡大と面積比率による音高補正を組み合わせ，検出符頭中の音高認識率を 99.47%まで向上させた．

梶原による音価認識のための符尾認識の高精度化の研究 [10] では，符尾の認識方法を改善するため，符頭と符尾を区別する条件や連桁に対する符尾認識手法を改良し，符尾の認識率が 90.78%に向上した．その後の音価認識に必要となる，音符ごとに符尾数を数える手法も提案している．

筆者による演奏順を考慮した音価・音高認識の研究 [11] では，それまでに実装されていなかった五線外の音符に対する音高認識や音符・休符の音価認識を実装し，音符や休符の演奏順を決定することで，音価・音高認識の精度をレーベンシュタイン距離 [12] という指標を用いて評価する手法を提案した．さらに，自動認識システムの出力結果を用い，入力した手書き楽譜に対する MIDI 形式ファイルを生成することに成功した．

1.3　研究目的

先行研究では，音符の音価・音高認識において，符頭・符尾・符幹のパーツを連結成分単位で認識していた．そのため，和音のような複数の符頭が接触して描かれた音符や，符頭と符幹，または符幹と符尾が離されて描かれた音符に対して正しい音価・音高認識が行えなかった．本研究では深層学習を導入して連結成分に依存しない音符やその他の音楽記号の分類を行い，クラス分類における精度向上を図ることを目的とする．さらに，その分類情報を用いることによって従来手法より音価・音高認識の精度を向上させることを目指す．また，先行研究では未実装だった付点や臨時記号の認識を行うことで，元の手書き楽譜により近い MIDI 形式ファイルの生成を行う．本研究で作成する自動認識システムの出力は，入力された手書き楽譜に含まれる音符や休符を演奏順に並べ，その音高や音価の情報を付属させた表（以下「出力表」という．）である．出力表の例を表 1.1 に示す．

表 1.1: 出力表の例

演奏順	種類 (音符, 休符)	音高	臨時記号	音価	付点
1	音符	C	-	1	0
2	音符	E	シャープ	1	1
3	音符	G	-	1	0
4	休符	-	-	1	1
5	音符	A	フラット	1	1
6	音符	A	ナチュラル	0.5	1
7	音符	A	-	0.25	0
…					

第 2 章

準備

2.1　データセット

オフライン手書き楽譜のデータセットとして，手書き楽譜データベース CVC-MUSCIMA database [16] がある．CVC-MUSCIMA database に含まれる楽譜の例を図 2.1 に示す．このデータセットは，50 人の記譜者がそれぞれ 20 種の楽譜を描いたものであり，合計で 1000 枚の楽譜が収録されている．しかし，このデータセットが含む正解情報は五線の情報のみであり，従来手法ではこの内 60 枚の楽譜にアノテーションを施し実験を行っていた． J. j. Hajic らが発表した MUSCIMA++ [14] というデータセットは，CVC-MUSCIMA database のデータセットのうち 140 枚にアノテーションを施したものである．記号ごとにクラス分類情報やマスク情報がアノテーションされており，本研究ではこのデータセットを用いて実験を行う．また，音価・音高認識の精度を評価するために，楽譜の音符・休符を演奏順に並べ，音高や音価といった情報をまとめたデータセット（出力表の正解データに相当するもの．以下「正解表」という．）を独自に作成した．

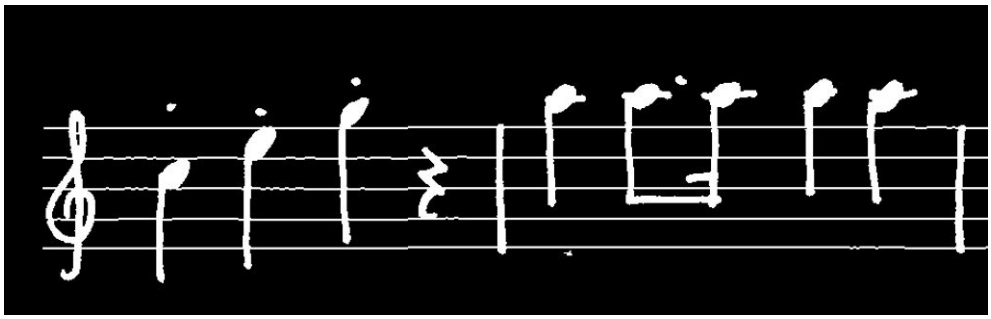

図 2.1: CVC-MUSCIMA database の楽譜画像

2.2 関連手法

同じく MUSCIMA++データセットを用いた研究に，Jan らによるバウンディングボックスを用いた符頭検出の研究 [6] がある．この研究では，楽譜内の音符周りの画像をパッチ画像として切り取り，その画像内に符頭が含まれているかどうかをバウンディングボックスで予測する検出ネットワークを学習させている．さらに，検出した符頭が誤検出でないかを確かめるためのフィルタを機械学習モデルを学習させることで作成し，Recall 値 0.96, Precision 値 0.97 の精度で符頭検出に成功している．

また，Faster R-CNN を用いた音楽記号の検出 [7] がある．この研究では，楽譜内に含まれる複数の音楽記号をバウンディングボックスによって検出することを目指しており，71 クラスの音楽記号を 93.34%の mAP (mean Average Precision) で検出し，特に音符のパーツに関しては表 2.1 から分かるように，高精度な検出に成功している．ここで AP (Average Precision) とは平均適合率のことであり，PR 曲線を積分して求められる値である．

表 2.1: Faster R-CNN を用いた音符パーツ分類の結果

クラス	サンプル数	AP
符頭（穴あり）	1669	99.51
符頭（穴なし）	21333	99.89
符幹	21417	98.52
符尾（連桁）	6593	93.00
符尾（8 分音符）	2200	94.05
符尾（16 分音符）	499	32.77

第 3 章

提案手法

3.1　自動認識システムの流れ

本研究で作成する自動認識システムは，先行研究 [11] 同様「クラス分類」「表の枠組み作成」「音高認識」「音価認識」の 4 段階に分けられる．クラス分類は，その結果がクラス分類後の処理すべてに影響を与えるため，本システムで最も重要な処理といえる．そのため，本研究ではクラス分類に深層学習を導入して連結成分単位ではない音楽記号単位の分類が行えるよう変更し，音価認識においても連結成分単位ではない音楽記号単位の認識が行えるよう処理方法を変更した．また，先行研究では未実装だった臨時記号の認識を音高認識処理に，付点の認識を音価認識処理に追加した．以下，本研究の自動認識システムで使用した手法を「新手法」と呼ぶ.

3.2　クラス分類

従来手法ではクラス分類において，主要な 19 クラスの音楽記号分類を行っていた．この分類は連結成分単位の分類であったため，複数の音符を接続して描く連桁に対しては 1 つの記号として認識した後，画像処理によって符頭・符幹・符尾といったパーツに分類していた．本研究では，連結成分単位ではなく音楽記号単位で分類を行える深層学習モデルを用いて，11 クラスの音楽記号分類を行う．今回分類の対象とした音楽記号一覧を図 3.1 に示す．これらの音楽記号は MIDI 形式ファイルへの変換時に最低限必要だと考えられるものであるため，本研究ではこれら 11 クラスの記号分類を扱う．全休符や 2 分休符は形状自体は全く同じものであり，五線上の座標が異なるだけの記号であるため，これらは「全・2 分休符」という同じクラスとして分類を行う．同様の理由で，8 分休符と 16 分休符，タイとスラー，付点やスタッカートなどの点も同じクラスとして扱う．

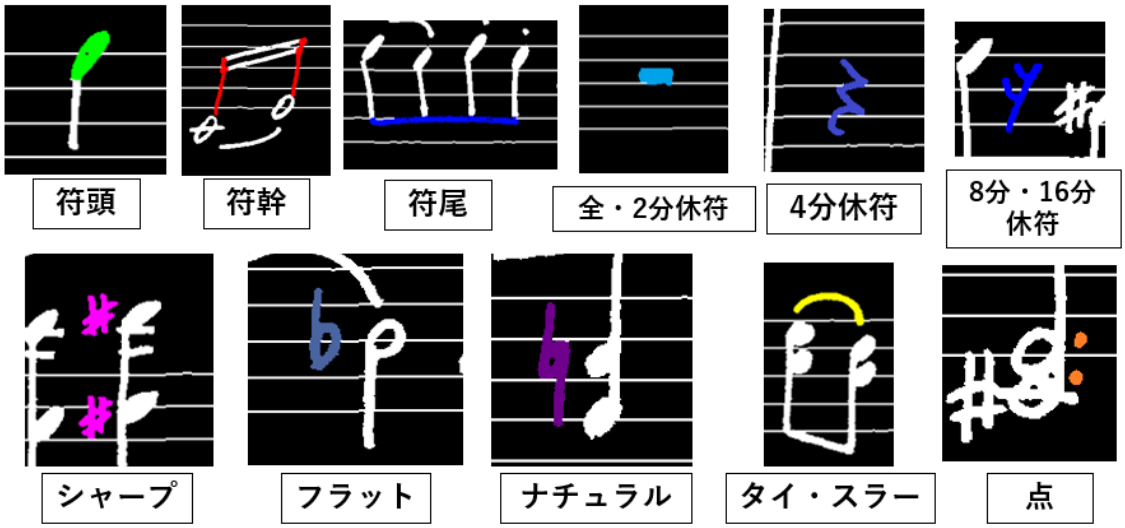

図 3.1: 分類クラス一覧

3.2.1　Mask R-CNN

深層学習モデルを利用して物体検知を行う手法として，バウンディングボックスでの検出とマスク情報を出力することによる画素単位での検出が考えられる．本研究では符頭の重心を求めて音高認識を行ったり，符幹についている符尾の数を数えたりすることを想定しているため，マスク情報を出力する画素単位での検出が必要だと考えた．バウンディングボックスでの検出では，バウンディングボックス内に別の記号が入り込み，それにより符尾の数を余分に多く数えてしまうといった誤認識につながる可能性がある．そこで，画素単位のマスク情報を出力でき，連結成分に依存せずに分類可能な深層学習モデル Mask R-CNN [13] を導入する．Mask R-CNN の構造を図 3.2 に示す．

Mask R-CNN ではバックボーンでの画像全体の特徴を CNN により抽出後，検出したい物体が存在する可能性が高い長方形の領域を畳み込みニューラルネットワークにより探索する Region Proposal Network，RoI の位置を維持しながら特徴マップを特定サイズにリサイズし，各物体のクラスやバウンディングボックスの推定を行う RoI Align の順に処理を行う．その後，各物体のクラスを推定してからバウンディングボックスを推定する Box Regression と各物体のマスクを推定する Mask Generation を並行して処理する [18]. Mask R-CNN はインスタンスセグメンテーションを行うモデルであるため，同じクラスの物体が接触している場合でもそれらを分割して検出することが可能であり，和音のように連結した符頭に対する適切な検出が可能になると考えた．

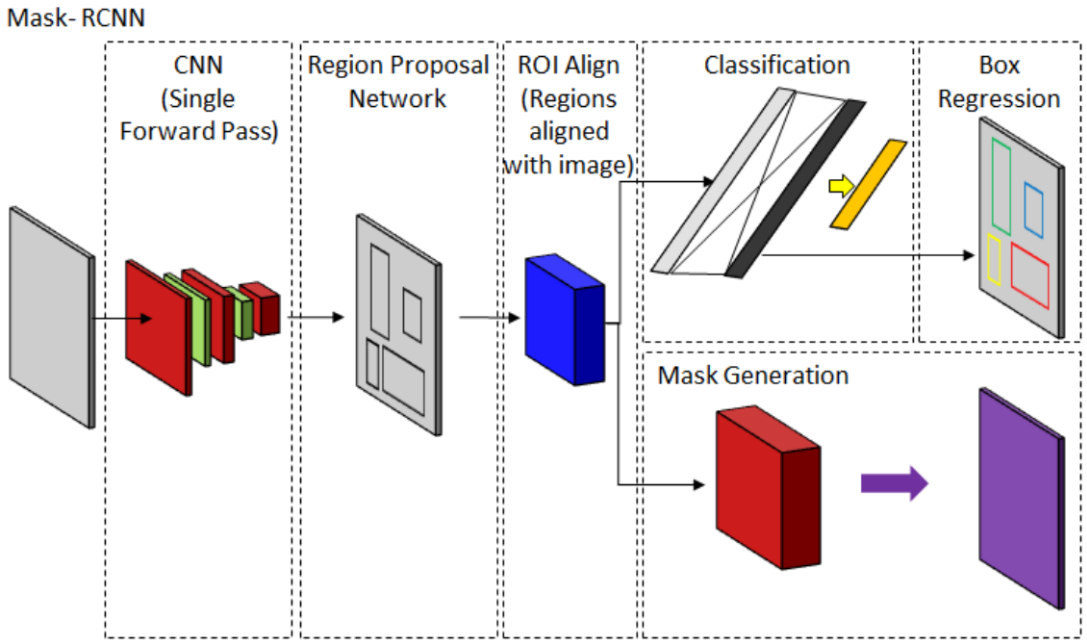

図 3.2: Mask R-CNN の構造 [17]

3.2.2　Segment Anything Model

Segment Anything Model (SAM) [15] は，画像内の座標やマスク，テキスト情報といったプロンプトを入力することで物体検出を行い，その領域を出力するモデルである．SAM の構造を図 3.3 に示す．このモデルは 10 億以上のマスク情報を持った 1000 万枚の画像といった膨大なデータセットで学習されており，Zero-Shot 性能，いわゆる学習用に使われていないような初めて見る画像に対する検出が高精度で可能である．

SAM の学習データに楽譜は含まれていないが，SAM に手書き楽譜を入力することで，一般的な物体を検知する場合と同様に音楽記号を検知することができると考えた．SAM に手書き楽譜を入力した場合の物体検出の結果を図 3.4 に示す．図 3.4 では異なる物体と判定したものを別の色で塗り分けており，この図から分かる通り，音符の符頭と符幹といったパーツを別の物体として検出できている．この色情報が Mask R-CNN による音楽記号のクラス分類の助けになるのではないかと考え，SAM で事前分割した画像をカラー画像として Mask R-CNN に学習させる実験を行った．

3.2.3　手書き楽譜のランダム生成

本研究で使用しているアノテーション済みデータセット MUSCIMA++に含まれる楽譜は 140 枚のみであり，Mask R-CNN の学習用として使用するにはデータ数が少ないと思

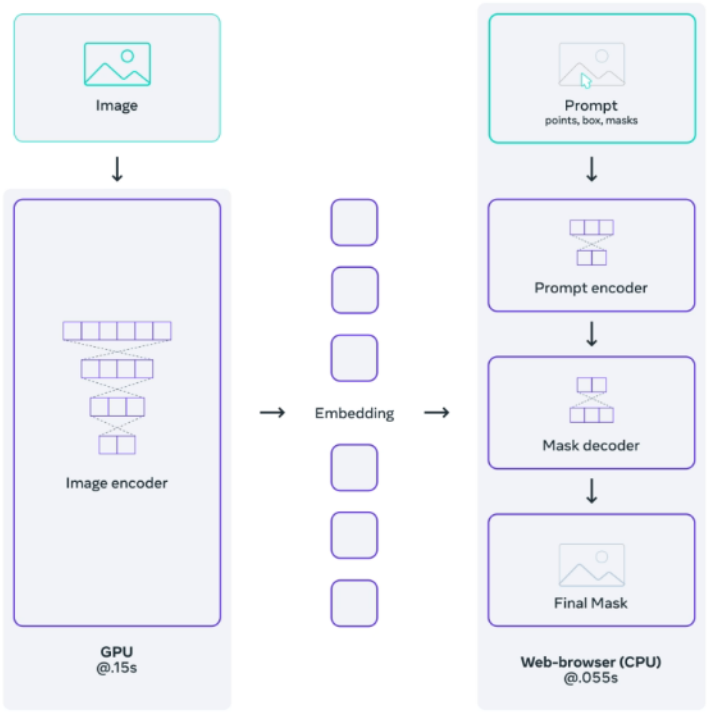

図 3.3: SAM の構造

図 3.4: SAM による楽譜の事前分割

われる．そのため，学習データを大幅に拡張するために手書き楽譜らしい楽譜をランダムで生成する手法を提案する．この手法では，まずランダムな音楽データを生成する必要がある．この音楽データは，自動認識システムが出力する出力表と同様の形式であり，今回は 1 小節を 4 分の 4 拍子とし，その範囲に収まるように音符や休符データをランダムで生成できるプログラムを作成した．

次にこの音楽データを参照し，手書き楽譜らしい楽譜を描画していく．最初に五線や小節線を直線で描画し，その後手書き楽譜データセットに含まれる音符や休符から抽出したものを，音楽データ通りに配置していく．今回手書き楽譜データセットから抽出した要素は，音符の符頭・単符の符尾・休符・臨時記号・付点・タイであり，符幹や連桁の符尾は直線を描画することで表現した．この手法によって，手書き楽譜らしい画像を大量に生成することができるようになり，これらを Mask R-CNN に学習させる実験を行った．ランダム生成した手書き楽譜画像を図 3.5 に示す．

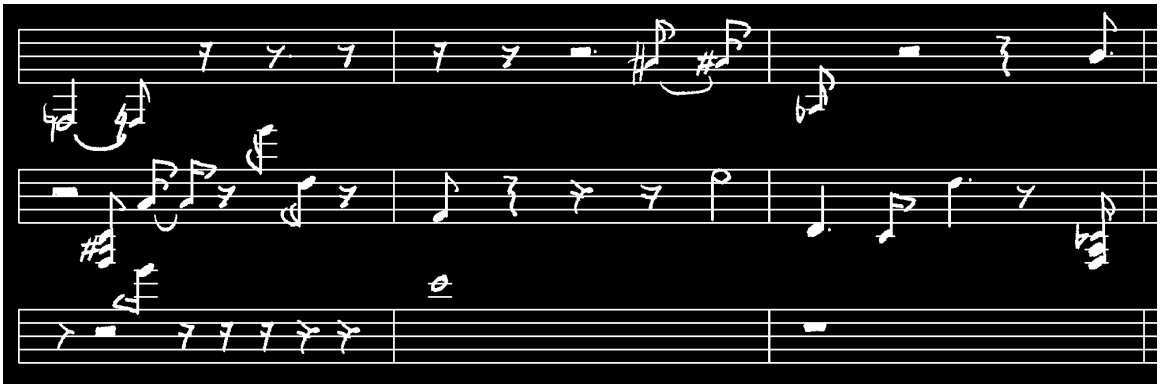

図 3.5: ランダム生成した手書き楽譜画像

3.3　音価・音高認識

臨時記号を除いた音高認識は符頭の重心を調べて五線との位置関係で決定されるため，その手法は新手法においても変更点はない．しかし，音価認識においては連結成分単位ではなく音符単位の処理に変更する必要がある．まず音価認識の初めに行う，符頭の穴の有無の確認は従来手法と同じである．次に符頭に符幹がついているかどうか確認するが，その方法は従来手法では符頭の連結成分をたどって符幹を探すという手法であった．そのため，図 3.6 のように符頭と符幹が離れて描かれる音符に対して，従来手法では符幹なしの音符であると判断していた．このような誤認識を防ぐために，新手法では符頭の上下一定間隔以内に符幹があるかどうか捜査し，符頭と符幹を紐づける．同様に図 3.6 は符幹と符尾も離れて描かれており，このような場合でも連結成分に依存せずに符尾数を決定する必

要があるため，符幹の端付近の一定間隔を捜査して符尾数を決定するものとする．

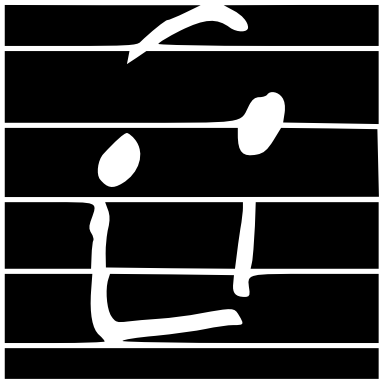

図 3.6: 符幹が符頭や符尾から離されて描かれる例

次に，従来手法では未実装だった臨時記号と付点の認識を実装した．臨時記号には「シャープ」「フラット」「ナチュラル」などがあり，音符の左側に描かれて音高を一時的に変化させる記号である．ダブルシャープなどの臨時記号も存在するが，出現頻度が低いため本研究では扱わないものとする．一方で付点は，音符や休符の右側に描かれ，その音価を 1.5 倍にするという記号である．臨時記号は音符・休符の左側に，付点は右側に描かれるため，単純な画像処理で音符・休符の左右を確認し，臨時記号や付点を紐づける処理を実装した．

3.4 MIDI 形式ファイルへの変換

MIDI 形式ファイルは，音楽の演奏情報をデータ化したバイナリファイルであり，電子楽器やコンピュータ上で再生可能である．新手法の自動認識システムで新たに可能になった臨時記号と付点の認識情報を利用し，より忠実に手書き楽譜データを MIDI 形式ファイルに変換するシステムを実装した．テンポや拍子については自動認識システムで認識できていないため，テンポを 100BPM に，表紙を 4 分の 4 拍子に設定した．

第 4 章

実験

4.1　音楽記号の 11 クラス分類

4.1.1　実験方法

音楽記号の 11 クラス分類実験を 3 つの異なる手法を用いて行った．実験を行うにあたって，手書き楽譜データセットの MUSCIMA++に含まれる楽譜を，学習用と評価用で同じ記譜者・同じ楽曲の楽譜が含まれないように分割し，学習用楽譜を 69 枚，評価用楽譜を 10 枚とした．

まずは基準となる実験として，69 枚の学習用楽譜を Mask R-CNN に学習させ，MIDI 形式ファイルへの変換に最低限必要だと考えられる 11 種類の音楽記号を分類する実験を行った．このとき，Mask R-CNN に入力する画像サイズが大きすぎるとメモリ不足により学習が実行できなかったため，元画像の 1/9 のサイズ（縦横それぞれ 1/3 サイズ）の画像を重なりのあるパッチ画像として切り出し，学習用画像として用いた．パッチ画像の切り出し方として，初めに画像の一番左上をパッチ画像として切り出し，次のパッチ画像は図 4.1 に示すように 1 枚目のものから横（もしくは縦）にパッチ画像サイズの 1/2 だけ移動した部分をパッチ画像として切り出す．この手法によって 1 枚の楽譜につき 25 枚のパッチ画像を生成し，音楽記号が 1 つも含まれないパッチ画像を削除することで，学習用に使われるパッチ画像は 1694 枚となった．

次に，学習用データセットの 69 枚の楽譜に対して物体検知モデル SAM を用いて楽譜内に含まれる音楽記号を事前に分割し，検出した記号をそれぞれ別々の色で塗りつぶした楽譜画像を用意する．これらを同様にパッチ画像に分割し，Mask R-CNN に学習させることで，SAM による事前分割の色情報が Mask R-CNN によるクラス分類に影響を及ぼすかどうかを調査した．

また，学習用データセットを拡張するために，手書き楽譜のランダム生成によって新た

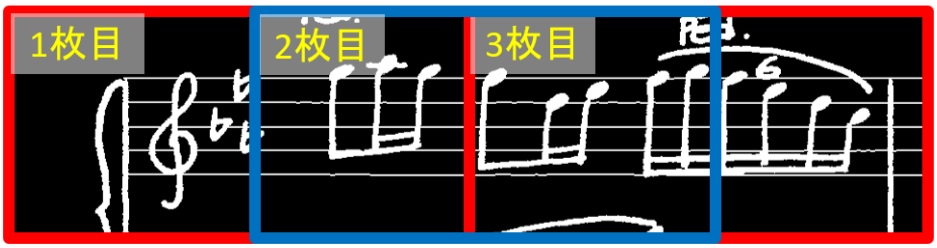

図 4.1: パッチ画像の切り出し方

に 1000 枚の楽譜画像を生成し，同様にパッチ画像に分割した後，Mask R-CNN に学習させて 11 クラス分類を行う実験を行った．このとき学習用のパッチ画像は，MUSCIMA++ のものと合わせて 26694 枚となった．

4.1.2　評価方法

Mask R-CNN は分類した 11 クラス分のマスク画像を出力するが，それらの分類精度を評価するために，まずはパッチ画像を元の楽譜画像に統合する必要がある．その際，重なりのある複数のパッチ画像の両方に描かれている音楽記号が存在し得るため，統合時にそれらのマスク情報が重複しないよう，重なる 2 つの要素のバウンディングボックスから求めた IoU が 0.2 以上の場合，両者を 1 つの要素として統合画像を作成する．パッチ画像統合の様子を図 4.2 に示す．

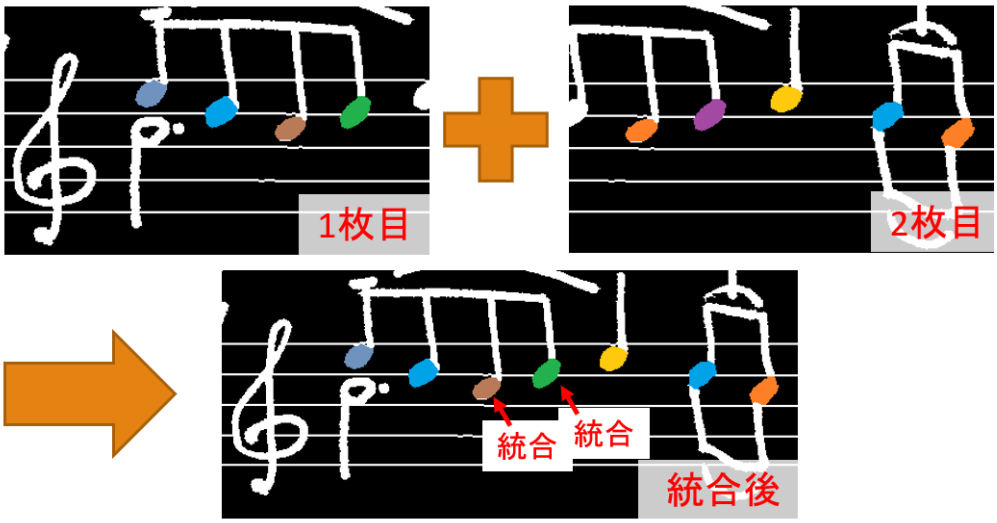

図 4.2: パッチ画像統合の様子

次に，正解マスク情報と Mask R-CNN が出力した予測マスク情報を比較する．比較す

るにあたって，まずは正解領域と予測領域の対応付けを行う．符頭のマスク画像を用いて正解領域と予測領域の対応付けを行っている様子を図 4.3 に示す．正解マスク画像には複数の正解領域が含まれているため，それぞれの正解領域に対応する予測領域を捜査し，その 2 つの領域同士が $IoU = 0.6$ 以上で重なっている場合，両者を対応する領域と決める．

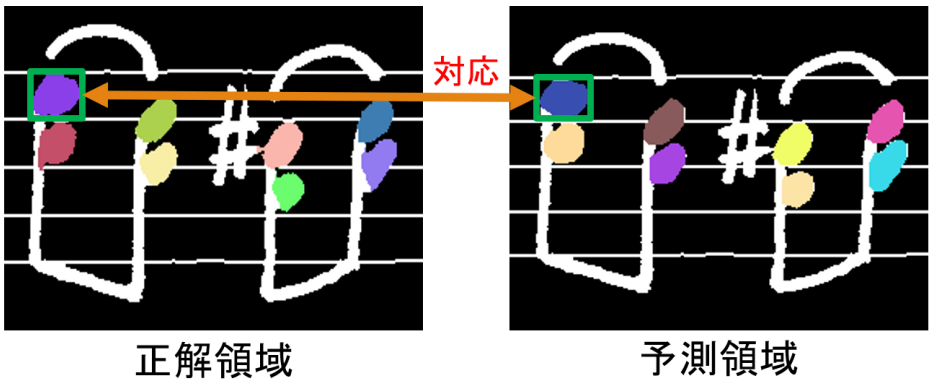

図 4.3: 正解領域と予測領域の対応付け

その後，対応付けされた要素だけを取り出し，正解領域と予測領域両方に存在する画素数を TP，正解領域のみに存在する画素を FN, 予測領域のみに存在する画素を FP とし，マスク画像すべての正解領域に対して求めた TP, FP, FN を合計する．その後，以下の式で $Precision$, $Recall$, $F{-}measure$ の評価指標を求める．

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

$$F{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.3}$$

4.1.3　実験結果

評価用楽譜画像 10 枚に対して，Mask R-CNN を用いて音楽記号の 11 クラス分類を行った．データセットの楽譜をそのまま Mask R-CNN に学習させた実験，SAM による事前分割画像を用いて Mask R-CNN を学習させた実験，手書き楽譜のランダム生成でデータ拡張を行い Mask R-CNN を学習させた実験の結果を表 4.1 に示す．

最もよい精度となったデータ拡張による実験時のクラスごとの結果を表 4.2 に示す．

表 4.1: 3 種類の音楽記号の 11 クラス分類実験

実験方法	Precision	Recall	F-measure
データ変更無し	0.768	0.857	0.810
SAM による事前分割	0.731	0.867	0.796
ランダム生成によるデータ拡張	0.729	0.949	0.824

表 4.2: データ拡張による 11 クラス分類実験

クラス名	Precision	Recall	F-measure	データ拡張前との差分
符頭	0.873	0.949	0.910	+0.009
符幹	0.762	0.904	0.842	+0.023
符尾	0.645	0.877	0.743	+0.021
全・2 分休符	0.873	0.949	0.910	-0.003
4 分休符	0.742	0.909	0.817	+0.005
8 分・16 分休符	0.700	0.904	0.789	-0.001
シャープ	0.815	0.907	0.859	+0.001
フラット	0.223	0.941	0.360	+0.053
ナチュラル	0.368	0.944	0.530	+0.055
タイ・スラー	0.700	0.767	0.732	+0.005
点	0.723	0.912	0.806	+0.114
全体	0.729	0.949	0.824	+0.014

4.1.4　考察

基準となるデータセットの楽譜をそのまま Mask R-CNN に学習させた実験での F 値は 0.810 であった．しかし，精度向上を図って行った SAM による事前分割を行う実験では F 値 0.796 という，基準の実験よりも低い精度となった．このような結果になった原因として，SAM が音符のパーツ分類を正しく行えていない場合が多く存在することが考えられる．SAM が音符のパーツ分類に失敗している様子を図 4.4 に示す．SAM は独立した音楽記号を正しく別の物体として検出する傾向がある一方で，図 4.4 のように，1 つの記号として描かれる音符の符頭・符幹・符尾といったパーツ分類において，不適切な検出を行う傾向にあった．SAM は Zero-Shot 性能に優れているため学習時に使用されていない物体の検出も可能であるが，学習に使用されている画像は風景画像が多く，手書き楽譜内の

記号分割には有効ではなかったと考えられる．

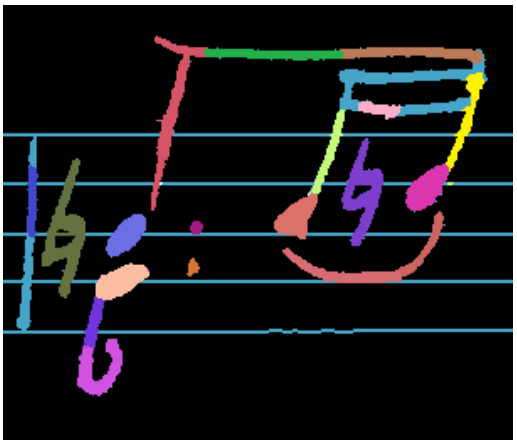

図 4.4: SAM による音符のパーツ分類の失敗例

次に，学習用データセットを拡張するために，手書き楽譜のランダム生成によって新たに 1000 枚の楽譜画像を生成し Mask R-CNN に学習させた実験では F 値 0.824 となった．生成した楽譜がランダムなものであっても，学習データの拡張が認識精度向上に寄与したことがわかる．さらに表 4.2 からは，クラスごとに見てもほとんどのクラスでデータ拡張前の実験より精度が向上していることがわかる．一方でフラットやナチュラルとといった記号に対する認識精度は比較的低く，特に $Precision$ が著しく低くなる傾向にあった．これらの記号の認識では過検出が多く，関係のない記号であるのにも関わらず，該当する記号であると判断してしまうことがあった．別の記号をフラットとして誤認識してしまっている例を図 4.5 に示す．図 4.5 では，強弱記号のフォルテや，符幹と符尾の一部がフラットとして誤認識されている．このような結果となる要因として，データセットに含まれるサンプル数の偏りが挙げられる．今回学習用に使った MUSCIMA++69 枚に含まれる 11 クラスのサンプル数を表 4.3 に示す．表 4.3 から分かる通り，フラットは最もサンプル数が多い符頭に比べて極端に少なく，手書き楽譜のランダム生成時に使用される画像パターンも少くなくなり，Mask R-CNN を適切に学習させられなかったと考えられる．解決策として，元々サンプル数が少なかったクラスの音楽記号を多く使用した手書き楽譜をランダム生成することや，GAN [19] といった生成モデルを使用することによって，サンプル数が少ない記号パターンを増加させたランダム楽譜を生成することが挙げられる．一方でフラットやナチュラルと同じく 4 分音符やシャープもサンプル数は少ない．それにもかかわらず 4 分音符やシャープの精度が低くならなかったのは，これらの記号の形状が他の記号とかけ離れているからであると考えられる．

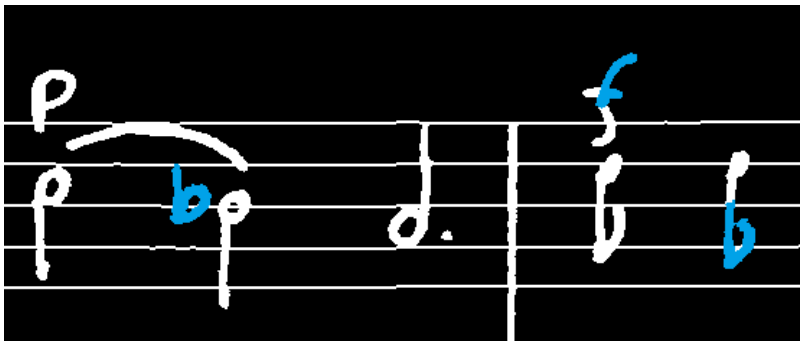

図 4.5: フラットの誤認識の例

表 4.3: MUSCIMA++69 枚における 11 クラスのサンプル数

クラス名	サンプル数
符頭	10637
符幹	10284
符尾	1842
全・2 分休符	4639
4 分休符	111
8 分・16 分休符	345
シャープ	764
フラット	712
ナチュラル	723
タイ・スラー	1447
点	1852

4.2　音価・音高認識

4.2.1　実験方法

音楽記号の 11 クラス分類実験で評価用として用いた MUSCIMA++10 枚の楽譜を，従来手法の自動認識システムと新手法の自動認識システム両方に入力し，それぞれ出力表を出力した．この出力表と正解表を用いて，音高認識と音価認識の精度評価を行い，新手法と従来手法の自動認識システムの精度を比較する．

4.2.2　評価方法

自動認識システムが出力した出力表がどれほどの精度を持つかを評価するために，レーベンシュタイン距離 [12] という指標を用いる．この指標は，2 つの文字列 A, B が存在するとき，A と B の文字列がどれだけ異なっているかを距離として表したものである．文字列 A に特定の操作を n 回行って文字列 B になるとき，考えられる最小の n が A と B のレーベンシュタイン距離となる．この特定の操作は，図 4.6 に示す「挿入」「削除」「置換」の 3 種類である．

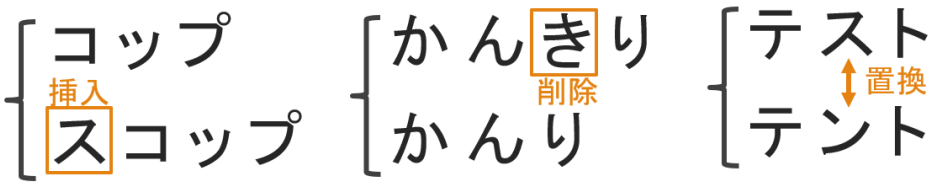

図 4.6: レーベンシュタイン距離を求めるときに使用する操作

本研究の自動認識システムをレーベンシュタイン距離で評価する方法は，出力表と正解表の特定の列を文字列として扱い，それらの間でレーベンシュタイン距離を求め，実際の楽譜に含まれる音符・休符の数で割ることによって誤読率を求めるというものである．出力表と正解表が表 4.4 であったとき，音高の精度を評価したい場合，出力表の音高の列から休符を除いた「CEGAB」という文字列が得られる．同様に正解表からは「CEGBBC」という文字列が得られ，両者の文字列のレーベンシュタイン距離を求めと，その距離は 2 となる．これを正しい音符数 6 で割った 0.333 という値を表 4.4 の場合における音高の誤読率とする．実際の音高は臨時記号を考慮したものであるため，音高と臨時記号の列を同時に評価することも考える．このとき評価する音高を完全音高とし，音価と付点を同時に評価する場合は完全音価とする．完全音高の評価方法としては，出力表の音高と臨時記号を 1 つのタプルとして「(C, -)(E, シャープ)(G, -)(A, フラット)(B, ナチュラル)」という文字列を用意し，同様に正解表から用意した「(C, -)(E, シャープ)(G, -)(B, フラット)(B, -)(C, -)」という文字列とのレーベンシュタイン距離を求める．(A, フラット) と (B, フラット) は違う文字として扱われるため，このときのレーベンシュタイン距離は 3 であり，完全音高の誤読率は 0.5 である．

表 4.4: 出力表と正解表の例

演奏順	音符 or 休符	音高	臨時記号
1	音符	C	-
2	音符	E	シャープ
3	休符	-	-
4	音符	G	-
5	音符	A	フラット
6	音符	B	ナチュラル

演奏順	音符 or 休符	音高	臨時記号
1	音符	C	-
2	音符	E	シャープ
3	休符	-	-
4	音符	G	-
5	音符	B	フラット
6	音符	B	-
7	音符	C	-

4.2.3　実験結果

音高認識に関する誤読率を表 4.5 に，音価認識に関する誤読率を表 4.6 にそれぞれ示す．

表 4.5: 音高認識の誤読率

記譜者番号 - 楽譜番号	音符数	音高	臨時記号	完全音高
W34-P16	168	0.179	0.042	0.202
W37-P17	191	0.298	0.058	0.319
W38-P18	188	0.144	0.122	0.207
W39-P20	275	0.105	0.178	0.225
W40-P19	140	0.343	0.171	0.414
W41-P16	168	0.137	0.036	0.137
W44-P17	191	0.246	0.120	0.288
W45-P18	188	0.138	0.133	0.191
W46-P20	275	0.185	0.149	0.222
W48-P16	168	0.137	0.036	0.149
平均	195.2	0.191	0.105	0.236

従来手法と新手法による自動認識システムの誤読率の比較を表 4.7 に示す．

4.2.4　考察

表 4.5 から分かるように，10 枚の評価用楽譜すべてに対して音高の誤読率を 0.3 未満に，臨時記号の誤読率を 0.2 未満に抑えることができ，完全音高の誤読率は平均で 0.236

表 4.6: 音価認識の誤読率

記譜者番号 - 楽譜番号	音符数 + 休符数	音価	付点	完全音価
W34-P16	194	0.129	0.098	0.191
W37-P17	224	0.563	0.170	0.594
W38-P18	233	0.236	0.112	0.249
W39-P20	275	0.262	0.062	0.273
W40-P19	148	0.243	0.007	0.243
W41-P16	194	0.160	0.103	0.232
W44-P17	224	0.241	0.290	0.464
W45-P18	234	0.205	0.077	0.209
W46-P20	275	0.335	0.113	0.385
W48-P16	194	0.201	0.108	0.289
平均	219.5	0.257	0.114	0.313

表 4.7: 従来手法と新手法の誤読率の比較

	音高	臨時記号	完全音高	音価	付点	完全音価
従来手法	0.455	0.318	0.484	0.443	0.319	0.477
新手法	0.191	0.105	0.236	0.257	0.114	0.313

であった．しかし，W40-P19 の楽譜において音高や完全音高の誤読率が比較的高い結果となった．この原因として，この楽譜には五線外に符頭が飛び出している音符が多く含まれることが考えれられる．W40-P19 の楽譜の一部を図 4.7 に示す．五線外に書かれた音符の音高認識では加線の本数や符頭との位置関係で音高を決定する必要があるが，現状では加線の認識はできておらず，加線が五線間隔と同間隔で引かれているものと仮定して音高認識を行っている．しかし，図 4.7 から分かる通り，加線は五線間隔よりも狭い間隔で描かれることが多く，符頭が五線から離れた位置に描かれるほど適切な認識ができなくなってしまう．そのため，五線外の音符に対しても適切な音高認識が行えるよう，加線の認識もクラス分類の時点で行う必要があると考える．

次に，表 4.6 から分かるように，音価の誤読率は 10 枚の評価用楽譜の平均 0.257, 付点の誤読率は 0.114, 完全音価の誤読率は 0.313 であった．完全音価の誤読率が高かった楽譜として W37-P17 の楽譜が挙げられる．W37-P17 の楽譜の一部を図 4.8 に示す．この楽譜の特徴として，符頭が楕円形ではなく傾いた直線のように描かれ，符幹が符頭からかな

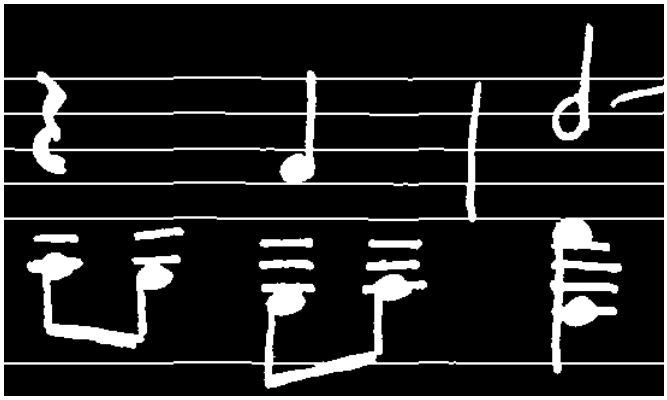

図 4.7: W40-P19 の楽譜の一部

り離れて描かれている．新手法の符頭と符幹の対応付け方法は，符頭の上下付近から符幹を探すというものであるが，このとき符頭から一定間隔はなれた場所しか捜査しないため，図 4.8 のような音符に対して正しく符幹が認識できない可能性がある．これらの問題の解決策として，符幹認識が終了した時点で，どの符頭にも対応付けされていない符幹があった場合に，その符幹からさらに遠くまで符頭が存在しないか操作するといった手法や，深層学習を用いてクラス分類の時点で符頭・符幹・符尾の対応付けまで行ってしまうといった手法が挙げられる．

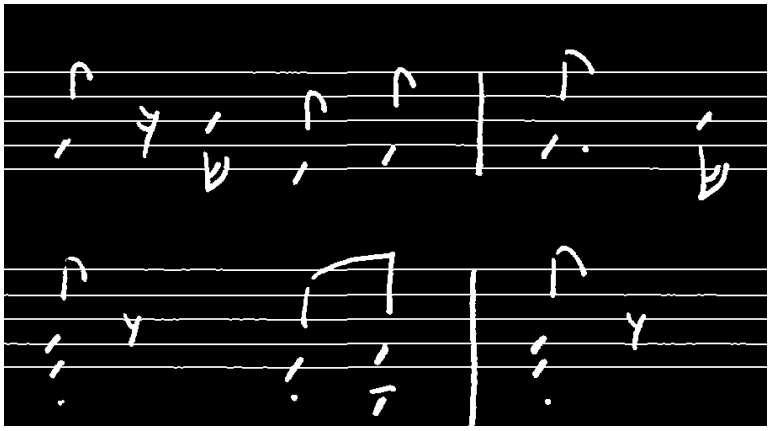

図 4.8: W37-P17 の楽譜の一部

最後に，表 4.7 では従来手法と新手法における自動認識システムの誤読率を比較することができる．この表から，音高認識と音価認識のすべてにおいて，従来手法より新手法の自動認識システムの方が誤読率が低くなったことが分かる．一方で，完全音高における誤読率の減少率よりも，完全音価における誤読率の減少率が少ないという結果になっている．音高認識は符頭の重心を調べるという単純な方法であるが，音価認識は符幹の有無や

符尾数を数えるなどの工程を踏む必要があり，さらに休符に対してもそれに対応した音価認識が行われる．そのため自動認識システム全体の精度を向上させるためには，音価認識に対する精度向上にむけた改善策の提案が必要だと考えられる．

第 5 章

結言

5.1　まとめ

本論文では，Mask R-CNN を用いた音楽記号の 11 クラス分類手法を提案し，SAM による事前分割や手書き楽譜のランダム生成によるデータ拡張などの実験を行った．その結果，手書き楽譜ランダム生成によるデータ拡張をしたうえで Mask R-CNN を学習させる手法の認識精度が最もよく，F 値 0.824 となった．また，従来の自動認識システムに臨時記号や付点の認識，連結成分に依存しない音価認識を実装し，レーベンシュタイン距離を用いた精度評価を行った．その結果，完全音高の誤読率が 0.236, 完全音価の誤読率が 0.313 となり，これらは全て従来手法の自動認識システムより低下した．さらに，臨時記号と付点を考慮した手書き楽譜データから MIDI 形式ファイルへの変換も可能にした．

5.2　今後の課題

手書き楽譜のランダム生成によるデータセット拡張を施すことによって，データ拡張前の実験よりも制度は向上したが，クラス別の認識精度を見てみると，フラットやナチュラルといった一部の音楽記号に対して認識率が著しく低い結果であった．フラットは最もサンプル数が多い符頭に比べて極端に少なく，手書き楽譜のランダム生成時に使用される画像パターンも少くなくなり，Mask R-CNN を適切に学習させられなかったと考えられる．そのため，GAN [19] などの生成モデルを使用することによって，サンプル数が少ない記号パターンを増加させたランダム楽譜を生成することや，サンプル数の偏りをなくすために元々サンプル数が少ない記号を中心にランダム生成させるという手法が解決策に挙げられる．

また，新手法における自動認識システムにおいても，五線外の音高認識や音符の符幹や符尾の認識精度が低く，課題が残る結果となった．深層学習を用いた手法によって，加線

の認識や音符の符頭・符幹・符尾の対応付けをクラス分類の時点で行うことができれば，その後の音価・音高認識の精度向上につながると考えられる．

第 6 章

付録

本研究に関するプログラムはすべて以下の github リポジトリ

- KiryuTakumi/music-score

6.1　コンパイル情報

プログラムのソースファイルのコンパイルは
KiryuTakumi/music-score
以下に存在する makefile を利用して行う.

6.2　プログラムの詳細

KiryuTakumi/music-score 以下に存在する vertical.sh 等を実行する．詳しい実行方法は README を参照のこと.

謝辞

本研究に関し，数多くのアイディアやアドバイス，その他様々なご指摘を熱心にご教授くださった若林哲史教授，盛田健人准教授，白井伸宙助教，日頃からお世話になった吉永みゆき事務員に深く感謝いたします．

また，日々の研究の中のみならず様々な相談に乗って下さったヒューマンコンピュータインタラクション研究室の先輩方，同期の友人の皆さん，後輩たちに深く感謝いたします．支えて頂いた皆様のおかげで，この研究生活は私にとって非常に有意義なものとなりました．

最後に，私の大学生活を支えてくれた両親に今一度深い感謝の意を表し，本論文の結びとします．

参考文献

[1] Albelt Gordo Josep Llados Alicia Fornes, Anjan Dutta. The icdar 2011 music scores competition: Staff removal and writer identification. *ICDAR*, 2011.

[2] et aj A.Dutta. An efficient staff removal approach from printed musical documents. *20th ICPR*, pp. 1965–1968, 2010.

[3] H.Eidenberger A.Pacha. Towards a universal music symbol classifier. *14th ICDAR*, Vol. 2, pp. 35–36, 2017.

[4] A.Fornes A.Baro, P.Riba. Towards the recognition of compound music notes in handwritten music scores. *15th ICFHR*, pp. 465–470, 2016.

[5] et al E. Shatri. Optical music recognition:state of the art and major challenges. *accepted to the International Conf. on Technologies for Music Notation and Representation(TENOR)*, 2020.

[6] Jan et al. Detecting noteheads in handwritten scores with convnets and bounding box regression. *Charles University*, 2017.

[7] Bertrand B. Coüasnon yann Ricquebourg Richard Zanibbi Horst Eidenberger Alexander Pacha, Kwon-young Choi. Handwritten music object detection: Open issues and baseline results. *HAL open science*, 2018.

[8] 早川優木. 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. 三重大学修士論文, 2018.

[9] 山田睦実. オフライン手書き楽譜の音高認識. 三重大学卒業論文, 2020.

[10] 梶原有紗. オフライン手書き楽譜 音価認識のための符尾認識の高精度化. 三重大学卒業論文, 2021.

[11] 桐生拓実. 音符・休符の演奏順を考慮したオフライン手書き楽譜の自動認識. 三重大学卒業論文, 2022.

[12] Levenshtein V. I. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii. NaukSSSR*, Vol. 163, No. 4, pp. 845–848, 1965.

[13] Gkioxari-G. Dollar P. He, K. and R. Girshick. Mask r-cnn. *he IEEE International*

Conference on Computer Vision (ICCV), 2017.

[14] J. j. Hajic and P. Pecina. The muscima++ dataset for handwritten optical music recognition. *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, 2017.

[15] Nikhila Ravi Hanzi Mao Chloe Rolland Laura Gustafson Tete Xiao Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár Ross Girshick Alexander Kirillov, Eric Mintun. Segment anything. *arXiv:2304.02643*, 2023.

[16] A.Gordo J.Lladoss A.Fornes, A.Dutta. Cvc-muscima: A ground-truth of hand-written music score image for writter identification and staff removal. *ICDAR*, Vol. 15, No. 3, pp. 243–251, 2012.

[17] Das N. Das I. Ghosh, S. and U. Maulik. Understanding deep learning techniques for image segmentation. *arXiv:1907.06119*, 2019.

[18] 友澤 弘充 田口 仁内藤 昌平. 複数種類の高解像度衛星画像を用いた mask r-cnn による建物抽出・被害分類モデル. AI・データサイエンス論文集 4 巻 3 号, 2023.

[19] et al. I. J. Goodfellow. Generative adversarial nets. *Process Syst*, 2014.