

物体検出結果に基づく凸包モデルを用いた音楽の自動生成

長田悠希[†] 望月茂徳[†]

概要：映像と音楽の結合を行うためには、両者の時間的な同期のみならず、映像の内容や文脈といった意味理解を音楽に結びつけることが重要な課題となっている。本研究では、映像をもとに自動的な音楽生成を行うための一つのアプローチとして、映像内の物体検出および凸包を用いた分析による手法を提案する。これを実装したシステムの展示を行い、51人にアンケート調査を行ったところ、73%の回答者が生成された音楽と映像が合っていると感じていたことがわかった。また、その要素として多くの評価を得た「物の動き」「物の種類」について考察を行った。

1. 背景と目的

1.1 映像と結合する音楽

映像と結びついた音楽の歴史は古く、トーキー技術が一般化した1930年頃から40年頃には、ハリウッドにおけるマックス・スタイナーやエーリヒ・ヴォルフガング・コルンゴルトらの活躍により、「ライトモチーフ」や「ミッキー・マウシング」の手法を取り入れた、現在の映画音楽の原型ともなる形が誕生している[1]。この場合、映像と音楽の時間的な同期のみならず、映像の内容や文脈といった意味理解を音楽に結合させることが重要な課題となっている。

本研究では、高度な映像の解析が可能となってきたことを背景とし、映像をもとに自動的な音楽生成を行うための一つのアプローチを提案する。

1.2 関連研究

映像をもとにした音楽の自動生成については、いくつかの研究が行われている。山本ら（2009）[2]は、音楽心理学に基づき、動画の色情報よりコード進行、リズム、伴奏、メロディラインのそれぞれを求めて生成を行う方法を提案している。また、佐々木ら（2017）[3]は、水槽の映像よりオブティカルフローによって魚の動きを求め、仮想弦を通過した際に発音を行う方法を提案し、これを用いたリアルタイムな音楽生成システムを制作している。

2. 提案手法とシステム

2.1 概要

本研究では、以下の二つの過程によって、任意の与えられた映像より音楽の生成が行われるものとする。

- (1) 映像を分析し、内容に対する解釈を行う。
- (2) 解釈をもとに、それに応じた音楽の生成を行う。

(1)は、与えられた映像に応じて音楽を構成するために、映像より何らかの要素を抽出し、それに対して適切な意味付けを行う過程である。今回は、映像（動画）の各フレー

ムに対して物体検出を行った後、検出結果に対して凸包を用いた分析を行う手法を提案する。

(2)は、(1)で得られた意味内容を反映させながら、アルゴリズムに基づき音楽を生成する過程である。今回は、音楽を「リズム」「旋律」「和声」の三要素[4]で捉え、凸包による分析結果をパラメータとしながら各要素の生成を行う。

これらの方針に基づき、本研究では映像に合わせて自動的に音楽の生成を行うシステムを制作する。なお、入力にはカメラより取得された映像を用いることとし、今回はリアルタイムな映像に応じて逐次的に音楽が生成されるものとする。システムの全体像を図1に示す。

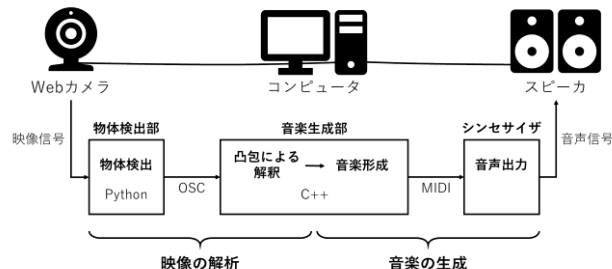


図1 システムの全体像

2.2 映像の解析

2.2.1 物体検出

物体検出とは、与えられた画像中より定められた物体を検出し、種類（クラス）および位置の識別を行う処理である。物体の位置は、バウンディングボックス（bounding box）と呼ばれる矩形によって表され、囲まれた領域内に任意のクラスの物体が存在することを予測する。ここでは、物体検出アルゴリズムとして、検出精度とリアルタイムでの実行速度の観点より、Single Shot MultiBox Detector（SSD）[5]を用いることとした。

なお、今回はデータセットに PASCAL VOC2007 および VOC2012 を使用し、VGG16 ネットワークで訓練された SSD300 モデルを使用する。

[†] 立命館大学映像学部

2.2.2 凸包モデル

物体検出は、動画のフレームごとに独立して行われるため、前後のフレームにおいて個々の検出物体の同一性は保証されない。この問題の解決を図るため、本研究では物体検出結果をクラス単位で分類し、凸包の概念を導入することにより、クラスごとに検出結果の一元化を試みる。

凸包とは、与えられた集合を包含する最小の凸集合で、2次元平面における点集合では凸多角形となる。ここでは、物体検出によって得られた矩形の頂点を使用し、同一クラスの物体ごとに凸包を求めることとする。

各物体の検出結果は、図2に示すようなクラスと矩形座標の組で取得される。これに対して凸包を求めた場合、図3が得られる。このような凸包どうしの比較を考えたとき、各々の凸多角形より一意に求まる値の大小比較とする方法が最もモデルを単純化でき、高速な処理が可能かつ、実装も容易であると考えられる。本研究では、凸多角形の面積ならびに重心を用いることとし、それによる物体検出結果の分析手法を「凸包モデル」と呼ぶ。

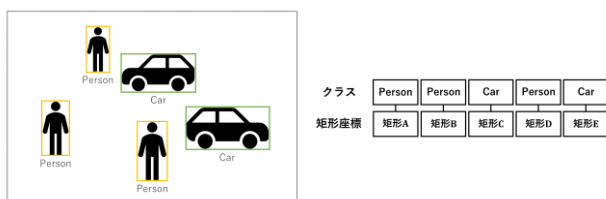


図2 物体検出結果



図3 凸包モデルによる物体検出結果の分析

あるクラスの凸包の面積が画像全体の面積に対して一定以上を占める場合、そのクラスの物体が画像における主要な構成要素であることが推測される。また、これをクラスごとの物体数とあわせて評価したとき、物体数が少ない

(1, 2個程度)場合は、比較的大きな物体が画像の一部を占有している様子を、物体数が多い場合は、物体が広範囲にわたって分布している様子をそれぞれ推測することができる。このように、一つの画像を対象とした場合には、凸包モデルによって各クラスの重要度や分布の様子をある程度推定できるものとする。

さらに、静止画像のみならず動画に対しても凸包モデルの適用は可能である。ここでは、前後のフレーム間における凸包の面積および重心座標の変化が大きい状況を「動」、

小さい状況を「静」とそれぞれ捉えることで、映像がどのような動きの状態にあるのか推定を試みる。

これらをまとめると、凸包モデルによる分析の手続きは以下の通りである。

第一に、物体検出によって得られたバウンディングボックスをクラスごとに分別し、各クラスにおいて矩形の頂点座標より凸包を求める。

第二に、得られた凸包の面積および重心を算出する。

第三に、動画のフレーム間における面積および重心の変化率を求める。ここでは、面積の変化率を前後のフレーム間における面積率の差の絶対値とし、重心の変化率を画像全体の対角線長に対する前後のフレーム間での重心の移動距離とする。

また、これにあわせて物体検出結果の誤差を吸収するために平滑化を行う。ここでは、クラスごとの物体数にも平滑化処理を適用し、瞬間的な検出漏れや誤検出の影響を緩和するため、整数値ではなく実数値として扱う。

これらの計算によって得られた値を映像の状況を表すパラメータと考え、映像に応じた音楽の生成を行うために使用する。本研究では、表1に示す値を用いることとし、これらを「凸包パラメータ」と呼ぶ。なお、「面積変化率」と「重心変化率」は、まとめて「変化率」と呼称する。

表1 凸包パラメータ

値	説明
クラス	凸包のクラス名
物体数	各クラスの凸包に含まれる物体の数
面積率	全体の面積に対する凸包の面積の比率
面積変化率	前後のフレーム間における面積率の差の絶対値
重心変化率	全体の対角線長に対する前後のフレーム間での重心の移動距離

2.3 音楽の生成

凸包モデルによる分析結果を用いて音楽を生成するにあたり、今回は1小節を生成の単位として逐次的にパラメータの反映を行うこととした。また、生成される音楽はホモフォニー様式の調性音楽とし、調と音階が固定された4分の4拍子のものとする。

音楽の生成は「ルールベース」[6]で行うこととし、次項以降で示される制約規則群に基づき、乱数生成を用いた選考を重ねることにより実行される。また、凸包パラメータは選考における確率分布の操作という形で適用される。

選考は、大きく「クラス選考」「リズムパターン選考」「和声進行規則選考」「暗意－実現パターン選考」の順に、前の選考結果を引き継ぎながら行われる(図4)。

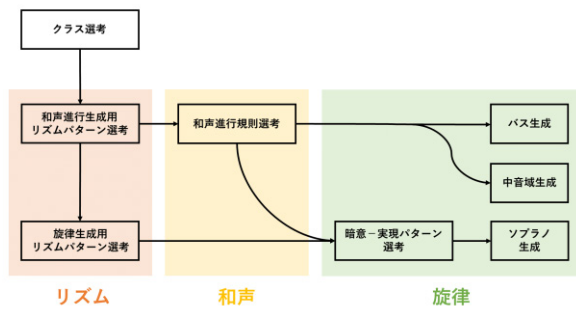


図 4 音楽生成における選考の流れ

2.3.1 クラス選考

ここでは、全てのクラスの凸包パラメータを比較し、単位小節の生成時点における映像中の主要なクラスを推定する。この選考結果をもとに、以降の選考において任意のクラスの凸包パラメータが適用される。

選考にあたり、今回は以下のような規則を設定した。

- (1) 選考時に物体が存在しないクラスは、選考の対象と
ならない。
- (2) 物体数の多いクラスほど優先される。
- (3) 凸包の面積率が大きいクラスほど優先される。
- (4) 凸包の変化率が大きいクラスほど優先される。

なお、検出可能な物体が存在しなかった場合は、以降の選考に対して「物体なし (none)」が渡され、各選考における凸包パラメータに基づく規則は適用されないものとする。また、今回の実装においてはホモフォニー様式の調性音楽としているため、選出されるクラス数は1クラスとし、主旋律を中心として反映が行われるものとした。

2.3.2 リズムパターン選考

本研究では、シェンカーの提唱した簡約仮説 (reduction hypothesis) [7]や Generative Theory of Tonal Music (GTTM) [8]における音楽への階層的な理解を参考としながら、任意の生成区間における拍節構造をある程度パターン化し、分節ごとのパターンを再帰的に選考することで、概念的な木構造の構築を試みた。そのための選考を、ここではリズムパターン選考と呼ぶ。リズムパターン選考には、「和声進行生成用リズムパターン選考」と「旋律生成用リズムパターン選考」の二つのサブ選考がある (図 5)。

1. 和声進行生成用リズムパターン選考

和声進行生成用リズムパターン選考は，単位小節内における和音進行の発生箇所を定めたいくつかの分節パターン（リズムパターン）より，和音進行数および各和音の音価を決定するための選考である．今回は，小節内を分節する最小の区間を 4 分音符一つ分の音価として，8 通りのパターンを求めた（図 6）

に分節パターンを設定する．今回は，分節の最小単位を 8 分音符の音価とし，以下の規則に基づきいくつかのパターンを選択した．

- (1) 上位階層の区間を全く区切らないパターンを含める．
- (2) 上位階層の区間は，過度に分節されてはならない．
- (3) 上位階層の区間長をある程度保つパターンを中心に，選考の終端となるパターンを含める．
- (4) 8 分音符の使用は，原則として他の分節方法が存在しない場合と，上位階層の区間長をある程度保ちながら弱拍に分節を追加する場合のみに限定する．

今回の使用パターンの一覧を図 7 に示す．インデックスに「T」がつくものは，そのパターンが対象区間の選考における終端となり，次の選考階層へは進まないことを表す．全ての区間においてパターンが終端に達した場合，リズムパターン選考が完了したものとみなす．

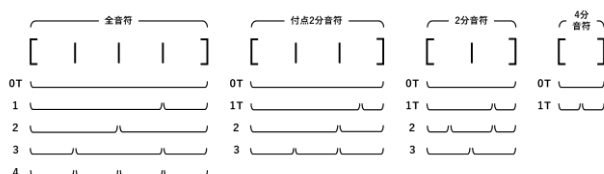


図 7 旋律生成用リズムパターン

2.3.3 和声進行規則選考

ここでは，和声学における機能と和声の考え方に基づき，和音の連結を行うこととする．本研究では，西田ら (2002) [9] による，HPSG (Head-driven Phrase Structure Grammar, 主辞駆動文法) に基づき和声解析を行った研究を参考としながら，これを逐次的な生成において実装するため，「カデンツ全体の主辞句の選考」「句の選考」「カデンツの開始和音の選考」「和音の選考」の四つのサブ選考を設けた．

まず，カデンツの構文木を構築するにあたり，最初のトニック句よりそのカデンツの主辞 (head) を決定する (カデンツ全体の主辞句の選考)．これにより木構造の最上位の親となる句が決定され，各句の下位範疇化規則に基づきながら，親子関係の構築 (句の選考) と木構造の末端となる和音の決定 (和音の選考) がなされる．これらの選考が最後のトニック句に達したとき，その句が次のカデンツの開始句となると同時に，その句によって下位範疇化された和音が決定される (カデンツの開始和音の選考)．このようにして，いくつものカデンツが結合されながら和声が行進していく (図 8)．なお，各句の下位範疇化規則は，原則として文献 [9] に従うこととする．

この選考における凸包パラメータの適用規則は，以下の通りとした．

- (1) 物体数が多く，凸包の面積率が大きいほど，I の和音が主辞となりやすい．

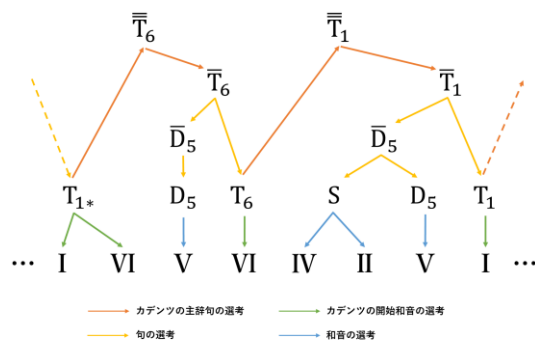


図 8 カデンツの形成

- (2) 物体数が少なく，凸包の面積率が小さいほど，ドミナントがⅢの和音となりやすい．
- (3) 凸包の変化率が小さいほど，IV → I に進行しやすい．
- (4) 凸包の変化率が小さいほど，同一の機能に留まりやすい (次の機能へ進行しにくい)．

2.3.4 暗意—実現パターン選考

旋律の生成においては遡行的な処理を行わず，ある音の決定には判断材料としてそれ以前の音のみを用いることとした．ここでは，局所的な 3 音に対しゲシュタルト的な理解を行う暗意—実現モデル (Implication-Realization Model) と，そこで提示される二つの原理 (方向原理，音程原理) を参考とする．文献 [7] では，これに基づく 3 音の関係を「暗意—実現パターン」としてまとめ，8 種類のパターン (D, ID, P, IP, VP, IR, R, IR) を挙げている．

本研究ではこれらのパターンを借用し，前 2 音の関係より次の音を求めるためのパターンの選考を繰り返すことで，ボトムアップに 1 音ずつ旋律を形成していくこととする．ここでは，選考において以下の規則を適用した．

- (1) 物体数が多いほど，第 2 音と第 3 音の音程が大きくなりやすい．
- (2) 凸包の変化率が大きいほど，第 2 音と第 3 音の音程が大きくなりやすい．
- (3) 方向原理と音程原理がともに成立するパターンは，選考において優先される．

2.3.5 旋律の生成

今回は，主旋律および伴奏を，「ソプラノ」「中音域」「バス」の 3 パートに分けて生成することとした．

主声部であるソプラノパートの旋律は，これまでの選考結果に加え，以下の条件に基づきながら 1 音ずつ選考によって決定される．

- (1) ソプラノ音域内の音であること．
- (2) 既定の調および音階の構成音であること．
- (3) 前の音より 2 度以内の音，もしくは 1 オクターブ以内の和音の構成音であること．

(4) 減5度/増4度の進行を避けること。

また伴奏は、一定の音域内で和音の演奏を行う中音域パートと、各和音の根音をとるバスパートに分けられる。それぞれ、和声進行生成用リズムパターン選考および和声進行規則選考の結果より、音高が一意に求められる。

2.4 パラメータ対応関係の整理

これまでに提示した、各選考に対する凸包パラメータの適用について整理する。2.2.2 項で述べたように、凸包パラメータは「①クラス」「②物体数」「③凸包の面積率」「④凸包の変化率(面積変化率および重心変化率)」で構成される。表2に、各選考においてどのパラメータが適用されるか、対応の一覧を示す。

表2 各選考と凸包パラメータの対応表

区分	選考	凸包パラメータ
クラス	クラス選考	③ ④
リズム	和声進行生成用リズムパターン選考	③ ④
	旋律生成用リズムパターン選考	③ ④
和声	カデンツの主辞句の選考	② ③
	句の選考	② ③ ④
	カデンツの開始和音の選考	④
	和音の選考	④
旋律	暗意一実現パターン選考	② ④

ここまでの選考では、「①クラス」が適用されていない。この対策として、今回は各クラスを異なる楽器に対応させ、クラスに応じて主旋律の演奏楽器を変化させることとした。

今回割り当てを行った各楽器とクラスの対応を表3に示す。なお、VOC データセット[10]に含まれる20種類のクラス全てに異なる楽器を割り当てることは困難であると判断し、ここではある程度のグループ化を行っている。

表3 楽器とクラスの対応表

分類	楽器	クラス
Animals	Flute	Bird
	Clarinet	Cat, Dog
	Oboe	Horse, Sheep
	Bassoon	Cow
Vehicles	Trumpet	Aeroplane, Bicycle, Boat, Bus, Car, Motorbike, Train
Household	Marimba	Bottle, Potted plant, TV/Monitor
	Cello	Chair, Dining table, Sofa
Other	Violin	Person
(none)	Harp	(none)

3. 展示

本研究で制作したシステムは、『物体認識を用いた自動音楽生成器《Amguor》』として、2018年11月14日、15日、立命館大学以学館多目的ホールにて展示した(図9)。

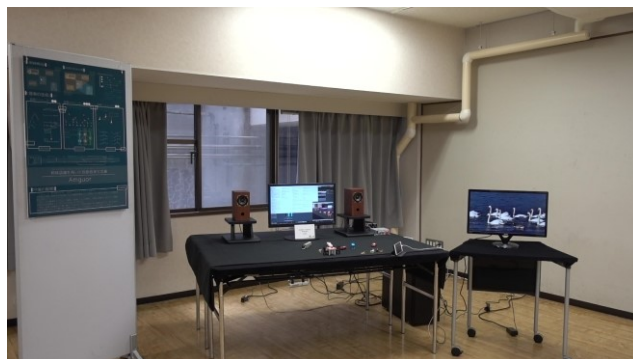


図9 『物体認識を用いた自動音楽生成器《Amguor》』

展示では、カメラの被写体となる物体検出対象として、いくつかの検出可能な物体の模型を机上に配し、カメラとあわせて作品の体験者が自由に触れて動かすことのできる環境を用意した。来場者は、スピーカより流れる音楽を聴取しながらカメラや模型を自由に操作し、撮影された映像に応じて音楽が変化する様子を確認していた。

作品を体験した来場者51人に対して実施したアンケート質問項目およびアンケート結果を以下に示す。

3.1 音楽は、映像と合っているように感じたか

この項目では、映像に合った音楽を生成するという目標がどの程度達成できたかを測るため、「生成された音楽は、映像と合っているように感じたか」について4段階での評価を求めた(図10)。その結果、回答者の73%が「とても感じた」「すこし感じた」と評価したことがわかった。

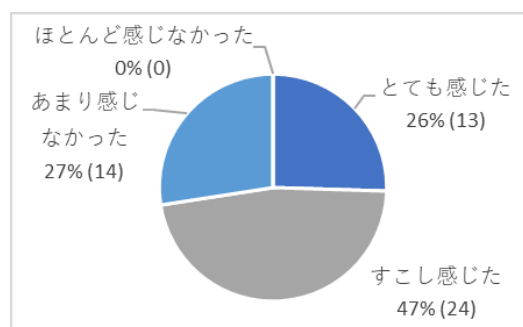


図10 音楽は、映像と合っているように感じたか

3.2 映像の中で、音楽と合っているように感じた点

3.1 節で「とても感じた」「すこし感じた」を選択した回答者に対し、「映像の、どのようなところが音楽に合ってい

ると感じたか」について、複数回答を許可した選択項目により評価を求めた（図 11）。その結果、「雰囲気」および「物の種類」「物の動き」が多く感じられたことがわかった。

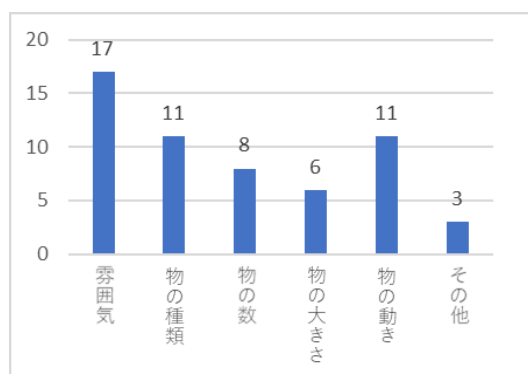


図 11 映像の中で、音楽と合っているように感じた点

3.3 音楽の中で、映像と合っているように感じた点

3.1 節で「とても感じた」「すこし感じた」を選択した回答者に対し、「音楽の、どのようなところが映像に合っていると感じたか」について、複数回答を許可した選択項目により評価を求めた（図 12）。その結果、特定の選択肢の突出は見られず、回答にばらつきがあることがわかった。

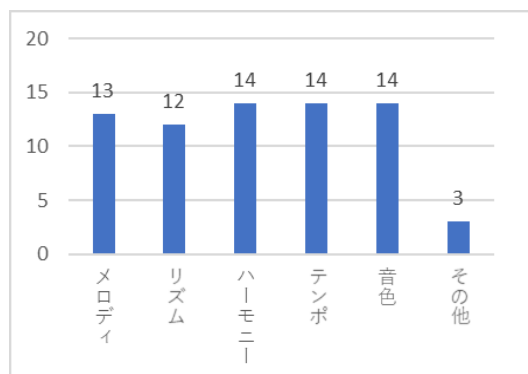


図 12 音楽の中で、映像と合っているように感じた点

4. 考察

3.1 節の結果より、本研究での提案手法を用いることで、映像に合った音楽の生成をある程度行うことができたと考えられる。また 3.2 節では、特に「物の動き」「物の種類」の要素が音楽に合っているという回答が見られた。これについて考察したい。

「物の動き」は、凸包パラメータにおける「変化率」と対応している。アンケート収集時には、「物の動きがテンポと連動している」という声が多く得られ、3.3 節においても「テンポ」は他の要素と同程度の回答が見られる。しかし、今回のシステムではテンポの変更は実装していない。それにもかかわらずテンポの変化が感じられた要因として

は、和音進行の速さがテンポ感に影響を与えている可能性が考えられる。和音進行の速さは、2.3.2 項の和声進行生成用リズムパターン選考によって決定されており、ここでは凸包の面積率および変化率が反映されている。このうち、面積率は「物の大きさ」との対応が考えられるものの、凸包は複数の物体を包含し得るため、一概に個々の物体の大きさが直接関係しているとは言えない。そのため、よりパラメータとの関連の強い「物の動き」の方が、大きく変化として感じられたのではないかと考えられる。

また、「物の種類」は凸包パラメータの「クラス」と対応しており、楽器との紐づけにより音楽へと反映される。今回のシステムでは、1 小節を単位として逐次的な生成が行われており、演奏楽器の切り替え時点はこれに対応している。「物の種類」がわかりやすい変化となったのは、音色の変化が聴衆にとって知覚しやすい変化であったと考えられることに加え、変化のタイミングが強拍で、かつ変化による影響が一定時間持続することも関係しているのではないかと考えられる。ただし、各クラスをどの楽器と対応づけるかについては検討が必要であり、今後の課題とする。

一方で、3.1 節では「すこし感じた」が約半数を占めることや、3.2 節では「雰囲気」が最も多い回答であることが結果として得られている。3.2 節における「物の数」や「物の大きさ」の反映が比較的弱いという結果と併せて、パラメータの適用規則や対応関係については妥当性の検討が必要である。

参考文献

- [1] 尾鼻崇, "映画音楽からゲームオーディオへ：映像音響研究の地平", 晃洋書房, 2016.
- [2] 山本敏生, 宝珍輝尚, 野宮浩揮, "動画をもとにした自動作曲", 平成 21 年度情報処理学会関西支部大会講演論文集, 2009.
- [3] 佐々木大輔, 中島武三志, 菅野由弘, "水族館における魚の動きを入力としたリアルタイム音楽生成システムの提案", インタラクシオン 2017, 1-505-24, 情報処理学会, 2017.
- [4] 菊池有恒, "楽典：音楽家を志す人のための", 音楽之友社, 1988.
- [5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "SSD: Single Shot MultiBox Detector", arXiv:1512.02325v5, 2016.
- [6] 松原正樹, 深山寛, 奥村健太, 寺村佳子, 大村英史, 橋田光代, 北原鉄朗, "創作過程の分類に基づく自動音楽生成研究のサーベイ", コンピュータソフトウェア, Vol.30, No.1, pp.101-118, 2013.
- [7] 東条敏, 平田圭二, "音楽・数学・言語：情報科学が拓く音楽の地平", 近代科学社, 2017.
- [8] Fred Lerdahl, Ray S. Jackendoff, "A Generative Theory of Tonal Music", The MIT Press, 1983.
- [9] 西田昌史, 東条敏, 佐藤健, "HPSG を用いた楽曲の和声解析", 情報処理学会研究報告音楽情報科学 (MUS), Vol.2002-MUS-047, No.100, pp.127-134, 2002.
- [10] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective", International Journal of Computer Vision, Vol. 111, No. 1, pp.98-136, 2015.