画像の類似度を用いたダンス動画モーション訂正手法

又吉 康綱* 小山 裕己† 深山 覚† 後藤 真孝† 中村 聡史*

概要．近年，SNS や動画共有サイトには多くのダンス動画の投稿が行われており，ダンスというコンテンツへの注目度は高い．こうしたダンス動画をダンスの内容に即して探索・推薦するには，ダンス動画を時系列のモーションデータとして扱う必要があるが，ダンス動画の量は膨大でありそのモーションを人手ですべて付与していくのは困難である．ここで動画からモーションを自動抽出する深層学習技術を利用することが考えられるが，自動抽出結果には多くのエラーが含まれるため，手作業でのモーションの訂正が必要である．そこで我々は，ダンス動画のフレーム間の画像の類似度が本来抽出されるべきポーズの類似度と高い相関を持つことに着目し，それを活用することで，モーション訂正作業を支援する手法を提案する．本稿では proof-of-concept であるウェブシステムの実装と，そのシステムに実装されたモーション訂正支援インタラクションについて述べるとともに，その特性について考察を行う．

1 はじめに

多くの人がダンス動画を SNS や動画共有サイトへ投稿することで，感情表現をしたり，コミュニケーションをしたりしている．そのため，バリエーション豊富なダンス動画がインターネット上には膨大に存在しており，例えばニコニコ動画 [1] 上にはダンス動画によく付与される「踊ってみた」タグの動画が 17 万件以上 [2] 存在している．

こうした膨大なダンス動画があることを背景として，ダンスのモーションデータ（本稿では人間の動作を表現するための仮想的な人体模型における複数の関節の位置や角度の時系列情報を指す）を利活用する研究は広く行われている．モーションデータを用いることで，動画の表層的な見た目の分析に加えて，動作の内容に即した検索や機械学習が行える．例えば Tsuchida ら [3] はユーザ自身のダンスモーションをもとにダンス動画を検索する Query-by-Dancing を実現している．また機械学習を用いることで，仮想キャラクタのダンスの動きを自然にする手法 [4] やダンスモーションからダンサーの感情を推定する手法 [5] などが提案されている．こうした研究が更に発展していくには，検索データベースや機械学習の学習データのために膨大なダンスのモーションデータが必要となるが，動画共有サイトに存在するのはダンスの動画のみであり，そのモーションデータは多くの場合存在していない．

ダンスモーションデータの作成方法の一つとして，手作業でダンス動画内の動作を再現するようにモーションデータを作成（トレース）することがある．具体的には，人が実際に踊っている動画と画面上に描画される人体模型を重ね合わせながら，キーフレームアニメーションを手動で作成する．この際，動画の見た目から奥行きを推測したり，人間として破綻した動きが生じないように考えたりと，多くの時間と手間がかかる．一方，映像制作や学術研究用途のダンスモーションデータ作成においては，精度が高いモーションキャプチャシステムを用いることが多い．しかし，実際に人に踊ってもらう必要があるため，データセットの新規作成に時間がかかるうえに，専用のスタジオが必要になるなど，その作成コストが高い．こうした問題により，モーションのデータセットの整備はまだ発展途上にある．

一方，OpenPose [6] などの近年の深層学習技術の飛躍的な進歩により，ダンスモーションの作成においてダンス動画からポーズを自動的に推定し活用できる見込みが高まってきた．しかし，連続した滑らかなモーションを含む動画であっても，その動画から得られたポーズ推定結果には不連続なちらつきのようなエラーが含まれることがある．また，類似の振り付けを踊っている動画であったとしても，人間の目には分かりにくい微細な違いから，例えば足の前後関係が入れ替わるなど，大きく異なるポーズ推定結果が得られることがある．このような推定の不安定性の問題に対応するためには，手作業での訂正が必要となる．モーションを訂正する方法は現時点で確立されておらず，1 フレーム毎に動画と見比べながらモーションを訂正していく必要がある．この訂正作業の負担は大きいため，推定したモーションを効率的に訂正できる仕組みが必要である．

そこで本研究では，動画から自動推定されたモー

* 明治大学，† 産業技術総合研究所

ションデータを効率的にエラー検知・訂正するための枠組みを提案する．また，その枠組みの一部を実装した proof-of-concept のウェブシステムを紹介する．将来的に，本手法と深層学習技術を組み合わせることで, SNS や動画共有サイトに存在するバリエーション豊かなモーションの利活用が可能になると期待できる．

このようなエラーを含むモーションデータの訂正を効率化するにあたって，我々は 1 つのダンス動画内において類似した部分ダンスモーション（振り付け）が複数回登場しうることに着目した．例えば類似した部分ダンスモーションが動画中に 4 つある場合，仮にその 1 つが推定エラーを含んでいたとしても，他の 3 つの推定結果を参照して効率的に訂正することが可能だと期待される．本研究のキーアイデアは，エラーを含むモーションを解析してそのような類似部分を探索するのではなく，元のダンス動画自体を解析して探索することである．具体的には，動画のフレーム間の類似度が，本来推定されるべきポーズ間の類似度と強い相関を持つと考えられるため，動画フレームに基づく類似度行列を計算し，それを用いて類似ポーズの探索を行うことで，モーションの訂正を支援する手法を考案した．ポーズ推定とは異なるアプローチで類似度の計算を行うことは，ある意味で推定の多重化を行っていると考えることができる. 多重の推定結果の相関に着目することで，ポーズ推定の不安定性を由来とするエラーをうまく検知・訂正するというのが，本研究の目標である．

本稿の貢献は以下の通りである．

- ダンス動画から自動抽出されたエラーを含むモーションデータに対し，手作業で訂正することを支援するという新しい問題に着目し議論すること
- 動画のフレーム間類似度に基づいてモーションのエラー検出及び訂正を支援する枠組みの提案
- モーション訂正インタラクションの提案とその proof-of-concept 実装の提示

2 関連研究

ダンスの動画やモーションを活用した研究として，岡田ら [7] のダンスモーションのセグメンテーションに関する研究，Senecal ら [8] の動きとダンサーの感情の関係性を機械学習した上での感情の自動認識に関する研究，古市ら [9] のダンスモーションから個性を抽出する研究などがある．これらの研究にはダンス動画やモーションが大量に必要であり，不足していることが課題になっている．そのため，本提案を活用することが課題の解決につながる．

モーションキャプチャ使用時に生じる特有のエラーを訂正する研究は多く存在する．Aristidou ら [10] は，モーションのフレーム間類似度を分析することによってエラー検出と訂正を行っている. また，Holden [11] は，モーションデータに人工的なノイズを乗算し機械学習を行うことによってモーションのエラーを除去する手法を提案している. 本研究は，モーションキャプチャによるエラー訂正とは性質が異なる自動推定によるエラー訂正を対象としているだけでなく，ダンス動画というモーションデータに付随するデータを活用してエラー訂正を行う点で，これらの研究とアプローチが異なる．

モーションの編集を行うインタラクションについても研究がある. SketchiMo [12] はモーションの軌跡を表示したり逆運動学を用いてポーズの推薦を行ったりすることで効率的な編集を実現している．これに対し，本研究ではダンス動画中のモーションを再現することが目標であり，自由な創作を支援することを目標とした研究とは問題の性質が異なる．

以上のように，モーションのエラー訂正と編集インタラクションに関する研究は多くなされている．これに対して本研究では，ダンス動画から自動推定されたモーションデータを訂正する問題に即したエラー検出手法とモーション編集のインタラクションのあり方を提案する．

また，本研究のように類似度行列を活用した研究は多く存在する．例えば，音楽情報処理における楽曲構造推定 [13] や，繰り返し構造を含む動画コンテンツの操作 [14] などにも使われる. 本研究では，モーションの訂正タスクに対して動画フレームの類似度行列を活用するという，ドメインの異なるものへの活用という点で工夫がある．

3 モーションを訂正する際の問題点

3.1 目視によるエラー箇所の発見

動画から自動推定したモーションデータには，当面の技術的な限界により間違った推定結果が含まれてしまうことが避けがたい．間違った推定がされている箇所は，動画とモーションデータを見比べながら確認し，発見する必要がある．このエラー箇所の発見の労力を軽減することが望ましい．

3.2 手作業でのモーション訂正

前節のエラー箇所の発見に加え，モーションの訂正は手作業で行う必要がある．動画から立体的な奥行きや傾きを推察しながら，手足の交差や関節の曲がり方などを細かい時間単位で訂正する必要がある. サビなどでは同じ振りであることも多く，訂正を終えた他の時刻からモーションをコピーすることもあ

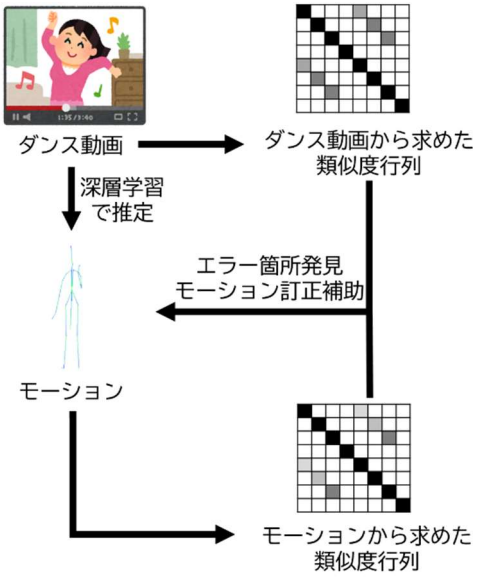

図 1. 提案する枠組みの概略図．類似度行列を活用することで訂正すべきエラーの発見及び訂正の支援を行う．なお，本稿ではダンス動画から求めた類似度行列（右上）を用いてモーションの訂正を行う部分について議論する．

るが，そのような操作を促すシステムはなく，動画を確認しながら判断するしかないため，負担を伴う．

4　動画のフレーム間類似度を活用したモーション訂正支援の枠組み

3.1 節で述べたエラー箇所の発見の労力を軽減するためには，「ダンス動画上では同じポーズをとっているにも関わらず，モーションデータ上では異なるポーズをとっているようなフレーム」を検出することが有効だと思われる．そこで我々は，モーションデータだけでなく，ダンス動画も併せて分析し比較することで，モーション訂正が必要であるエラー箇所の発見を促すという新しい枠組み（図 1）を提案する．より具体的には，ユーザが訂正したいモーションデータとそのダンス動画を指定すると，ダンス動画中のフレーム画像を解析して類似度行列を計算し（図 1 右上），またエラーを含むモーションに基づく類似度行列を計算し（図 1 右下），両者の情報を合わせることで，モーションデータ中のエラーを発見するという枠組みである．

更に，ダンス動画に基づく類似度行列を計算するという本枠組みには，もう一つ有用な側面がある．それは，モーションデータ中のある時刻のポーズを訂正する際に，そのダンス動画の類似度行列に基づき，本来そのフレームで推定されるべきだったポーズと類似するポーズを含んでいる可能性が高いフレームの探索が可能なことである．ダンスの振り付けには曲中の異なる箇所で同じ動きが繰り返されることが頻繁にあるため，このような動画を元にした探索を行うことにより，モーションのエラー訂正に有用な情報を得ることができる．これにより，3.2 節で述べたモーション訂正の負担を軽減するためのインタラクションを考案することが可能になる．

本稿では特に後者の側面，すなわちダンス動画に基づく類似度行列を活用することでモーション訂正時に有用な情報を取得し，モーション訂正作業を支援することについて議論する．なお本稿では，カメラの位置が固定されており，かつ 1 人で踊っているダンス動画を対象とする．

5　動画のフレーム間類似度

5.1　類似度関数

動画内の任意の 2 つの動画フレーム間の類似度を測る関数$\mathrm{sim}(\cdot,\cdot)$を

$$\mathrm{sim}(I_i, I_j) = \frac{1}{\sum_{x=1}^{w}\sum_{y=1}^{h}\left\|I_i(x,y) - I_j(x,y)\right\|_1 + 1} \quad (1)$$

と定義する．ただし，I_iはi番目の動画フレーム，I_jはj番目の動画フレーム，$w, h \in \mathbb{N}$はそれぞれ横方向と縦方向のピクセル数を表し，$I(x,y) \in \mathbb{R}^3$はその動画フレームの座標(x,y)におけるピクセルの RGB 値を表すとする．この関数の値が大きいほど対象とする 2 つの動画フレームが類似していることを表す．

5.2　類似度行列

類似度関数$\mathrm{sim}(\cdot,\cdot)$を用いて，対象とする動画の自己類似度行列（self-similarity matrix；本稿の他の箇所では単に類似度行列と呼ぶ）$\mathbf{S} \in \mathbb{R}^{n\times n}$を計算する．ここで，$n \in \mathbb{N}$は対象の動画のフレーム数を表す．具体的には，行列のi列j行要素を

$$S_{ij} = \mathrm{sim}(I_i, I_j) \quad (2)$$

とする．図 2 は公開されているダンス動画 [15] から計算した類似度行列の可視化結果である．左上から右下の対角成分が濃い青色となっているのは，同じフレーム同士を比較しており，類似度が最大値（具体的には 1）をとるためである．また，長さが様々な青色の斜線があるが，これはモーション中の対応する区間同士が互いに類似している（すなわち，似た振り付けが繰り返されている）ことを意味している．このような互いに類似したモーションを検出しておくことで，モーション中のあるフレームに訂正が必要な場合に，類似しているもう一方のフレームのポーズを参考にして訂正を行うことができる．

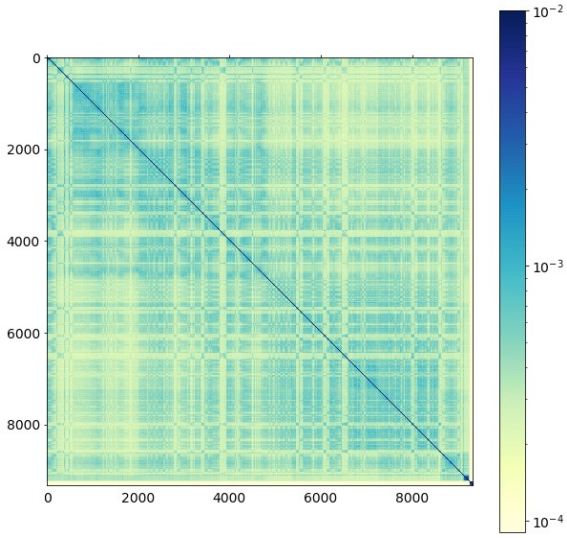

図 2. 動画フレームに基づく類似度行列の可視化．縦軸と横軸は動画のフレーム番号に対応している．

図 3. 左に示す動画に対し，類似度の上位 3 件を右側に表示した結果．フレーム番号 5302 はそのフレーム自身のため最も類似しており，フレーム番号 7826 と 8418 の順で類似していることを示している．

5.3 動画フレームの類似度を活用することの妥当性

前節の方法で計算した動画フレームに基づく類似度行列によって，実際に類似したポーズを含むフレームが検出可能かを確かめるため，ダンス動画中のあるフレームに対して，そのフレームとの類似度が高いフレームを表示するシステムの実装を行った．図 2 と同じダンス動画に対して適用した結果例を図 3 に示す．ただし，フレームが近い画像同士でも類似度が高くなってしまうため，前後 60 フレームを除外した．最も類似度が高いフレームは必ず左側に表示されているフレーム自身（ここではフレーム番号 5302）となる．2 番目及び 3 番目に類似度が高いフレームは，この例ではフレーム番号 7826 と 8418 であり，実際にこれらのフレームも似たポーズを含んでいることが観察できる．このように，動画のフレーム画像間の類似度は，当該フレームに含まれるポーズ間の類似度と相関があることが示唆された．

5.4 動画フレームに基づく類似度計算のリミテーション

5.1 節の類似度行列計算は，動画の背景に大きな影響を受ける．実際に，夜間にライトアップされた観覧車が背景にある動画で類似度行列を計算した場合，観覧車のライトの点滅パターンが類似度行列に反映されてしまい，類似の振り付けを行っている部分間でも類似度が低くなってしまった．他にも，別の人物が映り込んでいる動画や，カメラが固定されていない動画，曲の途中で別の場所で撮影したカットに切り替わるような動画では，現在の計算手法では効果的に類似したポーズを検出することが難しい．

図 4. システムスクリーンショット．

6 モーション訂正のためのインタラクション

実際にモーション訂正を行うためのインタラクションと，それを実証するためのプロトタイプシステムについて述べる．実装には，JavaScript でウェブシステムを作成し，公開されているダンス動画 [15] とトレースされたモーションデータを使用させて頂いた．

6.1 類似度行列の事前計算

実行時の処理を高速化するため，類似度行列を事前計算した．その際，高画質なダンス動画に対して類似度行列を計算すると，処理に多くの時間がかかってしまうため，Liu ら[16]の手法を用いて人が動いている範囲のみを切り出すことで画像サイズを小さくし，さらに縮小して使用した．計算した類似度行列から，フレームごとにそのフレームの画像と前後 60 フレームを除く類似度の高い画像を持つフレームを探索し，それをファイル保存して使用した．

6.2 インタラクション機能

システムの画面は大きく分けて 3 つで構成されている（図 4）．左上にダンス動画，上部にモーションデータの描画，下部にキーフレームアニメーションのタイムラインがある．モーションデータの描画画面では，作業中のモーションの動きをプレビューすることができる．タイムラインでは，アニメーション制作ソフトウェアと同様に，モーションデータのボーン毎のキーフレームアニメーションが表示されており，関節位置や角度などの基本的な編集をすることができる．また，タイムラインの再生ボタンを

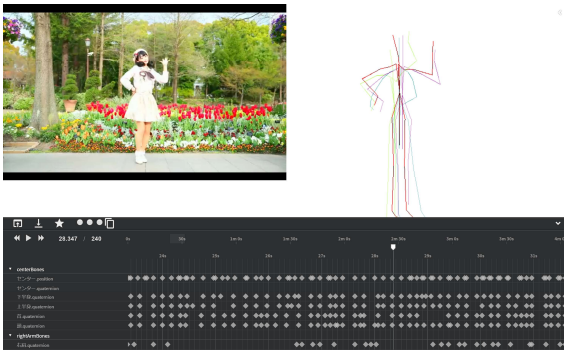

図 5. 類似度行列に基づいたポーズ候補の推薦.

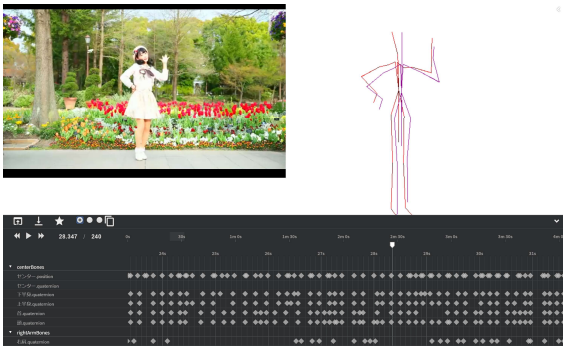

図 6. 1 つのポーズ候補の選択.

押すか，シークバーを移動することで，ダンス動画とモーションデータを同期させて再生できる．

本システムに固有のインタラクション機能として，類似度行列に基づく訂正内容の推薦機能を実装した．タイムライン上部にある星マークをクリックすると，現在の再生時間に対して，類似度行列に基づき選出した，モーションデータ内の他のフレームにおけるポーズ 3 つが薄くオーバーレイされる（図 5）．さらに，ラジオボタンを選択することによって，推薦内容を適用したポーズを 1 つずつプレビューすることができる（図 6）．また，コピーマークを押すことによって，推薦されたポーズを適用することができる．この機能により，3.2 節の手作業でのモーション訂正の効率化が期待できる．

6.3 プロトタイプシステムの有効性の検証

深層学習に基づくポーズ推定器の出力を用いて検討を行う前段階として，ここでは既存のモーションに人工的なノイズを加えることで深層学習の不安定性に由来するエラーを擬似的にシミュレートし，有効性の検討を行う．具体的には，サビ区間において類似した振り付けが複数回現れるダンス動画について，その一つのサビ区間を Songle [17] より求め，その区間のモーションの関節角度にランダムなノイズを与えたモーションを作成し，プロトタイプシステムで訂正することが可能かどうかの確認を行った．実際にモーションの訂正を行ってみたところ，図 7

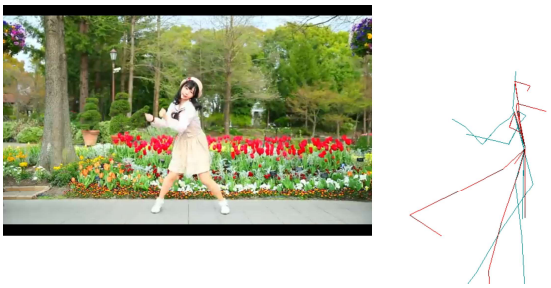

図 7. 赤色ボーンが擬似的に作成した訂正が必要なモーション．青色ボーンが類似度行列から推薦された訂正の候補.

のように適切な推薦がなされ，モーションの訂正を行うことができた．しかし，細かい手や首の動作を正確に推薦することはできなかった．また，訂正ができなかった例として，訂正が必要なサビ区間内の同様なモーションも推薦された事により，適切な候補を選ぶことができないこともあった．推薦精度の向上は今後の課題である．

6.4 今後検討していくインタラクション機能

本研究では類似度に着目したエラー検出・訂正の枠組み（図 1）を提案したが，現時点ではその全体の実装・検証には至っていない．未実装の側面として，モーションデータに関する類似度行列の計算と，それを活用したインタラクション機能の検討が挙げられる．Kovar ら [18] の手法などを利用することでモーションに関する類似度行列は計算できるため，これと動画フレームに基づく類似度行列との差異を分析することで，エラー訂正が必要なフレームを抽出してタイムライン上に可視化したりするインタラクションを実現できると期待できる．どのような可視化が効果的かは今後の調査が必要である．また，動画フレームに基づく類似度だけでなく，モーションに基づく類似度を用いて，身体の一部分のみを推薦や訂正に用いること，候補として挙がったポーズ同士を補間して訂正に用いることなど，様々なインタラクションの可能性がある．

7 まとめと今後の課題

ダンス動画を画像処理することによって，モーションデータ中の互いに類似している箇所を探索し，それに基づいてダンス動画に対応するモーションデータの訂正を支援する枠組みを提案し，その一部をウェブシステムとして実装した．

現在の動画フレーム間の類似度関数の定義は単純なものを採用しているため，入力に用いることができる動画はカメラが固定され 1 人でダンスしているものに限定されている．ニコニコ動画などで共有される「踊ってみた」動画にはこうした条件を満たすものが多いものの，より多様なダンス動画に適用で

きるようにすることは重要な将来課題である．深層学習技術により動画内の身体位置に応じて類似度計算の際に重み付けをすることや，色空間上の類似度だけでなくオプティカルフローに基づく類似度等を採用するなどの方法が考えられる．

ダンスの振り付けと楽曲構造は関連度が高いと考えられるため，ダンス動画に含まれる楽曲情報を活用したモーション訂正支援には大きな将来性がある．例えば，現在は 1 キーフレーム単位での推薦や適用を行っているが，ダンス動画に用いられている曲の楽曲構造を Songle [17] から取得し，小節単位同士や楽曲のサビ同士のモーションの類似度を活用したモーションの訂正を行えるようにすることが考えられる．また，Cheng ら [13] の類似度行列から楽曲構造を分析する手法をモーションデータに応用して，ダンスの振り付けの構造分析を行い，その結果に基づいてより効果的にモーションを訂正できる手法も検討していきたい．

本提案手法を深層学習に基づく動画からのモーション抽出モジュールと連携させ，ダンス動画の SNS や動画共有サイトの URL を入力するだけで，モーションデータの抽出と訂正がブラウザ上で手軽に行えるシステムの構築が可能である．このようなシステムによって，多くの人による手軽なモーション訂正に支えられた，世の中のあらゆるダンス動画に対応するモーションデータベースが整備されるようになり，それを活用することでダンスの内容に踏み込んだ様々なアプリへの応用が可能になると期待される．

謝辞

本研究の一部は JST ACCEL（JPMJAC1602）の支援を受けた．本稿では，ニコニコ動画上のダンス動画 [15] を，オリジナル振り付けで踊って投稿した「足太ぺんた」様の承諾を得て使用した．

参考文献

[1] niconico(ニコニコ). https://www.nicovideo.jp/. （2019/07/31 確認）

[2] 人気の「踊ってみた」動画 171,323 本 - ニコニコ動画. https://www.nicovideo.jp/tag/踊ってみた. （2019/07/31 確認）

[3] S. Tsuchida, S. Fukayama and M. Goto. Query-by-Dancing: A Dance Music Retrieval System Based on Body-Motion Similarity. In *Proc. MMM '19*, pp.251–263, 2018.

[4] C. Mousas. Performance-Driven Dance Motion Control of a Virtual Partner Character. In *Proc. VR '18*, pp.57–64, 2018.

[5] A. Aristidou, P. Charalambous and Y. Chrysanthou. Emotion Analysis and Classification: Understanding the Performers' Emotions Using the LMA Entities. *Comput. Graph. Forum*, 34(6):262–276, 2015.

[6] Z. Cao et al. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proc. CVPR '17*, pp.1302–1310, 2017.

[7] 岡田成美, 福里司, 岩本尚也, 森島繁生. 振り付けの構成要素を考慮したダンスモーションのセグメンテーション手法の提案. 情報処理学会 研究報告グラフィクスと CAD, Vol.2014-CG-156, Issue.9, pp.1–6, 2014.

[8] S. Senecal et al. Continuous body emotion recognition system during theater performances. *Comput. Animat. Virtual Worlds*, 27(3–4):311–320, 2016.

[9] 古市冴佳, 阿部和樹, 中村聡史. ヒップホップダンスにおける骨格情報を用いた個性抽出の検討. 情報処理学会 研究報告エンタテインメントコンピューティング, Vol.2018-EC-50, Issue.23, pp.1–9, 2018

[10] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou and A. Shamir. Deep motifs and motion signatures. *ACM Trans. Graph.,* 37(6):1–13, 2018.

[11] D. Holden. Robust solving of optical motion capture data by denoising. *ACM Trans. Graph.,* 37(4):165:1–165:12, 2018.

[12] B. Choi et al. SketchiMo: Sketch-based Motion Editing for Articulated Characters. *ACM Trans. Graph.,* 35(4):1–12, 2016.

[13] T. Cheng, J. B. L. Smith and M. Goto. Music Structure Boundary Detection and Labelling by a Deconvolution of Path-Enhanced Self-Similarity Matrix. In *Proc. ICASSP '18*, pp.106–110, 2018.

[14] A. Schodl et al. Video textures. In *Proc. SIGGRAPH '00*, pp.489–498, 2000.

[15] 【足太ぺんた】恋愛デコレート 踊ってみた【オリジナル振付】 - ニコニコ動画. https://www.nicovideo.jp/watch/sm29383900. （2019/07/31 確認）

[16] W. Liu et al. SSD: Single Shot MultiBox Detector. In *Proc. ECCV '16*, pp.21–37, 2016.

[17] 後藤真孝, 吉井和佳, 藤原弘将, Matthias Mauch, 中野倫靖. Songle: 音楽音響信号理解技術とユーザによる誤り訂正に基づく能動的音楽鑑賞サービス. 情報処理学会論文誌, Vol.54, No.4, pp.1363-1372, 2013.

[18] L. Kovar, M. Gleicher and F. Pighin. Motion Graphs. *ACM Trans. Graph.*, 21(3):473–482, 2002.