振りの理解を助けるためのダンス動画の自動分割

遠藤 輝貴 * 土田 修平 † 五十嵐 健夫 *

概要. ダンサーがダンスの振りを覚える際，振りが短時間の動きに分割されていると理解や習得が容易になる．しかしダンス動画から振りを覚える場合はそのような振りの分割が予め存在しないため，学習者は振りを理解するために適切な分割位置を自分で見つける必要があり，これが振りの習得を困難にしていると考えられる．そこで我々はダンス動画の振りを個々の動きへと自動で分割する手法を提案する．提案手法では動画中のダンサーの身体や手の位置から視覚特徴量を，動画中の音楽から聴覚特徴量をそれぞれ計算し，これらの特徴量を Temporal Convolutional Network (TCN) に入力して，出力として得られる分割可能性のピークを検出することで動画の分割位置を求める．AIST Dance Video Database の動画から作成した学習データを用いた実験の結果，ダンス動画の分割に提案手法の視覚，聴覚特徴量の両方が役立つことを確認した．本論文では提案手法の詳細や実験結果と，自動分割の応用例として開発したダンス学習支援システムについて述べる．

1 はじめに

ダンサーがダンスの振りを習得する主な手段として，指導者から直接教わる方法と，ダンス動画の振りを真似する方法とがある．前者の場合は指導者が振りを短時間の動きに分割してそれぞれの動きを 1 つ 1 つ順番に教えていくことが一般的である．例えばダンスの「技」と呼ばれる一連の動きや，切れ目なく滑らかに繋がる動きなどは 1 つの塊として練習すると効率が良く，このように振りの構成要素を考慮して分割することで，学習者は振りの理解や習得が容易になる．これに対し後者の場合，通常のダンス動画ではそのような振りの分割が存在しないため，学習者は自ら踊ることのできるレベルまで振りを理解するために適切な分割位置を自分で見つける必要がある．ダンス経験の浅い人は上述の技や身体の流れに関する知識が少ないため，自分で振りを分割することが難しく，このことがダンス動画からの振りの習得を困難にしていると考えられる．

そこで我々はダンス動画の振りを時系列方向に，個々の動きへと自動で分割する手法を提案する．ダンスは音楽に合わせて身体を動かすため，振りの分割にはダンサーの動きだけではなく動画中の音楽も有用な情報になり得る．そこで提案手法ではまず，入力となるダンス動画から視覚的，聴覚的な特徴量を抽出する．視覚特徴量としては動画内のダンサーの身体や手のキーポイントの位置を推定し，その速度成分を使用する．また，聴覚特徴量としては動画に含まれる音楽からメルスペクトログラムを計算し，CNN で畳み込んだものを使用する．次にこれらの特徴量を Temporal Convolutional Network (TCN) [1] の入力として与え，出力として得られる分割可能性のピークを検出することで動画の分割位置を求める．

提案モデルの学習には，ダンス動画に対してその動画の分割位置をアノテーションした学習データが必要となる．我々は AIST Dance Video Database [12] のダンス動画に対して振りの分割位置を手動でアノテーションし，合計 1410 本の学習データを用意した．この学習データを使い，視覚特徴量と聴覚特徴量の両方を使用した場合とどちらか一方のみ使用した場合とでモデルの性能を数値的に評価したところ，両方の特徴量を使用するモデルが最も性能が高くなった．この結果から，ダンス動画の分割には視覚的なダンサーの動きの情報と聴覚的な音楽情報の両方が重要であることが分かる．

提案手法の応用例として，自動分割の結果を利用したダンス学習支援システムを作成した．具体的には，推定した分割位置で区切った動画をループ再生することで，同じ動きの繰り返し練習を支援することができる．ループ再生するセグメント間には重なりを持たせることで，動きのつながりを理解しやすくしている．更にユーザが自動分割の細かさを調節できる機能も追加することで，ユーザの好みや熟練度に合わせた動画分割及び練習ができる．

2 関連研究

2.1 振りの分割の有効性

Rivière ら [9] は，ダンサーが動画から振りを覚える際の行動を調査し，ダンサーは動画の振りを分割して覚えていることを明らかにした．更に彼らは，手作業でダンス動画を分割するためのツールを開発し，ダンサーや指導者に対する被験者実験を行った．その結果，経験の浅いダンサーの場合は，自ら分割

* 東京大学
† 神戸大学

した動画よりも指導者が分割した動画のほうが振りを習得しやすいことを示した．この結果から我々は，ダンス動画を適切に分割することは振りの習得に有効であると仮定し，動画の分割作業を指導者の代わりにコンピュータで自動化することを目的とする．

2.2 ダンスモーション分割

Shiratori ら [10] は，日本舞踊のモーションデータと使用されている音楽情報を入力とし，ルールベースでモーション分割を行った．しかし，このルールは日本舞踊に特有の特徴を使用したものになっており，ヒップホップなどの一般的なダンスジャンルに適用することは難しい．Okada ら [6] は，CG キャラクタのダンスモーションを生成するために，既存のモーションデータをルールベースで分割する手法を考案した．しかし，この手法では音楽の拍の位置は既知でなければならず，拍の位置が分からない通常のダンス動画には適用できない．

ダンス動画の拍の位置を推定するために，Pedersoli と Goto [8] は，ダンス動画の映像情報を用いた手法を考案している．この研究は，動画から推定したダンサーのポーズ情報を TCN の入力に使っている点で我々の手法と類似しているものの，動画中の音楽情報を使用していない点や推定する対象が大きく異なる．

2.3 ダンス動画からのダンス学習支援システム

ダンス動画からの振りの習得を支援するアプリはいくつか存在する．ウゴトル [14] では，ダンス動画を閲覧しながら振りを覚えるための左右反転や速度変更の機能が充実しているが，動画を分割して一部だけ再生することができない．SymPlayer [11] は動画の左右反転や速度変更に加え，ループ再生機能が存在する．但し，ループ区間の始点と終点を手動で設定しなければならない上に，指定できる区間も 1 つのみなので，振りを複数の短い動きに分割して順番に練習するには不向きである．

また，ダンス動画に対して処理を行うことで振りの習得を支援する研究も存在する．例えば，斎藤ら [16] は，ダンス動画に対して動きのニュアンスを表すオノマトペを表示することで，学習者が動きのニュアンスを理解してダンスを学びやすくしている．Zhou ら [15] は，複数人で踊っているシンクロダンスの動画に対して，動画に映る複数人の動きの類似度を推定し，身体部位別，時間別に動きのずれを表すヒートマップを表示することで，シンクロダンスの練習を支援するシステムを提案している．Tsuchida ら [13] は，深層学習技術を用いて学習者が手本となるダンスをマスターして踊っている動画を生成し，それを学習者自身に見せることでダンスの学習を支援することを提案している．

Rivière ら [9] は，手作業でダンス動画を分割するためのツール「MoveOn」を開発した．このツールでは動画を分割して複数のセグメントを作成し，各セグメントに対してループ再生や速度変更の設定ができる．しかし，ダンス動画の分割を人手で行う必要があるため，適切な分割位置が分からないダンス初心者が使用するのは難しい．また，たとえ熟練者であってもセグメントを 1 つ 1 つ作成するのは手間がかかる作業である．

これに対し，我々が 6 章で提案するシステムは，自動分割した動画セグメントをユーザに表示するため，手動分割の手間が省けるだけでなく，初心者によるシステムの利用やダンスの練習支援ができると考えられる．

3 提案手法

3.1 アルゴリズムの概要

本研究の目的は通常のカメラで撮ったダンス動画を入力として，その動画の振りの分割位置を推定することである．提案手法の概要を図 1 に示す．はじめに，動画内のダンサーの動きから視覚特徴量を，使用されている音楽データから聴覚特徴量を，それぞれ後述の方法で計算する．次にこれらの特徴量を TCN の入力として使用する．TCN の出力は，動画の各フレーム t に対してそのフレームで動画が分割される可能性の高さを表す t の関数 $p(t)$ であり，この $p(t)$ のピークを検出することで，動画の分割位置を決定する．以降の節ではアルゴリズムの各段階の詳細を記す．

3.2 視覚特徴量の抽出

視覚特徴量は動画内のダンサーの動きから計算される．まず，動画内のダンサーの身体や手のキーポイントを AlphaPose [4] を用いて検出する．本研究で用いるキーポイントは身体が 26 個，左右の手がそれぞれ 21 個の合計 68 個であり，キーポイントの位置は動画の左上が $(0, 0)$，右下が $(1, 1)$ になるように正規化した座標値で求める．さらにダンサーの動きの速度に注目するため，1 フレーム前からのキーポイント位置の変化を求め，これを視覚特徴量とする．即ち，フレーム t で検出した i 番目のキーポイントの位置を $\boldsymbol{k}_i(t) \in \mathbb{R}^2$ とすると，視覚特徴量 $\boldsymbol{v}(t) \in \mathbb{R}^{68\times 2}$ の i 番目の要素 $\boldsymbol{v}_i(t) \in \mathbb{R}^2$ は以下の式で求めることができる．

$$\boldsymbol{v}_i(t) = \frac{1}{2}\big(\boldsymbol{k}_i(t) - \boldsymbol{k}_i(t-1)\big) \tag{1}$$

但し，最初の $\frac{1}{2}$ は特徴量の値を $[-0.5, 0.5]$ の範囲に収めるための定数である．

提案手法では 1 人のダンサーのみが映っているダンス動画を対象としているが，AlphaPose の誤検出によって動画の 1 フレーム中に複数の人間が検出されてしまうケースや，1 人も検出されないケースも

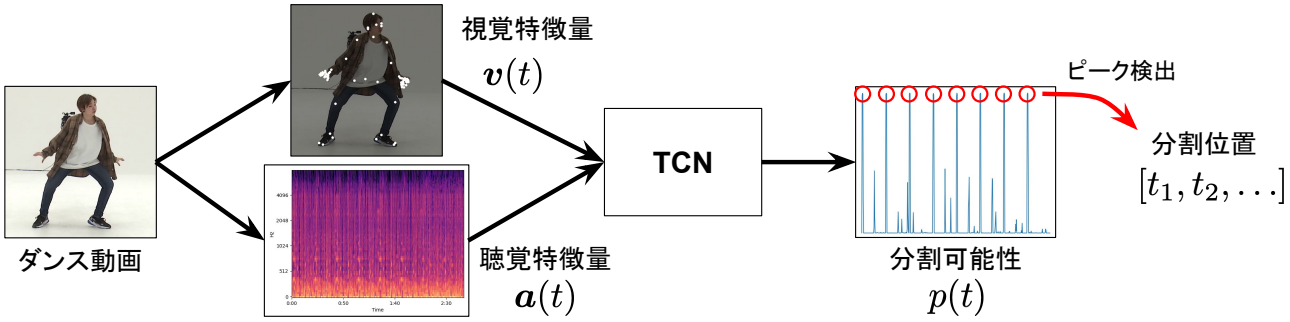

図 1. 提案手法のアルゴリズム概要．ダンス動画から抽出した視覚特徴量と聴覚特徴量を TCN の入力とし，出力として得られた分割可能性のピークを検出することで動画の分割位置を決定する．

存在するため，その場合の対策も必要である．前者の場合は AlphaPose がキーポイントの位置と合わせてそのキーポイントの検出の信頼度を出力することを利用し，各フレームに対して全身のキーポイントの信頼度の総和が最大となる人間を選択する．後者の場合は，検出されなかったフレームでのキーポイントの位置は前後のフレームでのキーポイント位置から線形補間によって求める．

3.3 聴覚特徴量の抽出

聴覚特徴量は動画に含まれる音楽データから計算される．まず，動画中の音楽を短時間フーリエ変換 (STFT) してメルスペクトログラム S を作成する．これは音の各時間での各周波数成分の強さを表す 2 次元配列であり，各要素は dB 単位で求めた後に $[-0.5, 0.5]$ の範囲に正規化している．ここで得られたメルスペクトログラムのサンプル数が動画のフレーム数よりも多いため，CNN を用いて情報を圧縮し，フレーム t に対する聴覚特徴量 $\boldsymbol{a}(t) \in \mathbb{R}^{16}$ を計算する．具体的には，t に時間的に最も近いスペクトログラムのサンプルのインデックスを i とすると，$\boldsymbol{a}(t)$ は

$$\boldsymbol{a}(t) = \mathrm{Conv2d}\left(S_{i-2,\ldots,i+2}\right) \tag{2}$$

で求めることができる．ここで Conv2d は CNN による 2 次元の畳み込みを表す．メルスペクトログラムや Conv2d の計算におけるパラメータは [3] と同様に設定した．

3.4 Temporal Convolutional Network

以上で求めた視覚特徴量 $\boldsymbol{v}(t) \in \mathbb{R}^{68\times 2}$ と聴覚特徴量 $\boldsymbol{a}(t) \in \mathbb{R}^{16}$ をまとめて 152 次元の特徴量を作成し，TCN への入力として使用する．TCN は時系列データに対して CNN を適用するモデルであり，層が深くなるにつれて隣の要素との間隔を広げて畳み込むことで過去や未来の情報を考慮した計算ができる．TCN の構造は Davies と Böck [3] を参考に作成した．[3] では TCN を用いて音楽から拍の位置を推定しており，元々の TCN のモデル [1] と比較すると，過去だけでなく未来の情報も畳み込む点や活性化関数に exponential linear unit (ELU) [2] を使用している点が異なる．TCN による計算は各次元で独立に行い，最後に全結合層で全ての次元をまとめ，活性化関数としてシグモイド関数に通すことで出力 $p(t) \in [0, 1]$ を得る．

3.5 ピーク検出

TCN から出力された $p(t)$ は，動画がフレーム t で分割される可能性であるため，最終的な分割位置を決定するために $p(t)$ からピーク値を検出する必要がある．$p(t)$ がフレーム t^* でピーク値を取ることは，$p(t^*)$ が閾値を超える局所最大値であると言えるため，ピーク検出の条件は以下のように表すことができる．

$$p(t^*) > h \ \wedge \ p(t^*) = \max_{t-w/2 \le s \le t+w/2} p(s) \tag{3}$$

ここで w は局所最大値を計算する窓サイズ，h はピーク検出の閾値である．この条件を満たす t^* を $p(t)$ から全て求め，分割位置 $[t_1, t_2, \ldots]$ とする．

4 学習データの作成

4.1 使用したダンス動画

提案手法のネットワークモデルを学習させるためにはダンス動画と分割位置をペアとする学習データが必要となる．本研究では，AIST Dance Video Database [12] のダンス動画に対して手動で分割位置を指定することで学習データを作成した．使用した動画は基礎的な動きを行う 23 秒程度の基本ダンス 1200 本と様々な振りを含む 52 秒程度のフリーダンス 210 本の合計 1410 本であり，いずれも正面のカメラから撮影されたものである．なお，分割位置の指定作業は約 10 年のダンス経験を持つ第 1 著者が手作業で行い，所要時間は約 50 時間であった．分割の際は全体的な基準として，技と呼ばれる一連の動きや切れ目なく滑らかに繋がるような動きは 1 つの区間となるように，その動きの開始，終了のタイミングで分割を行った．

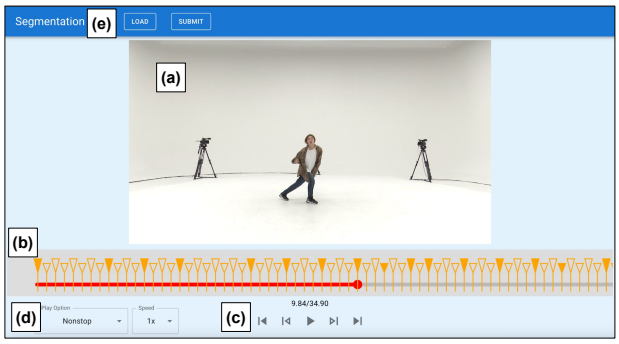

図 2. 学習データ作成ツール．(a) ダンス動画．(b) 分割候補点．(c) 再生，スキップボタン．(d) 再生モード，速度変更用プルダウン．(e) 動画読み込み，分割結果登録ボタン．

4.2 学習データ作成ツール

我々は作業効率化のために図 2 のようなツールを作成し，これを用いて動画の分割を行った．このツールでは中央にダンス動画を表示する画面 (図 2a)，その下にシークバーと分割の候補点 (図 2b) が表示され，ユーザは各候補点をクリックすることで分割位置を指定する．分割位置として指定された候補点はオレンジ色で塗りつぶされる．下部中央 (図 2c) には動画の再生ボタンがあり，その両隣には現在の再生位置の 1 つ前または後ろの分割候補点まで再生位置をスキップするボタンがある．また通常の動画再生だけでなく，指定した分割位置で動画を一時停止する機能を追加している．再生モードは左下のプルダウン (図 2d) から変更できる．動画の読み込みや分割結果の登録は左上のボタン (図 2e) から行う．

4.3 分割候補点の絞り込み

分割の候補点は少なすぎるとユーザの望み通りに分割できない一方で，多すぎると作業効率が悪くなるため，適切な数を設定することが望ましい．そこで我々は，動画中の音楽の拍の位置とその半分の位置を分割の候補点として設定した．AIST Dance Video Database で使われている音楽はテンポが既知であり，動画に含まれる音声はノイズのない純粋な音楽データであるため，最初に音が大きくなった位置を音楽の開始位置としてテンポに従って区切ることで分割候補点の位置を容易に計算することができる．ツール上では図 2b のように，拍の位置を大きい逆三角で，その半分の位置を小さい逆三角で示している．

5 評価実験

5.1 実験設定

第 4 章で作成した学習データを用いて提案手法のモデルを学習し，その性能を評価した．学習データはそれぞれに含まれる基本ダンスとフリーダンスの割合が等しくなるよう注意しながら 3:1:1 の比でランダムに分割し，それぞれ訓練用，検証用，テスト用データとした．

表 1. テストデータに対する分割位置の推定結果．

モデル	訓練終了エポック	適合率	再現率	f 値
視覚	17.4	0.383	0.407	0.395
聴覚	29.2	0.594	0.625	0.609
視覚+聴覚	18.6	**0.696**	**0.682**	**0.689**

損失 $\mathcal{L}$ は重みつきの binary cross-entropy を採用し，モデルの出力する分割可能性 $p(t)$ と正解ラベル $l(t)$ から以下の式で計算する．

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\{\alpha\, l(t)\log p(t) + (1-l(t))\log(1-p(t))\} \quad (4)$$

ここで T は動画の総フレーム数である．正解ラベル $l(t)$ は第 4 章での分割結果から計算され，指定された分割位置に最も近いフレームで 1，その両隣で 0.5，それ以外で 0 を取る．α は重みを表す定数であり，ラベルの要素がほとんど 0 であることを考慮して，学習が偏らないように設定している．本実験では $\alpha = 100.0$ とした．

以上の設定で，訓練データを用いてモデルの学習を行った．訓練時の最適化手法は Adam [5]，バッチサイズは 1，学習率は 0.001 とし，検証データによる平均損失が直近 10 エポックで改善されなければ学習をストップし，その時点でのモデルをテストデータで評価した．また比較手法のため，提案手法 (視覚特徴量と聴覚特徴量を両方使うモデル) に加えて，視覚特徴量のみ，聴覚特徴量のみを用いたモデルを用意した．但し，視覚のみ，聴覚のみのモデルで実験するときは，提案モデルの実験で用いた訓練，検証，テストデータと同じものを使用した．モデルは PyTorch [7] を用いて実装し，学習は Google Colaboratory 上で GPU を使用して実行した．

5.2 結果と考察

5.1 節の手順で実験を行い，モデルの訓練にかかったエポック数と，訓練後のモデルを用いてテストデータから検出した分割位置の適合率と再現率を記録した．この実験を独立に 10 回繰り返し，訓練エポック数と適合率，再現率の平均値を求め，平均適合率と平均再現率から f 値を求めた結果を表 1 に示す．ピーク検出時の窓サイズ w は 20 とした．また閾値 h は 0.90 から 0.99 まで 0.01 ずつ変更しながら数値評価を行ったところ，全てのモデルで $h = 0.98$ の時に f 値が最大となったため，0.98 を採用した．

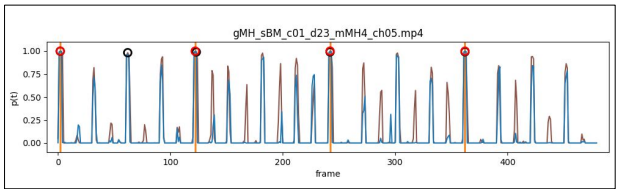

図 3. あるテストデータに対する推定結果の比較．タイトルは動画ファイル名，オレンジの縦線は正解の分割位置を示す．視覚＋聴覚のモデルで推定した $p(t)$ と分割位置はそれぞれ青線と赤丸，聴覚のみのモデルの推定結果は茶色線と黒丸で示す．

表 1 から，提案手法 (視覚特徴量と聴覚特徴量を両方使うモデル) が適合率，再現率，f 値が最大となった．一方，視覚特徴量のみのモデルでは各指標の値が大きく低下した．聴覚特徴量のみのモデルは提案モデルに近い性能を出しているものの，訓練終了までのエポック数が大きくなる傾向が見られた．この結果から，ダンス動画の分割には視覚的なダンサーの動きの情報と聴覚的な音楽情報の両方を用いるのが有益であると言える．

また，提案モデルと聴覚特徴量のみを用いたモデルとで，あるテストデータに対する推定結果を比較したものを図 3 に示す．図 3 を見ると提案モデルで推定した分割可能性は聴覚のみのモデルと比べて，正解位置以外での値が小さい．他のテストデータでも同様の傾向が見られたことから，視覚情報は聴覚情報のみでは判断できない不要な候補を削除する役割を果たしていると考えられる．

さらに提案手法において，テストデータの動画の中で f 値が最大，最小の動画の分割位置を表すグラフを図 4，5 に示す．図 4a のダンス[1]は同じ動きの繰り返しになっており，音楽の拍位置に動きのアクセントが来るような振りのため，モデルによる予測が容易であったと考えられる．その一方，図 5b のダンス[2]は，はっきりと音楽に合わせるのではなく柔らかく流れるような動きが多いため，分割位置の予測が難しかったと考えられる．

6 応用例

提案手法の応用例として，我々は自動分割の結果を利用してダンスの理解や習得を支援するシステムのプロトタイプを開発した．プロトタイプのインタフェースを図 6 に示す．読み込んだダンス動画を表示する画面 (図 6a) やシークバー (図 6b)，再生，スキップボタン (図 6c) は，学習データ作成時のツール (図 2) と同様の設計だが，シークバー上に表示

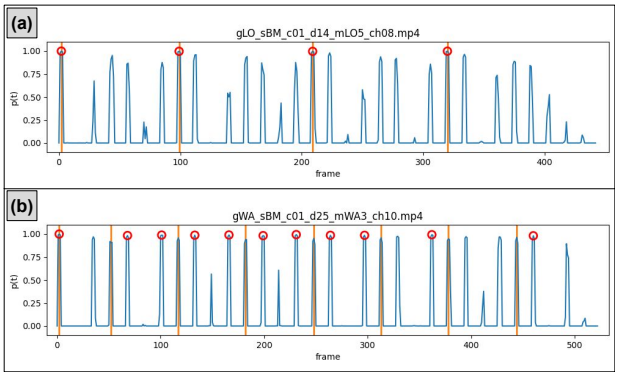

図 4. テストデータ (基本ダンス) に対する分割位置の推定結果 ((a) f 値が最大，(b) f 値が最小のもの)．グラフの見方は図 3 と同様．

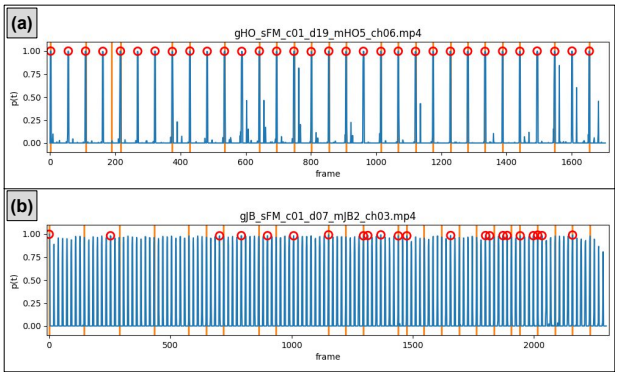

図 5. テストデータ (フリーダンス) に対する分割位置の推定結果 ((a) f 値が最大，(b) f 値が最小のもの)．グラフの見方は図 3 と同様．

される分割位置は自動推定した分割位置となっている．再生モードは 4.2 節で述べた 2 種類に加えてループ再生モード (図 6d) を追加した．ループ再生モードでは，再生開始時の再生位置の前後にある分割位置の間をループ再生することで，ユーザが同じ動きを繰り返し練習することを支援することができる．ループ再生の範囲はシークバー上に鍵括弧 (図 6e) で表示する．但し，ループ再生の範囲は，分割位置ちょうどではなく，少し前後にはみ出るように設定している．この理由として，分割したセグメント同士に少し重なりを持たせることで，ユーザがセグメントの境界での動きのつながりや予備動作を理解しやすくなることが挙げられる．現在の実装でははみ出す部分の長さは定数 (0.2 秒) としている．また，学習に最適な分割の細かさはユーザの好みや熟練度によって変わりうるため，ユーザが分割の細かさを調節できるように，右下部にスライダ (図 6f) を用意した．このスライダを操作することで，3.5 節のピーク検出時の閾値 h の値を変更し，分割位置の数を調整することができる．動画の読み込みは左上の

[1] https://aistdancedb.ongaaccel.jp/v1.0.0/video/10M/gLO_sBM_c01_d14_mLO5_ch08.mp4

[2] https://aistdancedb.ongaaccel.jp/v1.0.0/video/10M/gJB_sFM_c01_d07_mJB2_ch03.mp4

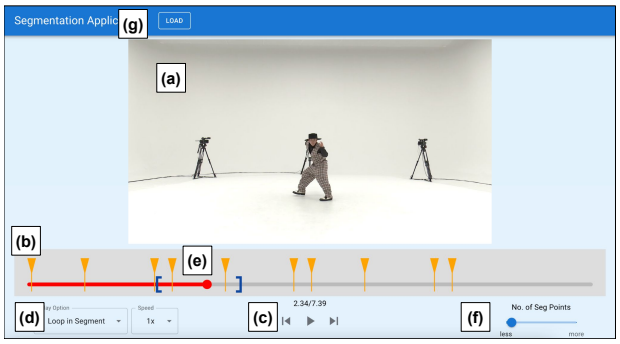

図 6. 自動分割の結果を利用したダンス学習支援システム．(a) ダンス動画．(b) シークバー．(c) 再生，スキップボタン．(d) 再生モード，速度変更用プルダウン．(e) ループ再生の範囲を表す鍵括弧．(f) 分割の細かさ調節スライダ．(g) 動画の読み込みボタン．

ボタン (図 6g) で行う．

本システムは学習者が個人で使用し，自力で振りを覚えるための分割動画を作成することを想定している．振りの自動分割機能は主に初心者のダンス習得に役立つと考えられるが，自力で振りを分割できる熟練者が動画の振りを練習する場合でも，動画を手で分割する手間が省けるという点で役に立つと考えられる．現在はプロトタイプ実装のため，事前に提案モデルで分割可能性を推定した動画しか読み込めないが，将来的にはユーザがアップロードする任意の動画に対しても自動で分割位置を推定できるように改善する予定である．

7　まとめと今後の課題

本研究では，ダンス動画の振りを個々の短い動きへと自動で分割する手法を提案した．提案手法は，動画内から推定したダンサーのキーポイントの速度情報により視覚特徴量を求め，動画の音楽データのメルスペクトログラムを CNN で畳み込むことで聴覚特徴量を取得する．これらの特徴量を TCN に入力して分割可能性を計算し，その中のピークを検出することで分割位置を計算した．我々は，学習データ作成のためのシンプルなアノテーションツールを作成し，AIST Dance Video Database の動画に対して人手で分割位置を指定することで学習データを用意した．このデータを使用した実験の結果，提案手法 (視覚と聴覚両方の特徴量を組み合わせたモデル) は視覚のみ，聴覚のみの特徴量を使用するモデルよりも適合率，再現率，f 値の項目で優位な結果が得られた．この結果から，ダンス動画の分割には視覚的なダンサーの動きの情報と聴覚的な音楽情報の両方が役に立つことが分かる．

今後の課題として，まず人による分割位置の差異への適応が挙げられる．現在の学習データは第 1 著者のみで作成しているが，振りの分割位置は絶対的なものではなく人によってある程度の差異がある．この差異に対応するためには，まず複数のダンス経験者に動画の分割を行ってもらって学習データを集め，大多数の人間の好みにある程度沿う「平均的な」正解ラベルを作成する．さらにこの平均的な正解ラベルを出発点として，[17] のように Human-in-the-Loop の手法で個々人の好みに沿った分割を生成することが考えられる．

また，現在は学習データの都合によりストリートダンスやジャズダンスを対象としているが，提案手法はダンスの種類やジャンルに特有の知識を必要としないため，民族舞踊やコンテンポラリーダンスなど他のダンスにも適用できる可能性がある．これらのダンスについて学習データを集め，提案手法の有効性を検証することも今後の発展として考えられる．性能評価については，提案モデルによる分割の結果をダンス初心者や熟練者が手動分割した結果と比較することも考えられる．更に，提案手法の応用例として開発したダンス学習支援システムのプロトタイプに関するユーザテストを実施し，自動分割やプロトタイプシステムが実際にダンスの理解や習得に役立つかどうかを検証していく予定である．

謝辞

本研究は JST CREST JPMJCR17A1 の支援を受けたものである．

参考文献

[1] S. Bai, J. Z. Kolter, and V. Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271*, 2018.

[2] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In Y. Bengio and Y. LeCun eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[3] M. E. P. Davies and S. Böck. Temporal convolutional networks for musical audio beat tracking. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.

[4] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional Multi-person Pose Estimation. In *ICCV*, 2017.

[5] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In Y. Bengio and Y. LeCun eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[6] N. Okada., N. Iwamoto., T. Fukusato., and S. Morishima. Dance Motion Segmentation Method based on Choreographic Primitives. In *Proceedings of the 10th International Conference on Computer Graphics Theory and Applications - GRAPP, (VISIGRAPP 2015)*, pp. 332–339. INSTICC, SciTePress, 2015.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett eds., *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 2019.

[8] F. Pedersoli and M. Goto. Dance beat tracking from visual information alone. In *Proceedings of the 21st International Society for Music Information Retrieval Conference, (ISMIR 2020)*, pp. 400–408, Montreal, Canada, Oct. 2020.

[9] J.-P. Rivière, S. F. Alaoui, B. Caramiaux, and W. E. Mackay. Capturing Movement Decomposition to Support Learning and Teaching in Contemporary Dance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019.

[10] T. Shiratori, A. Nakazawa, and K. Ikeuchi. Detecting dance motion structure through music analysis. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pp. 857–862, 2004.

[11] M. Tachikawa. SymPlayer -動画ミラー反転でプロの動きをマスター- (最終閲覧日：2022 年 9 月 24 日). `https://apps.apple.com/jp/app/id1048785434`.

[12] S. Tsuchida, S. Fukayama, M. Hamasaki, and M. Goto. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, (ISMIR 2019)*, pp. 501–510, Delft, Netherlands, Nov. 2019.

[13] S. Tsuchida, H. Mao, H. Okamoto, Y. Suzuki, R. Kanada, T. Hori, T. Terada, and M. Tsukamoto. Dance Practice System That Shows What You Would Look Like If You Could Master the Dance. In *Proceedings of the 8th International Conference on Movement and Computing*, MOCO '22, New York, NY, USA, 2022. Association for Computing Machinery.

[14] Ugotoru Inc. ウゴトル (最終閲覧日：2022 年 9 月 24 日). `https://ugotoru.com/ugotoru`.

[15] Z. Zhou, A. Xu, and K. Yatani. SyncUp: Vision-Based Practice Support for Synchronized Dancing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3), Sept. 2021.

[16] 斎藤 光, 徳久 弘樹, 中村 聡史, 小松 考徳. ダンス動画へのオノマトペ付与によるダンス習得促進手法. 情報処理学会 研究会報告グループウェアとネットワークサービス (GN), 2020-GN-109(33):1–8, Jan. 2020.

[17] 山本 和彦. Human-in-the-Loop 型適応によるインタラクティブな音楽的拍節解析. 第 29 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS2021), pp. 23–29, Dec. 2021.