Vol.2020-HCI-190 No.7
Vol.2020-UBI-68 No.7
2020/12/8

人体ポーズ分析を応用したシンクロダンス練習支援システム

周　中一[1,a)]　矢谷　浩司[1,b)]

概要：シンクロダンスの美しさは，複数のダンサーの間で体のポーズが同期していることにある．ダンサーは練習にカメラ録画を利用しているが，標準的な動画インタフェースでは，ポーズの同期が十分でない部分を特定し，さらなる練習を行うことを効率的に支援できていない．今回提案する SyncUp は，一般的なカメラで撮影された動画を用いるシンクロダンス練習支援システムである．SyncUp は，ユーザがアップロードした動画を解析することで，動画中の複数ダンサーのポーズの類似性を定量化する．そしてより良い同期性を実現するために，体のどの部分の練習する必要があるのかを可視化する．本稿では，SyncUp の実装手法，およびシンクロダンスに特化したポーズ類似性推定手法の詳細と評価について報告する．

A Synchronized Dance Practice Support System Using Human Pose Analysis

Zhongyi Zhou[1,a)]　Koji Yatani[1,b)]

1. 背景

シンクロダンスは，複数のダンサーによる同期した運動の連鎖で構成されたダンスであり，動きの同期性が視覚的な美を創造する．プロのダンサー以外にも，人気アイドルやアニメのキャラクターの振り付けを再現して，SNS（ソーシャルネットワークサービス）で動画を公開するアマチュアのダンサーも多くいる．シンクロダンスへの関心が高まっているにも関わらず，その練習を支援するためのインタラクティブな技術はほとんど研究されていない．例えば既存の技術では，主に一人での利用に焦点が当てられており，シンクロダンスを明確な対象としていない．また，これらのシステムでは，正解例のポーズとの部位間のユークリッド距離の単純和をダンスの良さの指標として用いることが多い．しかし，ダンスにおいては一部の体の部位がより視覚的注意を集めやすい場合があり（例えば腕を回しているときは，腕に注意が行きやすい），既存の手法ではこの点を明示的に考慮していない．

また，アマチュアのシンクロダンスでは，SNS 上での公開や練習の振り返りのために，ビデオ撮影を行うのが一般的である．既存のダンス練習支援システムでは深度が測定できるデプスカメラを利用したものがあるが，一般的なカメラでシンクロダンスの練習支援が可能となれば，その利用範囲も広がることが期待できる．

本研究では，一般的なカメラによって撮影されたシンクロダンスの動画を用いてシンクロダンスの練習を支援するシステム SyncUp を提案する（図 1）．また．これらのシステムでは，与えられた動画から複数ダンサーのポーズを推定し，体の部位間でのユークリッド距離を計算した上で，人間が感じる同期度の高さを機械学習によって推定することにより，どの場面においてダンスのズレが生じているかをユーザに提示する．この同期の度合いはインタフェース上において，折れ線グラフと 1 次元ヒートマップで可視化される．さらに SyncUp は動画中のダンサーの上にヒートマップを重畳し，ユーザに対してどのダンサーのどの体の部位がズレを生じさせているかを視覚的に提示する．これらのインタフェースの機能により，ユーザはより練習を深める必要がある部分を効率よく同定し，練習の指針を効率的に得られることが期待できる．

[1] 東京大学 IIS Lab
IIS Lab, University of Tokyo
[a)] zhongyi@iis-lab.org
[b)] koji@iis-lab.org

ⓒ 2020 Information Processing Society of Japan

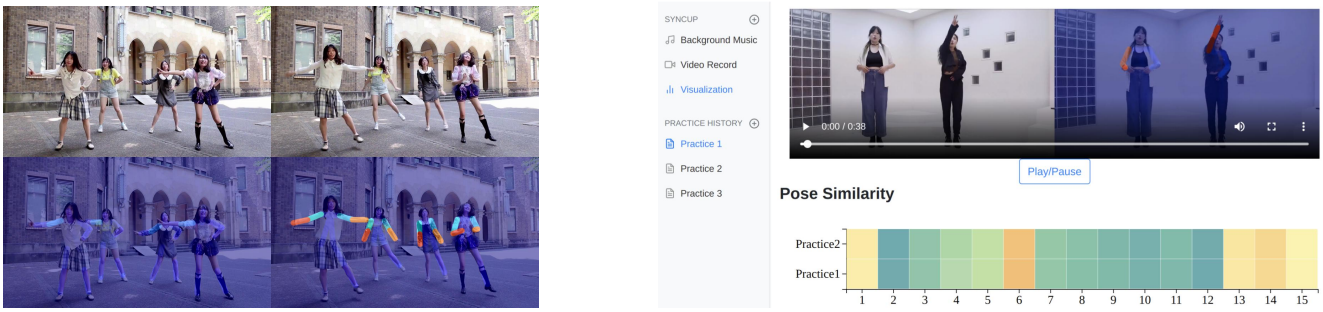


図 1: SyncUp システムの概要．左：SyncUp のヒートマップがダンサーに重畳され、どの体の部位がポーズの同期度を下げているかを強調する．色が各体の部位のズレの度合いの大小を表す（赤色ほどズレの度合いが大きい）．右：SyncUp の Web インタフェース．ダンサーが練習の動画をアップロードすると、システムがダンスの同期度を数値化し，フィードバックを提供する．

2. 関連研究

2.1 ダンス練習支援システム

本研究はシンクロダンスの練習支援を目的としている．既存のインタラクティブなシステムや可視化は，ダンスにまつわるさまざまなアプリケーションを探求してきた（例えば，人々が一緒に踊ることを促進するもの [13] など）．先行研究ではダンス学習の複雑さが明らかにされており [3,14]，ダンサーのための練習システムの研究が進められてきた．そのようなインタラクティブなシステムの多くは，知覚の強化を通してダンサーの練習を支援している [5,18]．Drobny ら [8] によると，ダンス学習者にとっての共通の課題は，音楽のリズムに自分の動きを同期させることであることを発見した．この問題に対処するために，彼らは Saltate を作成した．Saltate は，ビートのタイミングでドラム音のフィードバックを初心者に提供する．また YouMove [1] は，ダンスの重要な場面において生徒と教師のポーズ比較を可視化し，ダンス学習を支援する拡張現実鏡である．このシステムでは，ダンスの定量的な評価を行い，次の練習でどのように改善すればよいのかを生徒が素早く理解できるようにしている．

上記のシステムは主に一人での利用を対象としたものであり，シンクロダンスへの展開は十分に検討されていない．また，シンクロダンスの練習を支援するために，スマートフォンやノートパソコンなどの汎用的なデバイスとコンピュータビジョンを用いたシステムの実現可能性を明らかにすることで，ユーザはデプスカメラや拡張現実鏡など特別な機器を必要とせず，より幅広い場面でシステムを使うことができることが期待される．

2.2 コンピュータビジョンによる人の動作解析

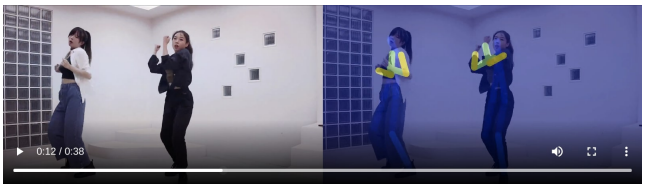
コンピュータビジョンの分野では，人の正確な姿勢検出に関する研究が活発に行われており，デプスカメラを用いたリアルタイムで頑健なポーズ検出手法が開発されている [17, 24, 25]．最近のコンピュータビジョンの進化の一例 [16] として，通常の RGB カメラで撮影された画像から人間の骨格を検出する頑健なアルゴリズムが確立された [4, 11, 27, 29]．これらの手法は，大規模なラベル付きデータセット [2, 19] を用いた畳み込みニューラルネットワーク (CNN) によるものである．OpenPose [4] は，CNN を用いた，最も初期のリアルタイム多人数ポーズ推定器の 1 つで，与えられた画像において，人物の 18 個の身体部位（キーポイント）のピクセル位置を予測する．これによって，開発者は OpenPose を使って簡単に人物のスケルトンを作成することができる．SyncUp では，OpenPose よりも優れた性能を持つ AlphaPose [11] を採用している．

2.3 コンピュータビジョンによるポーズ類似度推定

ポーズ類似度推定は，行動認識 [9,20]，運動技能学習 [1,6]，動き検索 [22] などの様々なアプリケーションで重要な役割を果たしている．基本的な方法の 1 つとして，2 つのポーズの体の部位間のユークリッド距離を利用するものがあげられる．例えば，Chan ら [6] は，仮想現実を利用したダンス練習システムを作成した．彼らは，ダンサーと正解例との間の各身体部位のユークリッド距離を用いた単純な閾値ベースの手法を採用した．Chen ら [7] は，このような身体部位のユークリッド距離を特徴量とする手法の欠点を指摘し，人間の動きを記述するための高次元のポーズ特徴量（1683 次元）と，マハラノビス距離に基づく類似度を学習するアルゴリズムを導入し，このような高次元特徴量を用いた場合により頑健であることを示した．最近では，人手で作成した特徴量の代わりに，グラフィック CNN を用いて，ポーズの類似性の概念を符号化することも検討されている [23, 28]．

しかし，これらの方法は大規模なデータセットを必要とし，シンクロダンスへの直接的な応用はまだ現実的ではない．ポーズ類似度推定はダンス練習支援システムの重要な要素であるが，先行研究の多くは未だに身体部位のユークリッド距離の単純和に留まっている．しかし，この手法はスケーリングの問題（身長の異なるダンサー間での計算など）があり，アプリケーションの実用性が低下する恐れが

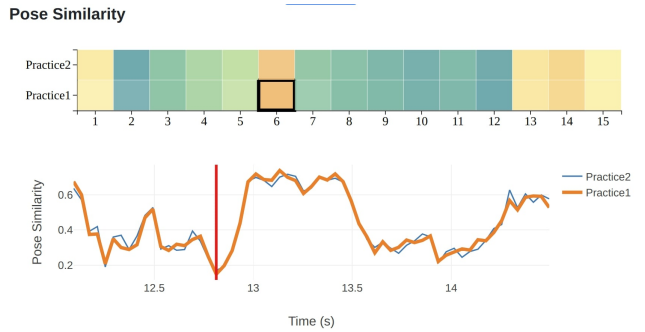
ⓒ 2020 Information Processing Society of Japan



(a) グループ練習時における動画インタフェース.



(b) 個人練習時における動画インタフェース.



(c) ダンサー間のポーズ類似度の可視化．値が高いほどポーズの同期度が良いことを表す.

図 2: SyncUp 上で提供される可視化.

ある．正確で頑健なポーズ類似度推定を構築するためには，さらなる検討が必要である．

3. SyncUp

SyncUp は，スマートフォンやタブレット等で動作するウェブベースのインタフェースとして設計されている．すべての機能が 1 つのウェブページに含まれており，ダンサーはシステムからのフィードバックを素早く確認することができる．後述するフィードバックを受け取るためには，ダンサーは自分の練習をカメラで録画し，システムにアップロードする．過去の研究では，ダンサーはリアルタイムよりも事後のフィードバックを通常好むことがわかっている [26]．また，このシステムではポーズ検出のために deep CNN を用いており，リアルタイムでの利用は不可能ではないが，頑健性が低下する可能性があるため，ダンスが終了したあとでフィードバックを見ることを想定したインタフェースとなっている．

3.1 グループでの練習

3.1.1 動画上でのフィードバック

SyncUp では，オリジナルの練習映像とダンサーの上にポーズのズレの大きさを重畳したものを横に並べて見ることができる（図 2a）．図 2a の例では，重畳されたヒートマップがダンサーの腕を黄色で強調しており，腕がダンサー間でより大きくずれていることを可視化している．このフィードバックを通じて，ユーザは次の練習で身体のどの部分を改善する必要があるかを素早く認識できる．

3.1.2 ポーズ類似度のグラフ

ポーズ類似度とは，与えられた動画フレームで複数のダンサーのポーズがどの程度同期しているかを定量化したものである．ダンサーの上に重畳されるヒートマップに加えて，SyncUp では 1 次元のヒートマップも提供している．このヒートマップでは，各場面の平均的なポーズ類似度を視覚化しており，暖色系の色は同期度が低いことを示している．

ユーザが 1 次元ヒートマップのブロックをクリックすると，ポーズ類似度スコアの折れ線グラフが表示され，類似度の変化を詳細に調べることができる．図 2c は，ポーズ類似度スコア結果の一例を示している．スコアの値が高いほど，ポーズの同期度が高いことを示している．ユーザはこの折れ線グラフを確認することで，同期度の低い瞬間を確認することができる．また，折れ線グラフ上でクリックすることにより，対応するフレームを動画インタフェース上で確認できる．さらに以前の練習におけるパフォーマンスも可視化され，どのように改善されたかを見直すことができる（図 2c）．

3.2 個人での練習

SyncUp は個人での練習にも対応している．インタフェースはグループでの練習の場合と同じであるが，正解とするダンサー（リーダーと呼ぶ）のダンスを参照してポーズの類似性を計算するように変更されている．練習を行うダンサー（フォロワー）が自身の練習動画をアップロードすると，システムは BGM の情報を用いてリーダーとの動画を同期し [10]，ポーズ類似度を計算する．動画インタフェースでは，左側にはフォロワーのダンスが，右側にはリーダーのダンスとヒートマップの重畳が表示される（図 2b）．

4. ポーズ類似度推定手法

SyncUp のポーズ類似性分析では，特定のフレーム内で複数のダンサーのポーズがどのように同期しているかを調べる．そして最終的には，図 2 に示すインタフェースを通じてユーザにその結果を提示している．

4.1 動画の前処理

SyncUp では，ダンサーのポーズを識別するために，複数人のポーズ推定器である AlphaPose [11] を使用している．AlpahPose では，人物のポーズは OpenPose と同様に，18 個のキーポイント (p) [4, 19] のピクセル位置のリストで表される．これらのキーポイントは，目 2 つ，耳 2 つ，鼻，首，肩 2 つ，肘 2 つ，手首 2 つ，腰 2 つ，膝 2 つ，足首 2 つの 2 次元上のピクセル位置である．図 3a は，AlphaPose



(a) 元のビデオフレームに AlphaPose [11] を用いて推定されたスケルトンを重ねたもの（この 4 つのフレームはランダムに抽出されたもの）.



(b) ヒートマップの重畳を用いた体の部位部レベルのポーズ類似性の可視化.

図 3: ポーズ類似度定量化手法の結果例.

キーポイントから作成されたスケルトンデータを録画されたダンス動画に重ねて表示した例である．ポーズ検出の精度については，本研究の範囲外であるため，正式な実験は行っていない．しかし，AlphaPose では，人が大きく隠れている場合を除いて，ダンサーのポーズの検出に成功していることが, 我々の予備的な調査で確認された.

4.2 ポーズ類似度推定

SyncUp では，まず Body-part Pose Similarity（BPS）と呼ばれる，ダンサー間の体の部位レベルでのポーズの違いの度合いがアルゴリズムによって算出される．動画上に重畳されるヒートマップではこの BPS が使用されている（図 3b ）．さらにシステムは，全てのダンサーの BPS の計算がなされた後，それらを統合して，折れ線グラフで使用される Overall Pose Similarity（OPS）を推定する.

4.2.1 体の部位レベルのポーズ類似度の定量化

SyncUp は，14 のキーポイント（目と耳を除く）を使用して，13 のベクトルとしてダンサーのポーズを表現する．目と耳の 4 つの特徴を除いたのは，顔の方向ではなく体のポーズに注目するためである．これらの特徴ベクトルを対応する単位ベクトルに正規化し，各身体部位の方向情報のみを保存する．これにより，人の高さやカメラからの距離が異なることによるスケールの問題を回避することができる．これら 13 個の単位ベクトルをアルゴリズムの入力として使用し，与えられたフレーム t での各身体部位 i に対する Body-part Pose Similarity score ($BPS(i,t)$) を算出する.

$\vec{v}_j(i,t)$ を j 人目 ($j \in \{1,2...,J\}$) のダンサーの i 番目の単位ベクトル ($i \in \{1,2...,13\}$) とする．SyncUp では，全ダンサーにわたる各身体部位の累積絶対差（$d(i,t)$）を以下の式で算出する.

$$d(i,t) = \sum_{j}^{J} |\vec{v}_j(i,t) - \vec{v}_R(i,t)|$$

$\vec{v}_R(i,t)$ はリーダーの i 番目の単位ベクトルであり，グループでの練習のモードにおいては，全ダンサーの平均値に設定される．ただし，この $d(i,t)$ は制限を持たない値となり，ダンサーの数によって大きく変動する．このため，以下の式で正規化を行い，BPS を算出する.

$$BPS(i,t) = \left(\frac{d(i,t)}{J}\right)^{\lambda} = \left(\frac{\sum_J |\vec{v}_j(i,t) - \vec{v}_R(i,t)|}{J}\right)^{\lambda}$$

$BPS(i,t)$ が大きければ，i 番目の部位のポーズの差異が大きいことを意味する．本システムでは，この $BPS(i,t)$ の値を線形に COLORMAP_JET クラス[*1]の色スペクトルに割り当てて，重畳されるヒートマップを生成している．また，λ は重畳されるヒートマップの感度を調整する変数であり，この値が大きい場合はポーズの大きな違いのみを強調し，一方でこの値が小さい場合はポーズの細かな違いも提示することになる.

4.2.2 OBS の導出

全身体部位における全ての $BPS(i,t)$ が求められた後，練習の各瞬間におけるダンスの類似性をまとめた OPS ($OPS(t)$) を計算する．我々の研究では，サポートベクターマシン（SVM）が最も頑健であることがわかったため，SVM を使用することにした（5，6 節を参照のこと）.

$BPS(i,t)$ を単純に加算する場合，知覚されるポーズのズレの大きさに対する重みは全ての体の部位で同一となる．一方，SVM の回帰モデルを用いて $OPS(t)$ を導出することで，ユーザの実際の知覚により沿った計算を行うことができる．この方法では学習データが必要となるが，ダンサーの練習映像において人間が知覚するポーズの類似度に関するデータは存在しないため，本研究ではその情報を収集した上で学習を行った（後述）.

[*1] https://docs.opencv.org/2.4/modules/contrib/doc/facerec/colormaps.html

ⓒ 2020 Information Processing Society of Japan



(a) $\lambda = 0.5$ (b) $\lambda = 1$ (c) $\lambda = 2$

図 4: 感度を制御する変数 λ の 3 つの値における可視化結果の違い．λ が大きくなると，体の部位の小さな違いの可視化が抑制されていることがわかる．

5. ポーズ類似度推定に関する比較評価

5.1 ダンスビデオのデータ収集

SyncUp には個人での練習とグループでの練習と 2 つの練習モードがあるため，評価のために 2 種類のダンスビデオを収集した．

5.1.1 個人練習動画ペア（Data-IPV）

ダンサーが個別に練習している動画を 9 本収集した。その中から同じ部分を抽出し，リーダーのダンスを基準としたペアと，フォロワーのダンスを比較対象としたペアを作成した．さらに，評価のために完全に同期したデータと完全に同期していないデータとなっている人工的なペアも作成した．以上により，完全な同期のペア 1 つ，意図的にずれたペア 2 つの動画を含む，計 7 つの動画ペアが得られた．

5.1.2 グループ練習動画（Data-GPV）

複数のダンサーが同じ動画の中で踊っている動画に関しては，4 つのダンスグループ（各グループ 2〜4 人）の動画を入手し，合計 11 本のグループ練習動画を評価に利用した．

5.2 人間の評価者によるポーズ類似度の評価付け

我々のポーズ類似度推定手法はダンス映像の各フレームを利用している．そこで，フレームの集合を作成し，評価者に映像中の 2 人のダンサーのポーズがどれくらい一致しているかを評価してもらった．Data-IPV と Data-GPV からそれぞれ 100 フレーム，合計 200 フレームを抽出した．次に，オンラインフォームを通じて，ボランティアの方に 2 人のダンサーのポーズの類似性を 5 段階（Excellent, Good, Fair, Poor, Very Bad）で回答してもらい，それぞれ 1, 0.75, 0.5, 0.25, 0.0 の数値に割り当てた．値が大きいほど，2 人のダンサーのポーズの類似度が高いことを示している．この評価付けには合計 161 人のボランティア（男性 98 人，女性 60 人，3 名は回答せず）を募集した．各ボランティアには，無作為に選択された 25 フレームが割り当てられた．結果，各フレームについて平均して 20 名からの評価を得た．収集したデータにおいて，各サンプルの平均評価から標準偏差の 3 倍を超えたものを外れ値として削除し，今回のシステム評価においては，残りのデータの平均を用いた．

5.3 比較するポーズ類似度推定手法

この実験では，人間の評価の平均値を基準とすることで，*OBS* の予測精度と頑健性の両方を評価することである．SVM を用いた手法の他，比較のために以下の手法を実験することとした．

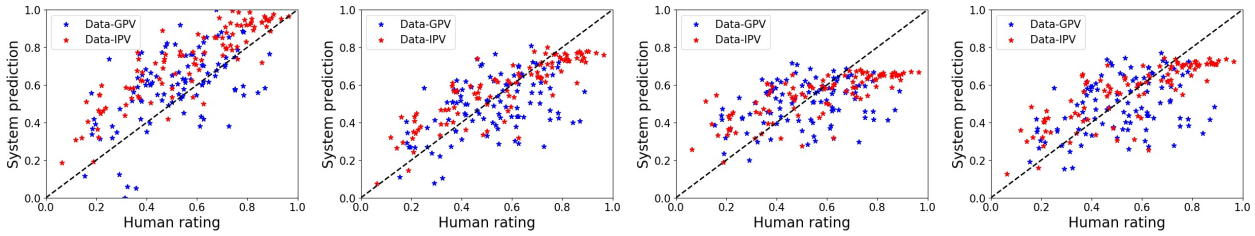
- **単純和**: *OBS* を計算する最も簡単な方法は，各身体部位におけるすべての *BPS* を合計することであり，既存の研究でもよく用いられている [7,30]．しかし，この手法の根底には，すべての特徴が最終的な総合スコアのために等しく重み付けされるべきである，という仮定があるが，これはポーズの類似性に関する人間の知覚に必ずしも適合していない [7,21]．しかし，機械学習法を用いた我々のアプローチと比較するために，この方法を実験に組み入れることとした．
- **ニューラルネットワーク（NN）**：ニューラルネットワークは物体検出や顔認識のような複雑な知覚タスクに適していることが多い [12,16]．本研究では，このようなニューラルネットワークの性能比較を行うために，short-NN と long-NN の 2 種類のニューラルネットワークを用いた．short-NN は，入力層（13 次元の *BPS*）と出力層（1 次元の *OBS*）の 2 つの層から構成される．また Long-NN は入力層，10 次元と 5 次元の 2 つの隠れ層，そして出力層の 4 つの層で構成される．より複雑なネットワーク構造も考うるが，本研究では学習するサンプル数が限られているため，比較的単純なネットワークを用いることとした．各線形変換後の活性化関数として，*ReLU* を用いた．学習には RMSE（Root Mean Squared Error）損失関数と Adam [15] を用い，学習率 0.01 とした．全体の訓練には 50 エポックを要した．

6. 実験結果

λ 値の異なる 4 つの計算手法（単純加算, SVM, Short-NN, Long-NN）を用いて予測を行い，人間の評価の平均値と

表 1: OPS 予測精度の比較

Method	λ									
	0.333	**0.426**	**0.543**	**0.693**	**0.885**	**1.13**	**1.44**	**1.84**	**2.35**	**3**
単純和	0.167	0.169	0.173	0.180	0.190	0.204	0.239	0.281	0.318	0.349
SVM	0.161	0.158	0.154	0.152	**0.151**	0.152	0.157	0.160	0.171	0.181
short-NN	0.225	0.226	0.187	0.174	0.170	0.167	0.190	0.178	0.184	0.189
long-NN	0.170	0.168	0.167	0.164	0.165	0.167	0.167	0.184	0.188	0.207



(a) 単純な足し算　**(b)** SVM　**(c)** short-NN　**(d)** long-NN

図 5: 人間の評価と 4 つの OPS 計算方法による予測値とのプロット．$\lambda = 0.885$ とした．

の比較を行った．SVM と NN を用いた手法の評価には，10-fold cross validation を採用した．また，λ の値の候補として，対数スケールで 0.333 から 3 の間に一様に分布する 10 の値を選択した．これ以外の範囲では実用的ではない (感度が強すぎたり弱すぎたりする）ため，除外することとした．

表 1 は，10 の λ 値での人間の評価平均と 4 つの方法による予測値の間の二乗平均誤差 (RMSE) を示している．すべての条件下で，SVM ベースの手法が最も小さい RMSE を示していることがわかった．図 ，λ=0.885 で 200 個のサンプルすべての詳細な結果をプロットにしたものである．このプロットは，人間による評価（x 軸）と SyncUp の予測（y 軸）を比較している．赤と青のサンプルは，それぞれ Data-IPV と Data-GPV の結果を表している．Data-IPV の結果の方が黒い点線により近い傾向があることがわかった．

7. 考察

ポーズ類似度手法の比較評価結果は，SVM を用いた手法による予測の RMSE（0.151）が，人間の評価によるもの（0.223）よりも小さいことを示している．この結果から，我々の手法は，人間の評価者よりも，人間の評価の平均により近い値を推定できることがわかった．しかし，それは必ずしも我々の方法が人間の評価者よりも正確であることを意味するものではない．人間の評価者でさえ，ある程度の不一致を示すこともあるように，ポーズの類似性を定量化することは，本質的に難しいタスクである．あるケースでは，人間の評価の標準偏差が 0.378 と大きくなっていた．このような場合においては，人間の評価者はダンスの

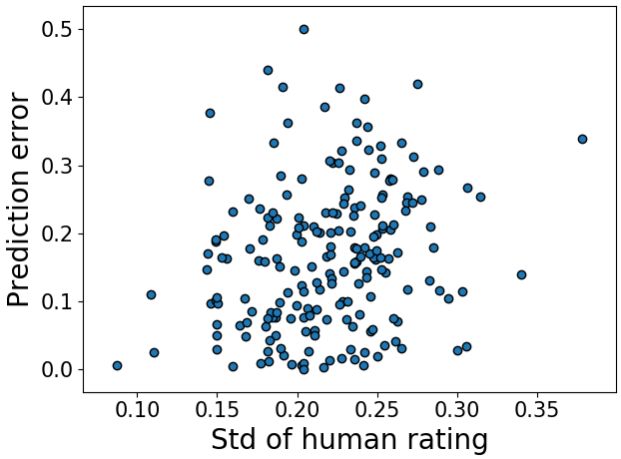


図 6: 人間の評価値の分散とシステムの予測の誤差のプロット．ピアソンの相関係数は 0.220 ($p < .005$) であった．

異なる部分を見て評価を下したと考えられる．今後の研究では，何がこのような大きな不一致の原因となったのか，また，アルゴリズムがどのようにしてこのような不一致を考慮に入れることができるのかを調査する必要がある．

よい定量化手法は，人間の評価者と同等の評価精度を持つことに加えて，人間の評価の分散と相関のある予測をすることが望まれる．例えば，人間の評価が乖離している場合は，予測の精度が低くなることが予想される．この点において我々のポーズ類似度推定がどのような結果となっているかを見るために，最良の結果の条件（SVM，λ=0.885）を用い，200 個のサンプルすべてについて，予測誤差と人間の評価の標準偏差との相関を可視化した（図 6 ）．図 6 のデータでは，ピアソンの相関係数は 0.220(p<.005) となっている．このことから，評価者の一致度と予測精度との間に明確な関連性を確認することはできなかった．この結果

は，人間の評価者によるこのような不一致をどのように予測アルゴリズムが考慮可能かを今後研究すべきであることを示唆している．

Data-GPV のデータにおいて精度が低下する原因として，ダンサーの体の一部が隠れてしまう（オクルージョン）ことが挙げられる．他のコンピュータビジョンの手法と同様に，オクルージョンはポーズ検出の信頼性を低下させる．Data-GPV に収録されている全てのクリップを精査したところ，14 個のサンプルの中にダンサーが他のダンサーの後ろに隠れている場面があることがわかった．この 14 サンプルを除外したところ，Data-GPV の RMSE は 0.174 から 0.146 に改善された．

8. 予備的なユーザ実験評価

さらに，実際の練習に SyncUp がどのように役立つかを理解するために，予備的なインタビュー調査を行った．インタビューの最初に，事前に録音されたダンスビデオを使って SyncUp のデモを共有し，システムの機能を説明した．その後，実験参加者には，現在の練習における負担を軽減を SyncUp がどのように軽減しうるか，実際の練習において SyncUp をどのように利用するのか，その理由などについて意見を求めた．この予備的なユーザ実験では，2 人のダンサーを募集した．インタビューは外国語で行われ，分析のために日本語に翻訳された．

実験参加者は SyncUp の機能に対して肯定的な態度を示していた．特に同期度が悪くなった瞬間を視覚化できていることがグループ練習に役立つ，ということに同意していた．彼らは SyncUp を使えば，ダンス動画を手動でチェックする手間から解放されると明言した．

> 「同期が取れておらず，ダンスを細かく修正する必要がある場合にシステムを使用したいと思う．コンピュータは人間よりも細部まで調べることができるので，人間が気が付かないような問題点を見つけることができると思う．」
>
> 「長いダンスを何度も何度も練習している時に，このシステムがあれば，多くの手間が省けると思います．例えば，3 分間の動画を 10 本チェックするのは，手間がかかります．このシステムがあれば，単に動画送って，濃い色を示している部分をより細かくチェックするだけでよくなるように思います．」

また，折れ線グラフでは過去の練習のパフォーマンスと比較することができ，改善点を実感できる点が良い，という意見も得られた．

9. 結論と今後の課題

シンクロダンスはアマチュアダンサーの関心を集めているが，その練習のためのインタラクティブな支援はまだ不足している．SyncUp は，シンクロダンスのパフォーマンスを定量的に評価し，ダンサーに対してフィードバックすることを可能にしたシステムである．本研究のシステム評価では，SyncUp によるパフォーマンス予測は人間の評価と高い相関性を持っていることが確認された．

本稿におけるシステム評価では，特定のタイプのシンクロダンス（アジアのポップスターやアニメのキャラクターが踊るダンス）のみを対象としていた．我々のポーズ類似度推定手法は，特定のダンススタイルを前提としていないが，今後の研究では，異なるタイプのシンクロダンスにどのように適用できるかを検討する必要がある．また，新型コロナウィルスの流行により，簡易的なユーザ実験のみしか実施できなかったが，将来的には，SyncUp のユーザ体験を理解するためにより詳細なユーザ実験を行う予定である．

謝辞

本研究の一部は，国立情報学研究所ロバストインテリジェンス・ソーシャルテクノロジー研究センターとの共同研究により支援されました．また本研究にコメントをいただいた Carla F. Griggio，Arissa Janejera Sato（佐藤 安理紗 ジエンジエラ），Zefan Sramek，本研究のデモビデオ作成を手伝っていただいた Minghui Chen（陳 明輝）に感謝申し上げます．

参考文献

[1] Anderson, F., Grossman, T., Matejka, J. and Fitzmaurice, G.: YouMove: Enhancing Movement Training with an Augmented Reality Mirror, *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, New York, NY, USA, ACM, pp. 311–320 (online), DOI: 10.1145/2501988.2502045 (2013).

[2] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[3] Bläsing, B., Calvo-Merino, B., Cross, E. S., Jola, C., Honisch, J. and Stevens, C. J.: Neurocognitive control in dance perception and performance, *Acta Psychologica*, Vol. 139, No. 2, pp. 300 – 308 (online), DOI: https://doi.org/10.1016/j.actpsy.2011.12.005 (2012).

[4] Cao, Z., Hidalgo, G., Simon, T., Wei, S. and Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *CoRR*, Vol. abs/1812.08008 (online), available from ⟨http://arxiv.org/abs/1812.08008⟩ (2018).

[5] Chan, J. C. P., Leung, H., Tang, J. K. T. and Komura, T.: A Virtual Reality Dance Training System Using Motion Capture Technology, *IEEE Transactions on Learning Technologies*, Vol. 4, No. 2, pp. 187–195 (online), DOI: 10.1109/TLT.2010.27 (2011).

[6] Chan, J. C. P., Leung, H., Tang, J. K. T. and Komura,

T.: A Virtual Reality Dance Training System Using Motion Capture Technology, *IEEE Transactions on Learning Technologies*, Vol. 4, No. 2, pp. 187–195 (2011).

[7] Chen, C., Zhuang, Y., Nie, F., Yang, Y., Wu, F. and Xiao, J.: Learning a 3D human pose distance metric from geometric pose descriptor, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 11, pp. 1676–1689 (2010).

[8] Drobny, D., Weiss, M. and Borchers, J.: Saltate!: A Sensor-based System to Support Dance Beginners, *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, New York, NY, USA, ACM, pp. 3943–3948 (online), DOI: 10.1145/1520340.1520598 (2009).

[9] Du, Y., Wang, W. and Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110–1118 (2015).

[10] Ellis, D.: The 2014 labrosa audio fingerprint system, *ISMIR* (2014).

[11] Fang, H.-S., Xie, S., Tai, Y.-W. and Lu, C.: RMPE: Regional Multi-Person Pose Estimation, *The IEEE International Conference on Computer Vision (ICCV)* (2017).

[12] Girshick, R., Donahue, J., Darrell, T. and Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[13] Griggio, C. F., Romero, M. and Leiva, G.: Towards an Interactive Dance Visualization for Inspiring Coordination Between Dancers, *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, New York, NY, USA, ACM, pp. 1513–1518 (online), DOI: 10.1145/2702613.2732925 (2015).

[14] Karpati, F. J., Giacosa, C., Foster, N. E., Penhune, V. B. and Hyde, K. L.: Dance and the brain: a review, *Annals of the New York Academy of Sciences*, Vol. 1337, No. 1, pp. 140–146 (2015).

[15] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[16] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25* (Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 1097–1105 (2012).

[17] Krull, A., Brachmann, E., Michel, F., Ying Yang, M., Gumhold, S. and Rother, C.: Learning Analysis-by-Synthesis for 6D Pose Estimation in RGB-D Images, *The IEEE International Conference on Computer Vision (ICCV)* (2015).

[18] Kyan, M., Sun, G., Li, H., Zhong, L., Muneesawang, P., Dong, N., Elder, B. and Guan, L.: An Approach to Ballet Dance Training through MS Kinect and Visualization in a CAVE Virtual Reality Environment, *ACM Trans. Intell. Syst. Technol.*, Vol. 6, No. 2 (online), DOI: 10.1145/2735951 (2015).

[19] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *CoRR*, Vol. abs/1405.0312 (online), available from ⟨http://arxiv.org/abs/1405.0312⟩ (2014).

[20] Liu, J., Shahroudy, A., Xu, D. and Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition, *European conference on computer vision*, Springer, pp. 816–833 (2016).

[21] Müller, M., Röder, T. and Clausen, M.: Efficient Content-Based Retrieval of Motion Capture Data, *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, New York, NY, USA, Association for Computing Machinery, p. 677–685 (online), DOI: 10.1145/1186822.1073247 (2005).

[22] Sedmidubsky, J., Valcik, J. and Zezula, P.: A Key-Pose Similarity Algorithm for Motion Data Retrieval, *Advanced Concepts for Intelligent Vision Systems* (Blanc-Talon, J., Kasinski, A., Philips, W., Popescu, D. and Scheunders, P., eds.), Cham, Springer International Publishing, pp. 669–681 (2013).

[23] Shi, L., Zhang, Y., Cheng, J. and Lu, H.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[24] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A.: Real-time human pose recognition in parts from single depth images, *CVPR 2011*, pp. 1297–1304 (2011).

[25] Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A. and Blake, A.: Efficient Human Pose Estimation from Single Depth Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, pp. 2821–2840 (2013).

[26] Trajkova, M. and Cafaro, F.: Takes Tutu to Ballet: Designing Visual and Verbal Feedback for Augmented Mirrors, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 2, No. 1 (online), DOI: 10.1145/3191770 (2018).

[27] Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.-P. and Theobalt, C.: MonoPerfCap: Human Performance Capture From Monocular Video, *ACM Trans. Graph.*, Vol. 37, No. 2, pp. 27:1–27:15 (online), DOI: 10.1145/3181973 (2018).

[28] Yan, S., Xiong, Y. and Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition, *Thirty-second AAAI conference on artificial intelligence* (2018).

[29] Zhang, Y., Xie, B., Huang, H., Ogawa, E., You, T. and Yu, L.-F.: Pose-Guided Level Design, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, ACM, pp. 554:1–554:12 (online), DOI: 10.1145/3290605.3300784 (2019).

[30] Zhou, Z., Tsubouchi, Y. and Yatani, K.: Visualizing Out-of-Synchronization in Group Dancing, *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, New York, NY, USA, Association for Computing Machinery, p. 107–109 (online), DOI: 10.1145/3332167.3356888 (2019).

ⓒ 2020 Information Processing Society of Japan