Vol.2018-CVIM-212 No.13
2018/5/10

人体姿勢アノテーション困難な映像における類似姿勢学習の有用性

村上 達哉[1,a)]　三輪 誠[1]　浮田 宗伯[1]

概要：我々は，既存の大量映像コンテンツからの演技特徴データセット生成を目指している．こうした演技特徴は，人の姿勢により表現可能であるため，このデータセット生成のためには，映像からの姿勢推定が基礎技術として必要になる．しかし，現在の姿勢推定手法では，演技特徴を表現するに足るほどの高精度な姿勢推定は実現できていない．本研究では，深層学習による姿勢推定モデルにおいて，そのモデルを特定の動きに特化して最適化することで，姿勢推定の精度にどのような影響が与えられるかを検証する．本稿では，特徴的な動きが色濃く表れる時代劇の殺陣を解析対象とした．時代劇の演者のほとんどは，着物を着ており，その姿勢推定モデル学習のために姿勢アノテーションを与えることが困難である．そこで，姿勢推定対象となる演者と同様の動きを，モーションキャプチャシステムと同期させつつ映像撮影することで，各フレームへの姿勢アノテーションを自動化した．しかし，このアノテーション映像の演者は，モーションキャプチャ計測のために着物を着用していない．このように，姿勢推定モデルが学習している情報のうち「見え方」は大きく異なるが「人体パーツ間の相対的な位置関係」は類似している学習データを利用することで，姿勢推定モデルの最適化がどのように影響を受けるのかを検証するのが，本稿の目標である．実験では，人体パーツによっては最大 6 %程度の性能向上が得られることを確認した．

キーワード：姿勢推定，単眼画像，深層学習，畳み込みニューラルネットワーク

1. まえがき

3DCG アニメーションの需要の高まりに伴い，キャラクターを動かすアニメーターの負担が増大している．モーションキャプチャシステムを用いれば，アニメーターの負担は大きく減るが，マーカーの位置から姿勢を取得する分，得られるデータはマーカー位置のずれによるノイズが避けられない．さらに人物の動きをより魅力的にするために，アニメーターによる調整はかかせない．このように，アニメータにかかる負担は 3DCG アニメーション制作の発展を阻害する要因となっている．

一方で，デジタルデータである 3DCG アニメーションは，一度動きのデータセットを作成すれば制作費の大幅な削減が可能になる．しかし，人間の演技特徴の多様性を考慮すると，一つ一つ動きのデータを作ることは現実的ではない．そこで，既に人間の多様な演技特徴の情報を含んだデータである，既存の映像コンテンツの活用を考える．

本研究では，動きのデータセット作成のために，大量の映像コンテンツから演技の特徴を解析することを最終的な目標とする．そうした演技の特徴は，人の 3 次元的な姿勢で定義されるので，特徴解析には姿勢推定技術を利用する．姿勢推定技術とは，画像に写る人物の関節や端点（例：肩，肘，手首）といったキーポイントの位置を画像から推定する技術（**図 1**）である．キーポイントの位置を画像上の位

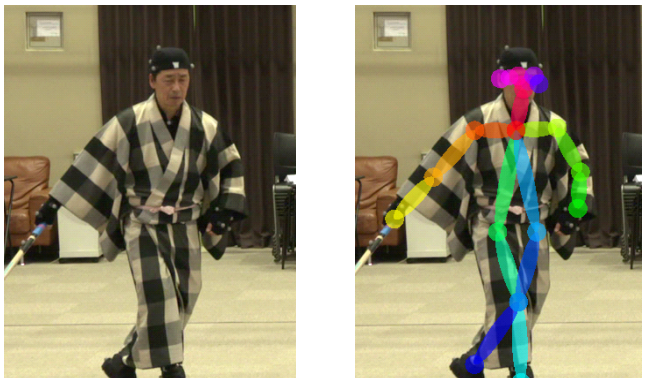


図 1　2 次元姿勢推定の結果．RGB 画像を入力として，その画像に写る人物のキーポイントの画像上における位置を推定している．図中では，キーポイントを色付きの円として表現している．

[1] 豊田工業大学
Toyota Technological Institute, Japan
[a)] sd14072@toyota-ti.ac.jp

置として推定する場合は 2 次元姿勢推定，空間上の位置として推定する場合は 3 次元姿勢推定と呼ばれる．

姿勢推定技術は，深層学習を用いた手法によって，推定精度は大幅に向上しており，特に 2 次元姿勢推定において顕著である．しかし，現状の 3 次元姿勢推定技術では，演技の特徴を得るに足る，高精度な推定は実現されていない．そこで，汎化ではなく，特定の個人/動きに特化することで 3 次元姿勢推定の精度を向上させる．推定精度の向上だけでなく，その対象が頻繁にとる姿勢や，他の人には見られない姿勢といった，動きの特徴を解析するための特徴抽出も可能になることが期待される．

このように，特定の個人/動きを表現するには，キーポイントの相対的な位置関係が重要である．例えば，ものまね芸人の動きは姿勢情報がオリジナルのものと物理的に異なるにも関わらず，われわれ人間にはオリジナルと似て見える．この時，ものまね芸人はオリジナルと似たような動きをするので，両者のキーポイントの相対的な位置関係は類似している．すなわち，似ていると感じる要因の一つは，両者のキーポイントの相対的な位置関係の類似性である．したがって，姿勢推定においてもこの情報を重視することで，より特徴を捉えた姿勢推定が可能になると考えられる．

この考えに基づき，今回は修士研究に向けた予備実験として，「特定の個人/動き」のキーポイントの相対的な位置関係を重視することによって，その「特定の個人/動き」を対象にしてどれほど姿勢推定精度を向上できるかを検証した．演技の特徴解析や 3 次元 CG への応用のためには，本来 3 次元姿勢推定が必要であるが，先に述べた通り 3 次元姿勢推定そのものの誤差が大きく，実験の際に誤差の原因を特定することが困難になりかねない．そこで本課題研究では，近年精度が急激に向上してきた 2 次元姿勢推定を対象として基礎検討を行う．

2. 関連研究

スポーツや医療目的の運動解析，ゲームなどのエンターテインメント用途への応用が期待され，姿勢推定に関する研究はコンピュータヴィジョンの分野で盛んに研究されてきた．現在の姿勢推定に関する研究は，入力となる画像の種類によって 2 種類に分けることができる．RGBD 画像と，RGB 画像である．

RGBD 画像では，RGB 画像に比べ，深度情報（Depth）を利用できる分，リアルタイムの 3 次元姿勢推定問題を単純化できる [1]．しかし，深度情報を得るには Kinect[2], [3], [4] に代表される RGBD カメラが必要であり，構築した姿勢推定モデルを適応する対象も RGBD 画像に限られてしまう．そうした制限を持つ RGBD 画像と比べ，RGB 画像からの姿勢推定は，既存の大量の画像データ（Youtube, Netflix など）を利用できるので，人間の動作解析の面で期待されている．

RGB 画像からの 2 次元姿勢推定研究の初期には，グラフィカルモデルをベースとした姿勢推定手法 [5], [6], [7] が多い．このモデルでは，構成されるグラフが木構造となるため，高効率な推定が可能であった．しかし，隣接していないパーツ間の依存関係が表現できていなかった．この問題を解決するために，様々な手法が考案されてきた．例えば，階層構造を取ったモデル [8] は，木構造による高効率性という利点を維持しつつ，パーツ間の依存関係の表現力を高めている．他には，パーツ間のエッジを増やすことによって，表現力を高めた非木構造のものがある [9]．このモデルはパーツ間の依存関係を十分に表現できるものの，グラフにループが存在することで高効率な計算を行えず，その予測値は近似した計算結果である．

このように，グラフィカルモデルベースの手法には隣接しないパーツ間の表現に問題があったが，現在優れた結果を残している 2 次元姿勢推定手法のほとんどが，CNN ベースの手法である．CNN を利用した姿勢推定手法の内で最も初期の取り組み [10] では，CNN を利用して回帰的に姿勢を推定している．CNN が利用され始めた頃にはグラフィカルモデルと CNN を組み合わせたものもあり，部分的なキーポイントの画像を，隣接するキーポイント位置の推定に役立てている [11]．こうした初期の取り組みと異なる，以降の 2 次元姿勢推定研究に大きな影響を与えたのが，ヒートマップと CNN を組み合わせた手法である [12], [13], [14]．代表的なものは，多段階の手法 [15] に CNN を利用することで，パーツ間の複雑な依存関係を捉えることを可能にした [14]．このように，CNN を利用した 2 次元姿勢推定手法は優れた結果を示し，その関心は「複数人に対する 2 次元姿勢推定」へと広がっている [16], [17], [18].

RGB 画像からの 3 次元姿勢推定でも CNN を用いた手法により，その推定精度を向上させている [19], [20], [21], [22]．時系列情報を用いてより安定で正確な姿勢推定を実現しているもの [23], [24] や，リアルタイム 3 次元姿勢推定 [25] の研究も行われている．一方で，3 次元姿勢推定は 2 次元姿勢推定と比べるといまだに難しいタスクであり続けている．2 次元的に体の姿勢が得られたとしても，3 次元的にみるとその姿勢を取りうる姿勢は幾通りも存在するからである．加えて，学習データが十分でないことも，3 次元姿勢推定を難しいタスクにしている．2 次元姿勢推定における深層学習の成功の大きな要因の一つは大量のデータセット [26], [27] が確保できたことであり，量と種類は 3 次元姿勢推定用のもの [28], [29] と比べると非常に多い．CG 技術を用いてデータセットを作る動き [30], [31] もあるが，これは本研究の主目的ではない．本研究では演技特徴を表現できるようなモデルを作りたいので，シーケンス情報を含んだ実際の人物の画像データセット [32] が重要になる．

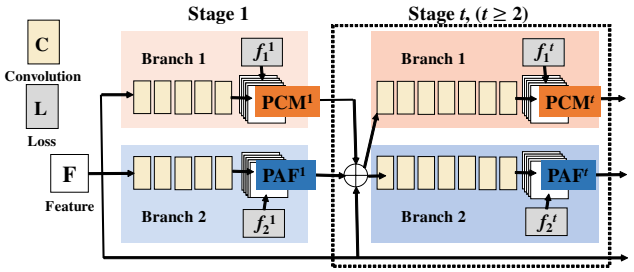


図 2 姿勢推定ライブラリ「OpenPose」で使われるアーキテクチャ．Branch 1（橙色）からは PCM が出力され，Branch 2（青色）からは PAF が出力される．Stage 1 において入力は画像から取り出した特徴量（F）のみであるが，点線で囲まれた Stage t ($t \geq 2$) においては特徴量 F に加え，Stage t-1 における出力も入力となる．黄色い四角は畳み込み層（C）を，灰色の四角は損失関数（L）を示している．

図 3 Part Confidence Map(PCM) と Part Affinity Field(PAF)．上段が右手首の PCM，下段が右前腕の PAF を示している．橙色の円で囲んだ場所は，ステージを経るごとに推定できるようになっているのが分かる．赤色の円で囲んだ場所は，ステージを経るごとに間違えなくなっているのが分かる．図は [18] より引用．

3. OpenPose の姿勢推定アルゴリズム

実験では姿勢推定ライブラリ「OpenPose」[14], [18], [33] を使用した．OpenPose は CNN ベースの精度の高い 2 次元姿勢推定を可能にするライブラリであり，細部にわたり整備され研究に際し扱いやすいため，実験に用いた．

図 2 に，OpenPose で使用される姿勢推定アルゴリズムの構造を示す．OpenPose の姿勢推定アルゴリズムは，Part Confidence Map (PCM), と Part Affinity Field (PAF) と呼ばれる二つの特徴量と，画像から直接得られる特徴量を CNN により多段階にわたって繰り返し抽出することで，姿勢推定の精度を高めている．ネットワークが PCM や PAF といった特徴量を Branch 1，Branch 2 から繰り返し出力できるように，各ステージの終わりにそれぞれ異なる損失関数 f1，f2 を適用している．具体的な PCM，PAF の画像を**図 3** に示す．

PCM とは，姿勢のキーポイントの位置の信頼度を示すヒートマップのことであり，画像上のスカラー場として表現される．図 3 では，上段に右手首に関する PCM の画像を挙げている．この図を見ると，Stage 1 では左手首を誤って右手首だと判断しているケース（図中赤丸）があるが，多段階にわたる処理を行うことで誤りを修正し，推定精度を向上させている．このヒートマップのピーク値を取得することによって，姿勢の座標を取得することができるが，OpenPose はボトムアップに PCM を作るため，画像上に複数人が存在する場合，問題が生じる．例えば，図 3 の上段には右手首の PCM を記載したのだが，最終的な Stage 6 において，二人の人物の右手首が画像から検出されており，このままでは検出したキーポイントを結んで一人の姿勢とすることができない．そのため，複数人の姿勢推定を行いたい場合には PCM だけでは不十分である．人体検知の手法を用いて，画像中の人物を一人検出したあとならば，その一人に対して CPM を用いた姿勢推定を行うことが可能である．しかしこうしたトップダウンのアプローチでは一人ずつキーポイント検出を行う必要があるため，時間がかかってしまう．

OpenPose の姿勢推定アルゴリズムは，PCM の他に PAF を用いることで，この問題を解決し，リアルタイムでの複数人の姿勢推定を可能にしている．PCM を用いて特定された複数人のキーポイント位置から，各人の姿勢を得るには各キーポイントを個々人に結び付ける必要がある．そこで使用されるのが PAF である．PAF は，ベクトル場であり，画像における個人の四肢上においては，その四肢が結ぶキーポイントに沿った方向を向いたベクトルになるように設計されている．よって，PAF を用いると PCM によって得られたキーポイントを結ぶことができ，各個人の姿勢を推定することができる．さらには PAF を利用してキーポイントを結ぶ際には貪欲法を用いた高速な処理が行われているため，リアルタイムでの複数人に対する姿勢推定が可能となっている．図 3 の下段には右上腕における PAF を記載した．PCM と同様に，多段階処理からその精度を向上させている．今回の実験では対象とする人物が一人だけなので，PAF による姿勢推定精度の向上は限定的であると予想される．しかし，今後の研究にて解析対象とする映画のような一般的な映像では，複数人が重なり合うことによる遮蔽が頻繁に生じる．そういった状況においては，PAF が姿勢推定の精度向上に大きく寄与することが期待される．

4. 類似姿勢学習の有用性の検証法

今回の実験では，「特定の個人/動き」のキーポイントの相対的な位置関係を重視することによって，「特定の個人/動き」に対する姿勢推定精度をどれほど向上できるかの検証を目標とする．

この検証のためには，実験結果における違いが際立つように，一般的な人物の動きの特徴ではなく，個人の演技特

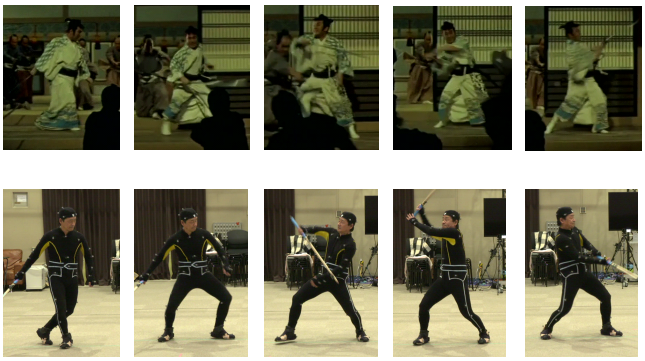
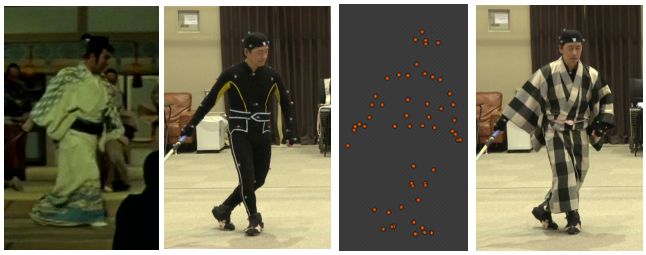


図 4 同じような動きのシーケンス．解析対象の動き（上段）と同じような動き（下段）をモーションキャプチャシステムによって計測することで，解析対象における，キーポイントの相対的な位置関係と似たような情報を取得できる．

図 5 モーションキャプチャシステム（ツークン研究所）．収録エリアサイズは 10m × 7m × 2.5m で，24 台の Vicon 社「T160」という光学式カメラを用いて撮影できる．



(a) **(b)** **(c)** **(d)**

図 6 計測したデータ．(a) 解析対象となる映像．着物の遮蔽によりキーポイントの位置が特定しにくい．(b) モーションキャプチャシステムでマーカー計測時の映像．解析対象と見え方は大きく異なるが，キーポイントの相対的な位置関係は類似している．(c) モーションキャプチャシステムから計測した，マーカーの 3 次元座標の可視化．橙色の点がマーカーに対応しており，その 3 次元的な座標を，CG ソフトウェアを介して取得することができる．(d) テスト用の映像．解析対象の映像は，ノイズや他の人物による遮蔽が大きく，姿勢推定精度に与える影響が強い．そのため，テスト用画像を作成するために，着物を着た状態での類似した動きを撮影した．

徴が色濃く表れる映像が有効である．そうした映像として，日本固有のコンテンツである時代劇，特に典型的な演技として，殺陣を実験対象とする．理想的には，時代劇の殺陣に対して高精度な姿勢推定モデルを作成したい場合，キーポイントの相対的な位置関係や，見え方に対する姿勢をモデル化するために，学習データにも時代劇の殺陣の画像を使用するほうが良い．しかし，時代劇では，登場人物が着物という遮蔽の大きな衣服を着用しているため，体のパーツが画像から確認しずらく，画像に対する姿勢のマニュアルアノテーションが困難である．そのため学習データを大量に獲得することが難しい．そこで，解析対象とする殺陣の動きと類似した動きをモーションキャプチャシステムを用いて計測した（**図 4**）．これにより，見え方は大きく異なるものの，殺陣の動きにおけるキーポイントの相対的な位置関係と類似した情報を持つ学習データを，マニュアルアノテーションなしに大量に作成することが可能である．計測したデータから画像上のアノテーションデータを得る具体的な方法は 5.2 節で述べる．こうして得られた学習データは，姿勢推定の対象となる画像（図 4 上段）を解析する際に有用な，キーポイントの相対的な位置関係に関する情報を含んでいる．一方で，学習データ中の人物はモーションキャプチャスーツしか着ておらず，見え方に関する情報が不足している．この情報不足を補うために，作成した学習データを用いて姿勢推定モデルを構築する際には，見え方について多様な情報を含んだデータセット [26] から学習したモデルをファインチューンする．作成したモデルを用いて，純粋に類似姿勢学習の有用性を検証するには，ノイズや遮蔽の少ない画像が望まれる．解析対象（図 4 上段）はノイズや遮蔽が多いため，検証に向いていない．よって，解析対象と似た動きを，着物を着た状態で撮影し，検証時のテストデータとする．

5. 実験

5.1 殺陣データセットの計測

実験に必要となる殺陣のデータは，東映デジタルセンターツークン研究所にて撮影を行った（**図 5**）．東映株式会社の長年にわたる映像制作の実績から，ツークン研究所では様々なジャンルのアクターを国内外から手配することができる．今回は，解析したい映像における殺陣の動き（**図 6**(a)）を，プロの殺陣師にアクターとして再現してもらうことで，その殺陣の動きと類似した動きをモーションキャプチャシステムにて計測（図 6(c)），それと対応した映像（図 6(b)）も撮影した．また，本来であれば構築した姿勢推定モデルの精度を検証するために，解析対象となる映像（図 6(a)）でテストを行いたいが，その映像はノイズや

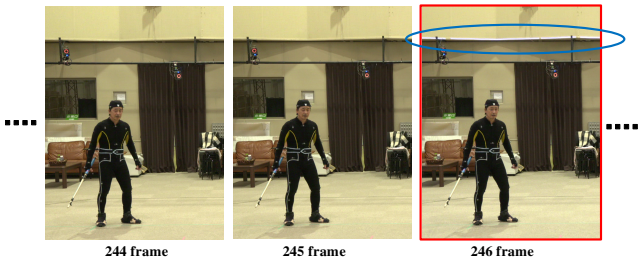


図 7 カチンコによる映像とモーションデータの同期．同期のためには，カメラとモーションキャプチャ用カメラから，撮影が開始する瞬間を確認できる必要がある．撮影開始時（図中 246 frame）に強い瞬間的な光を発する（図中青丸）ことによって，画像上からも，モーションキャプチャシステムからも，撮影開始のタイミングが判断できる．

他の人物による遮蔽が多く，実験の狙いに関係なく，姿勢推定精度に影響を与えてしまう可能性がある．そのため，今回はテスト用の映像として，着物を着た状態での類似した動き（図 6(d)）の撮影も行った．

データセットの計測に用いた機器はそれぞれ以下に述べる．マーカーの 3 次元座標（図 6(c)）の計測に用いたモーションキャプチャ用のカメラは Vicon 社の T-160 である．T-160 は 1600 万画素（4704 × 3456）の解像度を持ち，フル解像度で 120FPS のキャプチャが可能である．一方，学習用の映像（図 6(b)）と，テスト用の映像（図 6(d)）の撮影に用いたカメラは，Sony 社の PXW-FS7 で，約 1160 万画素の解像度で 60FPS の撮影を行える．

計測されたデータセットは，学習用，テスト用でそれぞれシーケンス数が 5 つである．本実験では，その中でも特に類似した動きを再現できている 1 シーケンスのデータを用いて，学習用データセットを作成した．

5.2 学習用データセットの作成

モーションキャプチャシステムから得られたマーカー（図 6(c)）のシーケンス情報と，それに対応する映像（図 6(b)）は，フレーム間の対応がついていないため，まずはデータ間の同期を行う必要がある．このために，映画撮影の際などに使われるカチンコから得られる情報を利用した．モーションキャプチャシステムにおいては，光がカチンコの役割を果たしている（**図 7**）．図中では 246 フレーム上で光が発せられたと確認できるので，学習用の画像のシーケンスには 246 フレーム目からの画像を使用する．撮影終了時も同じ原理でフレームを合わせる．結果，246 フレームから 1073 フレーム目までの計 828 枚の画像を学習データに使用することができる．画像に対する姿勢情報のアノテーションに関しては，計測されたキーポイントの 3 次元座標を，カメラ校正により得られたカメラパラメータによって，2 次元画像座標に変換する．カメラ校正のためには，モーションキャプチャスーツに着けるマーカーの位置情報を利用した．まずモーションキャプチャシステムによってマーカーの 3 次元座標が取得できる（図 6(c)）．さらにマーカーは画像上からもその 2 次元座標を確認することができる（図 6(b)）．したがって，空間上の 3 次元座標とそれに対応する画像上の 2 次元座標のペアを取得することができる．この座標のペアを 50 組ほど集め，空間座標から画像座標への透視投影行列を作成した．この作成した透視投影行列を用いることで，モーションキャプチャシステムから計測されたキーポイントを 3 次元座標から 2 次元座標に投影することができる．具体的には，「1 フレームごとに 3 次元座標を取得し，2 次元座標へ投影する」という処理をすべて CG ソフトウェア上で行うことができるので，1 つの映像シーケンスからアノテーション付きの画像 828 枚を容易にかつ大量に作ることができる．

図 8 再最適化したモデルを用いた姿勢推定結果．着物の遮蔽と形状変形によって推定エラー（図中赤丸）が生じている．

5.3 実験結果

実験では，一般的な人の画像と姿勢情報を含むデータセット COCO[26] で学習した姿勢モデルと，この姿勢モデルを殺陣の画像と姿勢の教師データ（Toei+COCO と記述）で再最適化した姿勢モデルを用いて，着物を着た殺陣の画像に対して姿勢推定を行った．COCO は，人が密集した状態，遮蔽が大きい状態など，様々な問題ケースの画像を含んだデータセットであり，330K の画像に対し，250,000 個の姿勢情報がアノテーションされている．

図 8 に推定結果の可視化を示す．赤い丸で囲んだ部分で推定エラーが生じているが，動きの激しさから着物が複雑に変形している画像におけるエラーが多かった．これは見え方から生じるエラーであり，連続的な動きの特徴を表現できていないことが分かる．

また，**図 9** に，COCO と，Toei+COCO で学習したモデルの PCKh による評価結果を示す．PCKh とは姿勢推定においてしばしば用いられる指標であり，Part Correct Keypoint (Head) の頭字語である．これは，頭の大きさを 1 として正規化した距離に基づき，ある閾値を定めたときに，キーポイントの正解座標と推測結果との間の距離が，その閾値の範囲にどれほど収まっているかを表したものである．図 9 を見ると，限定的だが微小な精度向上を確認できる．例えば図中右図の肘に関する PCKh では，横軸が 0.11 の所において，6 パーセント（16 パーセントから 22 パーセントに上昇）の推定精度の向上を果たしている．

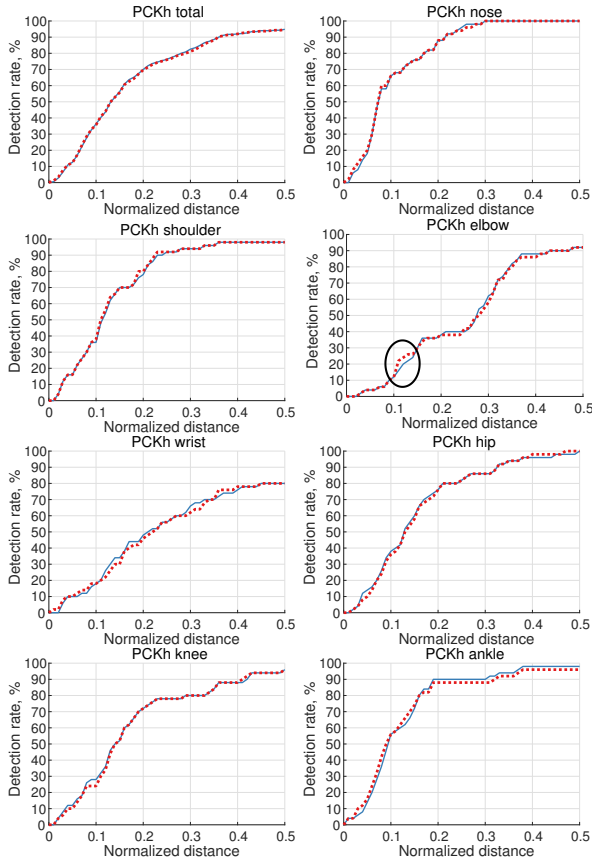


図 9 体の各部分に関する PCKh．青色の実線が，COCO のみで学習したモデルの推定結果，赤色の点線が，Toei+COCO で学習したモデルの推定結果である．体全体（PCKh total）でみるとあまり違いは見られないが，各パーツで見ると，推定精度が向上しているところも見られる．例えば，肘に関する PCKh（PCKh elbow）では，Normalized distance（横軸）が 0.11 の所において精度が 6 パーセント向上している（図中黒丸）．

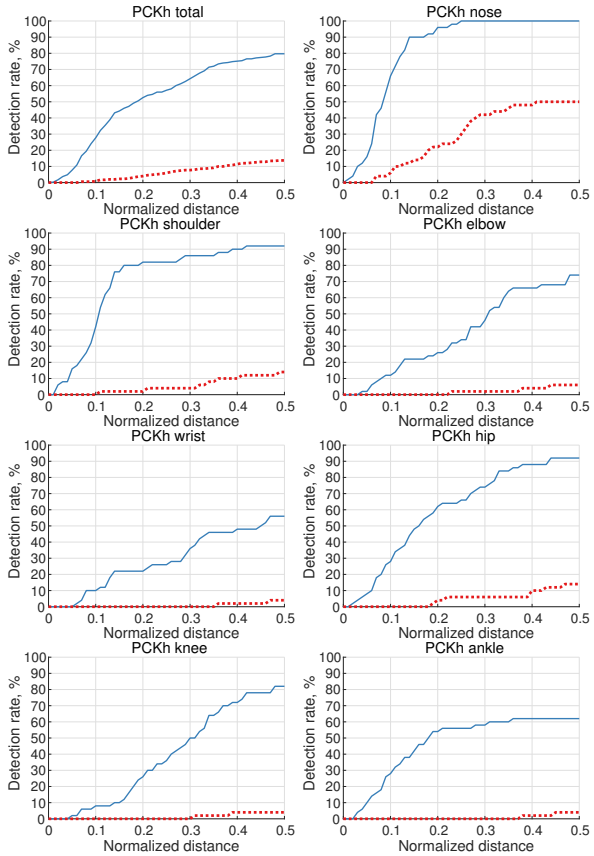


図 10 体の各部分に関する PCKh．青色の実線が，COCO のみで学習したモデルの推定結果，赤色の点線が，Toei のみで学習したモデルの推定結果である．

さらに，**図 10** に，COCO のみで学習したモデルと，Toei のみで学習したモデルの PCK h による評価結果を示す．多様な見え方の情報を含む COCO で学習したモデルの方が，より優れた推定結果を出している．

これらの実験結果より，特定の動きのデータセットで姿勢推定モデルを再最適化することで，特定の動きに対する姿勢推定の精度を高めることができる可能性を示せたが，大幅な精度向上は示せなかった．

5.4 考察

精度向上が限定的だった大きな理由として，姿勢を静止画で表現し，動きを表現できていなかったことが考えられる．着物とスーツでは見え方が大きく異なり，キーポイントの相対的な位置関係はスーツの画像に対して学習されているので，着物の画像に対しては学習結果を活かした姿勢推定を行えなかった．そのため，相対的なキーポイントの位置関係の情報を活用した姿勢推定を行いたいのであれば，その見え方に大きく依存しない手法が必要である．

6. むすび

6.1 まとめ

一般的な人物を対象にした 3 次元姿勢推定ではなく，「特定の個人/動き」に特化した 3 次元姿勢推定という最終目的への取り組みとして，今回は「特定の個人/動き」のキーポイントの相対的な位置関係を重視し，2 次元姿勢推定の精度を，「特定の個人/動き」に対して向上できるかどうかを検証した．この検証では，特徴的な動きが表れる「殺陣」を対象にして，モーションキャプチャシステムから得られる画像上の 2 次元姿勢と対応する画像のデータセットから，キーポイントの相対的な位置関係をモデル化した．このモデルを用いた検証のため，モーションキャプチャ撮影時の動きを着物を着た状態で再現し，再現画像に対して姿勢推定を行った．その結果，姿勢推定精度の向上は限定的であり，「特定の個人/動き」に特化した姿勢推定を行うためには別の手法が必要だと分かった．

検証に用いた姿勢推定アルゴリズムは，画像 1 枚のみを入力として画像上の人物の姿勢を推定しており，画像上の人物の見え方に大きく依存したアルゴリズムと言える．姿勢推定モデルの構築には，モーションキャプチャースーツの画像が使用されているため，着物を着た状態の画像に対

しては，モデル化したキーポイントの相対的な位置関係を，その見え方の違いから上手く姿勢推定に活かせなかった

6.2 今後の発展

検証に用いた姿勢推定アルゴリズムは，画像上の人物の見え方に大きく依存したものであった．そこで，今後の研究においては，時系列データを用いることで，着物映像の姿勢推定精度を向上させていく．時系列データの利用で姿勢推定精度を向上できることが分かっており，また，個人に特化した姿勢推定を行う観点でも，時系列データの使用は有効である．例えば，人の歩き方からその人物を特定する，歩容認証と呼ばれる問題では，個人の動きの特徴は時系列データに良く表れるこが明らかになっており，動きの時系列データは個人に固有のものである．したがって，「特定の個人/動き」の時系列データを利用することで，「特定の個人/動き」に特化して高精度に働く姿勢推定モデルの実現を目指していく．

謝辞 本研究で使用したデータは東映デジタルセンターツークン研究所から提供されている．また，本研究は京都大学インキュベーションプログラムの支援を受けて行った．

参考文献

[1] 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).
[2] Kinect for Xbox 360, Microsoft Corporation (2010).
[3] Kinect for Xbox One, Microsoft Corporation, (online), available from ⟨http://www.xbox.com/ja-jp/xbox-one/accessories/kinect/⟩ (2013).
[4] Kinect SDK, Microsoft Corporation, (online), available from ⟨https://developer.microsoft.com/ja-jp/windows/kinect/⟩ (2015).
[5] Felzenszwalb, P. F. and Huttenlocher, D. P. Pictorial structures for object recognition, *IJCV*, Vol. 61, No. 1, pp. 55–79 (2005).
[6] Ferrari, V., Marin-Jimenez, M. and Zisserman, A. Progressive search space reduction for human pose estimation, *CVPR* (2008).
[7] Andriluka, M., Roth, S. and Schiele, B. Pictorial structures revisited: People detection and articulated pose estimation, *CVPR* (2009).
[8] Sun, M. and Savarese, S. Articulated part-based model for joint object detection and pose estimation, *ICCV* (2011).
[9] Dantone, M., Gall, J., Leistner, C. and Van Gool, L. Human pose estimation using body parts dependent joint regressors, *CVPR* (2013).
[10] Toshev, A. and Szegedy, C. Deeppose: Human pose estimation via deep neural networks, *CVPR* (2014).
[11] Chen, X. and Yuille, A. L. Articulated pose estimation by a graphical model with image dependent pairwise relations, *NIPS* (2014).
[12] Tompson, J., Goroshin, R., Jain, A., LeCun, Y. and Bregler, C. Efficient object localization using convolutional networks, *CVPR* (2015).
[13] Pfister, T., Charles, J. and Zisserman, A. Flowing convnets for human pose estimation in videos, *ICCV* (2015).
[14] Wei, S.-E., Ramakrishna, V., Kanade, T. and Sheikh, Y. Convolutional pose machines, *CVPR* (2016).
[15] Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J. A. and Sheikh, Y. Pose machines: Articulated pose estimation via inference machines, *ECCV* (2014).
[16] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V. and Schiele, B. Deepcut: Joint subset partition and labeling for multi person pose estimation, *CVPR* (2016).
[17] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. and Schiele, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model, *ECCV* (2016).
[18] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *CVPR* (2017).
[19] Zhou, X., Zhu, M., Leonardos, S., Derpanis, K. G. and Daniilidis, K. Sparseness meets deepness: 3D human pose estimation from monocular video, *CVPR* (2016).
[20] Brau, E. and Jiang, H. 3D Human Pose Estimation via Deep Learning from 2D Annotations, *3DV* (2016).
[21] Tome, D., Russell, C. and Agapito, L. Lifting from the deep: Convolutional 3d pose estimation from a single image, *CVPR* (2017).
[22] Martinez, J., Hossain, R., Romero, J. and Little, J. J. A simple yet effective baseline for 3d human pose estimation, *ICCV* (2017).
[23] Lin, M., Lin, L., Liang, X., Wang, K. and Cheng, H. Recurrent 3D Pose Sequence Machines, *CVPR* (2017).
[24] Hossain, M. R. I. and Little, J. J. Exploiting temporal information for 3D pose estimation, *arXiv* (2017).
[25] Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D. and Theobalt, C. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera, (online), available from ⟨http://gvv.mpi-inf.mpg.de/projects/VNect/⟩ (2017).
[26] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. Microsoft COCO: common objects in context, *ECCV* (2014).
[27] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, *CVPR* (2014).
[28] Sigal, L., Balan, A. O. and Black, M. J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, *IJCV*, Vol. 87, No. 1-2, p. 4 (2010).
[29] Ionescu, C., Papava, D., Olaru, V. and Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, *TPAMI*, Vol. 36, No. 7, pp. 1325–1339 (2014).
[30] Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D. and Chen, B. Synthesizing training images for boosting human 3d pose estimation, *3DV* (2016).
[31] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I. and Schmid, C. Learning from Synthetic Humans, *CVPR* (2017).
[32] Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B. and Schiele, B. ArtTrack: Articulated Multi-person Tracking in the Wild, *CVPR* (2017).
[33] Simon, T., Joo, H., Matthews, I. and Sheikh, Y. Hand Keypoint Detection in Single Images using Multiview Bootstrapping, *CVPR* (2017).