DEIM2022 D23-2 (day2 p19)

画角不変な姿勢特徴量を用いた類似行動検出

吉田　登† 劉　健全†

† NEC バイオメトリクス研究所　〒 211–8666 神奈川県川崎市中原区下沼部
E-mail: †{n-yoshida14,jqliu}@nec.com

あらまし　人物の姿勢情報に基づく行動検索、行動検出技術が盛んに研究されているが、実運用を想定した場合、2 つの大きな課題がある。1 つは、姿勢情報の欠損への対処、もう 1 つは人物毎に異なる行動のバリエーション、すなわち個人差に対する頑健性の確保である。これらの課題を解決し、高精度な行動検索技術を実現するために、我々は画角変化に対して頑健であり、かつ情報の欠損や個人差に対して柔軟に対応できる姿勢特徴量と類似度関数を提案する。実験では、1 つの姿勢データをクエリに用いた類似姿勢検索と、時系列姿勢データをクエリに用いた類似行動検索において、最新手法よりも高い精度を実現した。また本発表では、本手法を実装した、オンデマンドにユーザーが検出したい行動を検出可能なデモシステムについてもインタラクティブに紹介する。

キーワード　画角不変な姿勢特徴量、行動検索、行動検出

1 背　景

人物の行動検出技術や行動認識技術は 20 年以上に渡って研究されており [1]、その用途は公共の場における転倒やしゃがみ込み等の不安全行動の検出や、車いす利用者等の要救援者検出、さらには映像編集のために大量な記録映像の中から特定の行動シーンを抽出するなど、多岐に渡っている。行動検出に着目した既存技術として、機械学習を用いた手法や [2–6]、近年では特に深層学習を用いた手法が多く提案された [7–16]。

しかしながら、様々な状況の変化に対応して頑健な検出を行うためには、例えば (1) 服装や持ち物などの人物の外観や、(2) 照明条件や背景などの撮影環境や、(3) カメラの焦点距離や設置位置などのカメラパラメータ等に頑健な特徴量を得る必要がある。そのためにはこれらのバリエーションを網羅した学習データが必要であるが、学習データの収集、及び正解付けは非常に高コストである。

この問題に対し、機械学習ベースの手法ではなく、クエリとの類似度判定による検索ベースの手法が提案された。本手法は学習データの収集が不要であるというメリットがある一方で、頑健な検索を実現するためには、上述した要素に対して頑健な特徴量を用いる必要がある。例えば、オプティカルフローやパターンヒストグラム等を特徴量に用いた行動検索手法が提案された [17–25]。

中でも、人物の姿勢情報は上述した要素の内 (1) 人物の外観、(2) 撮影環境、に対して頑健であることから、行動検索に適した特徴量として注目されており、姿勢情報を用いた検索手法が複数提案された [22, 23]。さらに、深層距離学習を用いて 3 次元姿勢とそれを様々な画角を想定して投影した 2 次元姿勢のペアを学習することで、2 次元姿勢情報を (3) カメラパラメータ等に不変な埋め込み特徴に変換する技術が提案され、(1)〜(3) の要素すべてに頑健な行動検索が実現された [24]。

しかしながら、実環境での利用を想定した場合、さらに 2 つの課題がある。

1 つ目は、同一の行動であっても、人物毎に姿勢やその変化の仕方に違いが生じることへの対処である。既存手法では全身の姿勢情報を均等に扱って類似判定を行うため、異なる姿勢を同一の行動と判定することができない。この課題を解決するために、我々は体の特定の部位に重みづけをした類似度判定手法を提案する。

2 つ目の課題は、姿勢情報の欠損への対処である。実環境の映像では、人同士の重なりや人と物の重なりによって、推定された姿勢情報が頻繁に欠損する。一般的な深層学習を用いた手法では、類似度判定の前処理として欠損情報を補完することが多い。しかしながら補完された情報の信頼性は限定的であり、不十分である場合もある。そのため、時には欠損点は補完せず、見えている情報のみを用いて類似判定を行う方が有効であると考えられる。

以上の背景から我々は、画角に対して不変になるように変換した姿勢情報に基づく、類似行動検索技術の開発を目標とした。そして上述した 2 つの課題を解決して高精度な検索を実現するために、特定部位への重みづけや欠損情報への対応を可能にする柔軟な類似度判定手法と、それを可能にする新しい特徴量として、特徴量の各次元と各部位の情報が 1 対 1 で対応する特徴量を提案する。

本手法を実装し、1 枚の画像や数秒の動画をクエリとして類似の行動を検索するシステムを構築した (図 1)。実験では、姿勢の正解情報であるモーションキャプチャーデータと実映像の公開データセットである UT-kinect [26] を用いて、定量的な検索精度の評価を実施した。その結果、我々の手法は既存の最新手法を上回る精度を実現した。

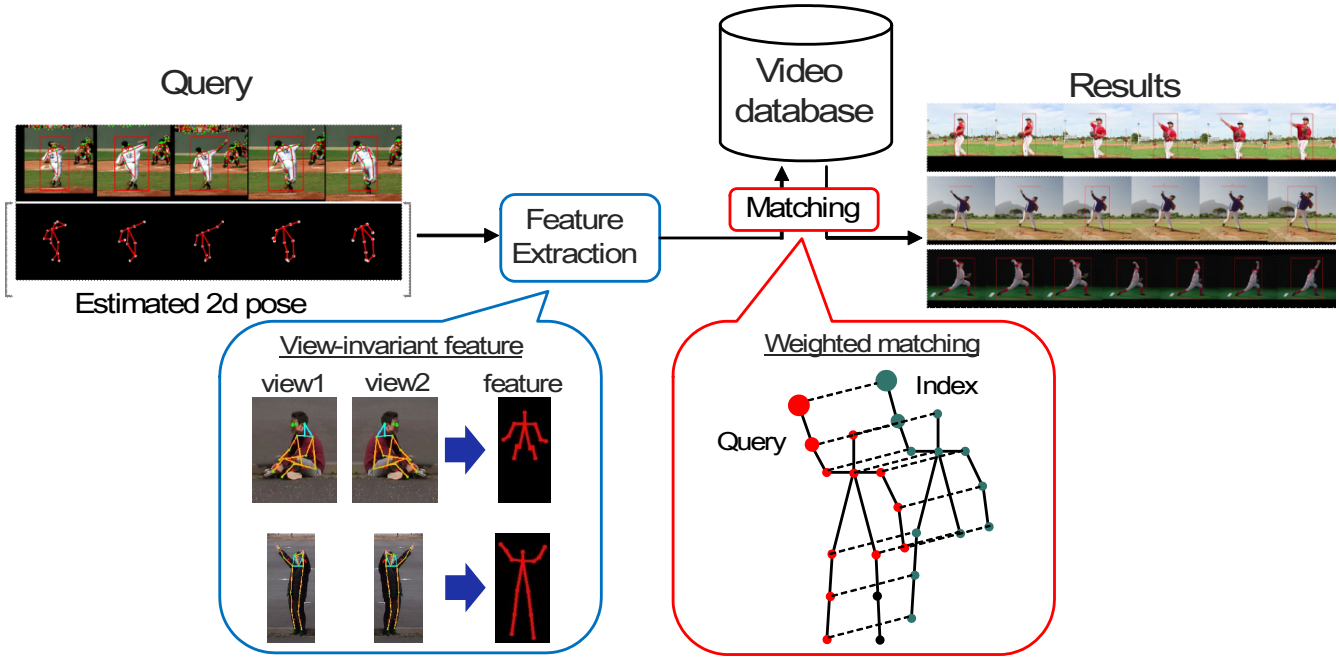

図 1　提案手法の全体像。画角変化に対して頑健な特徴量を抽出手法と、破線の有無や円の大きさで示すように、欠損を無視したり特徴的な部位に対して重みを付与したりすることが可能な、柔軟な照合方法を有する。

2　関 連 研 究

2.1　2 次元姿勢推定技術

近年、様々な姿勢推定技術が提案され、行動認識 [27–29] や人物のカメラ内追跡 [30, 31]、カメラ間追跡 [32] 等に応用されている。姿勢情報とは、画面上における人物の首や肘等の関節を含む 18 点のキーポイントの座標情報を示し、人の外観や照明条件や背景などに対して不変な性質を持つ。

姿勢推定アルゴリズムは大きく Top-down 型 [33,34] と Bottom-up 型 [35–38] に分けられる。Top-down 型では、初めに人体検出を行い、検出された人体矩形に対して関節点等のキーポイント検出を行う。しかしながら、実世界の防犯映像などにおいては、混雑による人同士の重なりが頻繁に発生し、結果として人体検出が失敗しやすい傾向にある。この問題は、Gkioxari らも自身の研究において指摘している [39]。

このような実環境での利用において生じる問題に対しては、Bottom-up 型の手法の方がより適している。Bottom-up 型では、初めに画像全体から全ての部位のキーポイントを検出した後、これらを関連づけることで個人ごとの姿勢情報を構成する。結果として、一部のキーポイントが欠損するような状況でも、頑健な検出が可能になる。

我々の研究においても、隠れに対して頑健な検索を可能にすることを目的に、Bottom-up 型の姿勢推定技術を用いた。

2.2　類似姿勢検索

ユーザーが指定した姿勢情報をクエリに用いて、類似した姿勢の人物を映像中から検索する手法が提案された [22, 23, 25]。クエリの設定方法としては、例えばユーザーが特別なユーザーインターフェース上で、人物の姿勢を模した棒人間を手動で操作する方法 [23] や、Kinect センサー等のデータを入力する方法 [23] や、上述したような姿勢推定エンジンの出力を用いる方法 [22, 23] や、ユーザーが書いた手書きの棒人間情報を用いる方法 [25] 等である。

ところで、これらの姿勢検索技術は、人とカメラの位置関係や人の向きが変化した場合に姿勢情報が大きく変化してしまい、結果として検索精度が大きく下がる。この問題を解決するために、3 次元姿勢とそれを様々な画角を想定して投影した 2 次元姿勢のペアを学習することで、2 次元姿勢情報を画角不変な特徴量に変換する手法が提案された [24]。しかしながらこの手法においても、実環境を想定した場合にはまだ以下のような問題が残る。

1 つ目は、同一の行動であっても、人物毎に姿勢やその変化の仕方に違いが生じることへの対処である。例えば、腕を組んで座っている人と、腕を組まずに座っている人を考えた場合、両者は座っているという同じ行動をしているにも関わらず、両者の姿勢は同一ではない。従来の類似姿勢検索手法 [24] では、全身の姿勢情報を均等に扱って類似判定を行うため、本ケースでは両者を同一の行動と判定することができない。この課題を解決するために、我々は体の特定の部位に重みづけをした類似度判定手法を提案する。

2 つ目の課題は、姿勢情報の欠損への対処である。実環境の映像では、人同士の重なりや人と物の重なりによって、推定された姿勢情報が頻繁に欠損する。従来の手法 [24] を含む一般的な深層学習を用いた手法では、入力となる姿勢情報の次元数が固定されているため、前処理として欠損情報を補完することが多い。例えば、欠損した部位以外の情報を用いて補完したり、時系列情報に基づいて補完する、などである。しかしながら補完された情報の信頼性は限定的であり、不十分である場合もあるため、時には欠損点は補完せず、見えている情報のみを用い

て類似判定を行う方が有効であると考えられる。

これら 2 つの課題を解決して高精度な検索を実現するために、特定部位への重みづけや欠損情報への対応を可能にする柔軟な類似度判定手法と、それを可能にする特徴量として、特徴量の各次元と各部位の情報が 1 対 1 で対応することを特徴とした新規特徴量を提案する。

2.3 類似行動検索

上述した姿勢検索の応用として、時系列姿勢情報をクエリとした類似行動検索技術が提案された [18, 25]。多くの研究は Kinect センサーや磁気センサー、モーションキャプチャー等で取得した 3 次元姿勢情報を用いることを前提としているが、一方で、Sun らは 2 次元姿勢のみを入力として、それを画角不変な特徴量に変換し、画角不変な行動検索を実現する手法を提案した [24]。

既存の類似行動検索技術では、2 つの映像に対して画像フレーム間の対応付けを Dynamic Time Warping (DTW) [40] アルゴリズムによって行い、フレーム間距離の合計を用いて映像間の類似度を求める方法が一般的である。

我々の研究においても、Sun らとは異なる方法によって新しい画角不変な姿勢特徴量を得る。さらに本特徴量と柔軟な類似度判定手法を用いた、映像から類似の行動を検索する手法を提案する。従来技術と同様に、フレーム間の対応付けは DTW によって行い、フレーム間距離の算出において、部位への重みづけや欠損への対処を考慮した。

3 提案手法

3.1 画角不変な姿勢特徴量

上述のように、我々の手法でも画像から推定された 18 個のキーポイントからなる姿勢情報を元に、これを画角不変な特徴量に変換して検索に用いる。既存の手法 [24] でも深層距離学習によって画角不変な姿勢特徴量に変換する方法が提案されたが、この手法で変換された特徴量は、元の姿勢情報が持つ部位のラベルを保持していない。すなわち、特徴量ベクトルの各次元と、元の姿勢ベクトルの各次元には対応関係がない。

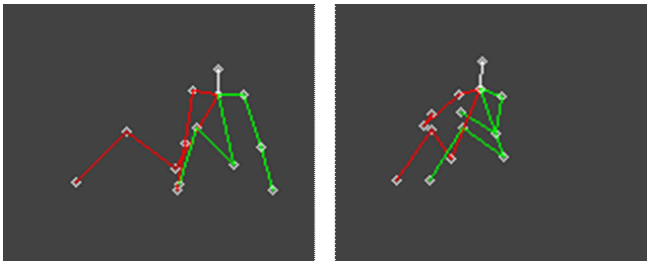

図 2　腕を組まずに（左）、腕を組みながら（右）、地面に座っている姿勢

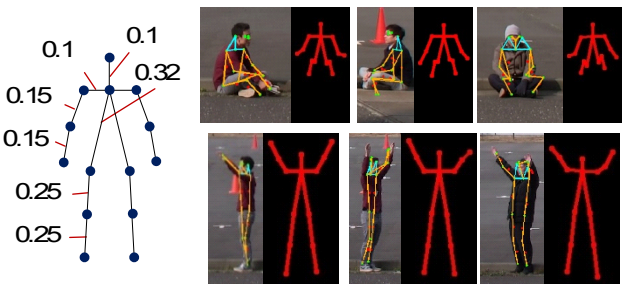

図 3　標準的な人体のモデル（左）と、2 次元姿勢とそこから変換され可視化された提案特徴量のペア（右）。

ここで、実世界の映像監視において、地面にしゃがんでいる人物を検索するというシステムを想定する。そして、例えば図 2 のように、腕を組んでしゃがんでいる人物と、腕を組まずにしゃがんでいる人物がいたとする。この場合、両者とも地面にしゃがむという同一の行動を行っているにも関わらず、従来の手法では、姿勢全体同士を均等に考慮して類似の判定を行うことから、片方の人物の姿勢情報をクエリとしてもう片方の人物を検索することはできない。

しかしこのケースにおいて、おそらくシステムユーザーはしゃがむという行動において重要な下半身の類似性に、より重きを置く、または上半身の情報は無視することで 2 つの姿勢を同一の行動と判定できる、というアイディアに至るだろう。ただし、このアイディアを実現するためには、変換された画角不変な姿勢特徴量が、元の姿勢情報が持つ部位に関するラベル情報を保持している必要がある。そこで我々は本思想に基づいて、特徴量ベクトルの各次元が元の姿勢ベクトルの各次元と 1 対 1 に対応するような特徴量を設計した。

そしてこの設計思想は、欠損情報の取扱いにおいても大きな優位性を持つ。従来手法では、深層ネットワークの入力次元数が固定であることから、前処理として欠損したキーポイント情報を補完する。例えば、周囲の見えているキーポイントや、前後の時系列情報に基づいて補完する等である。しかし、補完された情報の信頼性は限定的であり、不十分である場合もある。一方、我々の提案する特徴量は欠損情報の補完を必要とせず、さらに得られた特徴量の各次元が、対応する元の関節点が欠損していたという情報を保持している。この特性により、欠損している部位は無視し、見えている情報のみを用いて類似判定を行うことで、高精度な検索が実現できる。

では、具体的な特徴量算出方法を説明する。我々は、画面内における垂直方向と実世界における重力方向が平行になるようにほぼすべてのカメラが設置されていること、その状況下では画像から推定した姿勢情報の内、y 軸成分はカメラの水平角や人の向きの変化に対して概ね頑健であることに着目した。そしてこの値は、カメラの俯角や焦点距離と連動して変化する人物の画像中における身長（単位：画素）を用いて正規化することで、画角に対して不変な特徴量となる。従って、提案する特徴量は以下の式で与えられる。

$$f_i = \frac{pose\_y_i - core\_y}{p\_height},$$

ここで、$pose\_y_i$ は 2 次元姿勢情報のうち i 番目のキーポイントの y 成分であり、$core\_y$ は基準点となる首の y 成分、そして p_height は推定された画像内における人物の高さ（単位：画素）をそれぞれ示している。仮に i 番目のキーポイント情報が欠損した場合は、f_i は -1 となる。p_height も 2 次元姿勢情報を用

いて以下の方法で推定される。

2 次元姿勢情報は、重なりや隠れ等によって頻繁に欠損するため、身長推定手法もキーポイントの欠損に対して頑健である必要がある。この要件を満たすために、我々は人体の各関節間距離と身長との長さの比率を表した標準的な人体モデル (図 3) を定義し、2 次元姿勢情報の各キーポイント間距離から、本人体モデルを用いて身長を算出する手法を提案した。本手法の特性上、カメラの光軸に対して平行なボーン（キーポイントをつなぐリンク）から算出した身長は実際の身長より低くなるが、実際より高く推定されることはない。よって、各ボーンから推定された身長の内、上位の値を抽出して平均をとり、推定身長とした。本手法では、欠損しているボーンの情報は使われないため、欠損に対して頑健な推定が可能となる。

特徴量ベクトルの各次元が、元の姿勢ベクトルの各次元と 1 対 1 で対応している点、さらに元のキーポイントが欠損しているという情報をそのまま保持している点が、本特徴量の大きな特徴である。2 次元姿勢情報と、そこから変換された特徴量のペアを図 3 に示す。特徴量可視化の際には、x 成分は固定値とし、y 成分のみを特徴量の値に応じて変化させた。図から、提案した特徴量は画角や背景、人の向き等にロバストであり、かつ姿勢間の類似度を計量可能なことが確認できる。

3.2 姿勢間照合関数

特徴量間の距離関数としては、以下の式で示す一般的な $L1$ 距離を用いた。

$$L(f_a, f_b) = \frac{1}{n}\sum_{i=1}^{n}|f_{a,i} - f_{b,i}|,$$

ここで f_i は i 次元目の特徴量を示し、 n は特徴量次元数（＝キーポイント数）である。部分的にキーポイントが欠損している場合は、見えているキーポイントの情報のみを用いて照合を行う方がうまくいく場合がある。よって、キーポイントの欠損がある場合は、以下の距離関数を用いた。

$$L\_lack(f_a, f_b) = \frac{\sum_{i=1}^{n}|f_{a,i} - f_{b,i}| \times l_i}{\sum_{i=1}^{n} l_i},$$

$$l_i = \begin{cases} 1\ (f_{a,i} \neq -1\ and\ f_{b,i} \neq -1) \\ 0\ (f_{a,i} = -1\ or\ f_{b,i} = -1). \end{cases}$$

さらに、同一の行動であるにも関わらず個人ごとに異なる姿勢である場合において高精度に検索を行うためには、特定の部位に重みづけをして照合を行うことが好ましい。重みを考慮する場合は、以下の距離関数を用いた。

$$L\_lack\_weight(f_a, f_b) = \frac{\sum_{i=1}^{n}|f_{a,i} - f_{b,i}| \times W_i}{\sum_{i=1}^{n} W_i},$$

$$W_i = \begin{cases} w_i\ (f_{a,i} \neq -1\ and\ f_{b,i} \neq -1) \\ 0\ \ (f_{a,i} = -1\ or\ f_{b,i} = -1), \end{cases}$$

上述した式において、w_i は i 次元目の特徴量に対する重みである。

しかしながら、ユーザーが手動で各キーポイントへの重みづけを行うことは難しい。そこで我々は、クエリ姿勢に応じて自動で重みを設定する手法を提案する。まず、図 3 左に示すような、直立姿勢を参照姿勢として定義し、クエリ姿勢と参照姿勢の特徴量の距離に基づいて重みを設定した。このアイディアに基づき、w_i を以下の式で定義した。

$$w_i = \frac{|f_{a,i} - f_{r,i}|}{sum\_length_{r,i}} + 1,$$

ここで f_r は参照姿勢の特徴量を、$sum\_length_{r,i}$ は参照姿勢における i 番目のキーポイントと首のキーポイントとの距離の合計を表す。例えば、$sum\_length_{r,Rhand}$ は首-右肩 (0.1)、右肩-右肘 (0.15)、右肘-右手 (0.15) の合計で 0.4 となる。この手法により、それぞれのキーポイントに対する重みはおよそ 1 ～ 3 に正規化される。

3.3 シーン間照合関数

前述したように、姿勢間照合関数によってフレーム間距離が計算できる。次にこのフレーム間距離を用いて一般的な DTW アルゴリズムを適用し、フレーム間距離の和が最小になるように 2 つのシーンのフレーム対応付けを行った。そして最後に、フレーム間距離の和をシーン間距離として算出した。

4 実験と評価

我々の提案する特徴量と照合関数の有効性を評価するため、様々な画角で撮影された 2 次元姿勢データを用いた類似姿勢検索と、実映像を用いた類似行動検索において定量的な評価を行った。

4.1 データセット

類似姿勢検索の評価には、正解付けされた多視点 2 次元姿勢データセットを、類似行動検索の評価には汎用的な行動認識向けの公開データセットである UT-kinect [26] を用いた。

モーションキャプチャーデータセット

我々は様々な行動を行う人物をモーションキャプチャーシステムを活用して撮影し、キーポイントの 3 次元位置が正解付けされた 3 次元姿勢データセットを得た。次にこのデータセットを、様々な画角でカメラに投影した状況をシミュレートし、正解付けされた 2 次元姿勢データセットを得た。

本データセットは、椅子に座る、地面に座る、地面に寝る、右手を挙げる、左手を挙げる、両手を挙げる、右足を挙げる、左足を挙げる、という 8 つの行動を 10 人がそれぞれ実施したデータからなる。

得られた 3 次元姿勢データを、図 4 に示すように人物を中心とした半径 5m の半球上にカメラを設置した状況をシミュレートして、2 次元姿勢に投影した。その時、カメラの水平角 (pan) は 0° ～ 350° 、俯角 (tilt) は 0° ～ 40° の範囲で、それぞれ 10° 刻みに変化させた。これにより、一つの 3 次元姿勢から 180 パターンの 2 次元姿勢を生成した。

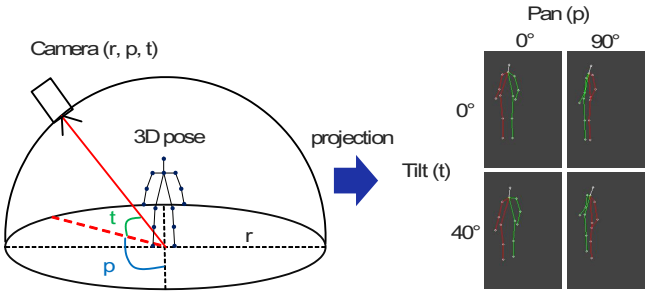

図 4　3 次元姿勢から 2 次元姿勢への投影方法

UT-kinect

このデータセットは、事前にトリミングされた 200 のビデオからなり、10 種類の行動（歩く、物を運ぶ、押す、引く、投げる、拾う、両手を振る、椅子に座る、椅子から立つ、拍手）を 10 人がそれぞれ 2 回ずつ行っている様子が撮影されている。

カメラの位置、向きは固定だが、それぞれの人物が行動を行う際の向きは制限されておらず、多様性がある。我々は本データセットに対し、Bottom-up 型の姿勢推定エンジンである NeoPose [35] を用いて 2 次元姿勢推定を行った。姿勢情報と動作を行っている人物との紐づけは、推定された姿勢の外接矩形と正解づけされた人物矩形との IoU (Intersection over Union) によって行った。

4.2　姿勢に基づく行動検索

モーションキャプチャーデータセットを用いて、特定のカメラ (pan = 0° , tilt = 0°) で撮影した姿勢情報をクエリとして、特徴空間において近傍探索を行った（この時、クエリは除外した）。初めに、距離関数には一般的な $L1$ 距離を用いた。検索を行う際、データセット全体から目的に応じて以下の 3 つのサブセット (a), (b), (c) を再構築し、多角的に提案手法の有効性を検証した。

(a) クエリと同一人物のデータのみを含む

(b) クエリと同一画角（すなわち pan = 0° , tilt = 0° ）のデータのみを含む

(c) 全てのデータを含む

データセット内すべての人物、すべての行動の姿勢をクエリとして用いて検索を行い、結果を平均して性能を評価した。

4.3　評 価 指 標

評価指標として、precision = 90%時の recall を評価した。recall は、データセット内における正解データの内、検索結果に含まれる正解データの割合である。なお、本実験においては、クエリ姿勢と検索結果の姿勢が同一の行動ラベルを有していた場合を正解と判定した。

4.4　比 較 手 法

姿勢全体同士での類似判定が可能な特徴量導出手法として、外接矩形の高さを正規化した 2 次元姿勢情報 (pose [23]) を特徴量として用いる手法と、Human3.6M データセットを用いて学習した、2 次元姿勢情報を画角不変な特徴量 (pr-vipe [24]) に変換する手法（我々のデータセットに対する追加学習なし）の 2 つを比較手法に用いた。まずは 3 種類の特徴量すべてに対して、同一の距離関数である $L1$ 距離を用いて精度を評価した。

4.5　姿勢情報の欠損なしの場合

表 1 に recall@precision = 90%を示す。また、サブセット a における我々の手法と pose の top-10 の結果を、サブセット b における我々の手法と pr-vipe の top-10 の結果を、それぞれ図 5 と図 6 に示す。

図 5 から、我々の特徴量を用いた場合の検索結果 top-10 には様々な画角の正解姿勢が含まれており、表 1 から pose よりも高い recall を有していることが分かる。これは、我々の提案特徴量が画角不変性を持つことにより、画角不変性を持たない pose よりも recall が高かったことを示唆している。しかしながら、精度は pr-vipe に劣る結果となった。

一方サブセット b では、我々の手法が他の 2 つの手法より優れる精度を達成した。この結果は、我々の特徴量が同一行動であるにも関わらず個人ごとに異なる姿勢という状況の検索に適していることを示唆している。事実、我々の手法を用いた場合の top-10 の結果には、異なる人物の正解データが多く含まれていることがわかる。

さらに、サブセット c でも、我々の手法が最も優れる結果となり、画角変化と姿勢の個人差への頑健性の両方が高い水準であることが検証できた。

表 1　モーションキャプチャーデータを用いた精度比較（欠損なし）

	recall (%) @ precision = 90%		
サブセット	a	b	c
pose [23]	31.5	84.0	19.6
pr-vipe [24]	**97.6**	49.7	49.0
ours	87.7	**91.0**	**72.4**

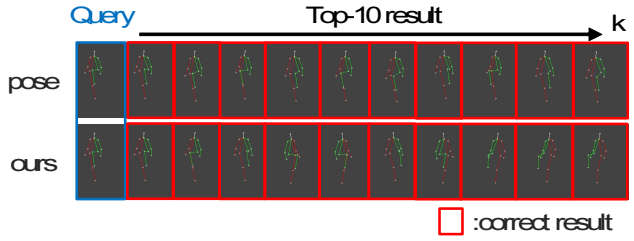

図 5　サブセット a を用いた時の、pose と我々の手法における、Top-10 結果の比較。クエリ行動は左足を挙げる。

4.6　姿勢情報の欠損ありの場合

キーポイントに欠損がある状況に対する頑健性を評価するため、クエリとデータセット内の姿勢情報両方に欠損がある状況を想定した評価を行った。具体的には、それぞれの姿勢情報から、実環境において頻繁に欠損が起こる人体の末端（右手首、左手首、右足首、左足首、頭部）のキーポントをランダムに一つ欠損させた。この時、クエリ姿勢と、照合されるデータセット

内の姿勢に対して、同一の箇所が欠損しないようにコントロールした。また、従来技術では照合のために欠損したキーポイントの補完が必要であることから、欠損キーポイントは一律に人物矩形の中心に配置することで補完した。

本実験における距離関数には、$L1$ 距離を使う場合に加えて、欠損キーポイントを無視した照合を行うために、*L_lack* (3.2 参照) を用いた。本距離関数を用いた結果を ours_lack とする。表 2 に示す結果からわかるように、欠損情報を含めて全身の情報を用いて照合を行った場合（$L1$ 距離を用いた場合）では、特徴量の種類に依らず精度が低かった。欠損情報の補完精度を上げることで検索精度の改善も見込めるが、補完の信頼性には限界がある。一方、ours_lack は最も優れた精度を示し、欠損がない場合と比較してわずかな精度低下にとどまっている。このことから、欠損していない情報のみに基づく照合方法の方が優れていると言える。

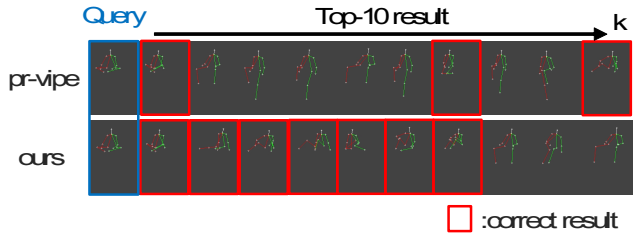

図 6　サブセット b を用いた時の、pr-vipe と我々の手法における、Top-10 結果の比較。クエリ行動は地面に座る。

表 2　モーションキャプチャーデータを用いた精度比較（欠損あり）

	Recall (%) @ precision = 90%		
サブセット	a	b	c
pose [23]	17.8	63.2	11.8
pr-vipe [24]	43.0	14.7	9.7
ours	18.4	12.9	5.8
ours_lack	**78.2**	**77.8**	**48.6**

4.7　時系列姿勢情報に基づく行動検索

評価には UT-Kinect データセット [26] を用いた。比較手法には、前回の実験同様 pose [23] と pr-vipe [24] を用いた。姿勢間の照合関数には、$L1$ 距離を用いて、我々の特徴量に対しては $L1$ 距離に加えて、欠損キーポイントを無視するために *L_lack*（結果は ours_lack）を、さらに特徴的なキーポイントへの重みづけを行うために *L_lack_weight*（結果は ours_lack_weight）を用いた。

2 つのシーン間の照合を行う際、まず DTW アルゴリズム [40] を用いてフレーム間距離の和が最小になるように対応付けを行い、フレーム間距離の和をシーン間距離とした。

表 3、表 4 に、top-k (k = 1, 5, 10) における precision と recall をまとめる。本結果は、我々の提案特徴量自身は既存手法である pr-vipe [24] に及ばないが、欠損等に柔軟に対応できる照合関数を用いることで、precision、recall 共に高い検索が可能であることを示している。さらに ours_weight_lack が最も高い精度を示しており、特徴的なキーポイントに重みをつけて照合する手法の有効性を検証できた。

表 3　UT-kinect データセット [26] を用いた行動検索における precision の比較

	Average precision (%) @ top-k		
k	1	5	10
pose [23]	86.2	67.6	54.0
pr-vipe [24]	91.9	78.9	67.3
ours	91.7	75.2	63.4
ours_lack	95.5	85.3	78.3
ours_lack_weight	**96.0**	**86.3**	**79.6**

表 4　UT-kinect データセット [26] を用いた行動検索における recall の比較

	Average recall (%) @ top-*k*		
k	1	5	10
pose [23]	3.7	15.4	24.9
pr-vipe [24]	4.2	18.8	32.2
ours	4.2	17.7	30.0
ours_lack	4.6	20.7	38.3
ours_lack_weight	**4.6**	**21.0**	**39.0**

5　結　　論

本論文では、新しい画角不変な姿勢特徴量と、柔軟な照合関数によって実現した高精度な行動検索手法を紹介した。本特徴量は 2 次元姿勢情報を入力として変換され、特徴量ベクトルの各次元が元の姿勢情報の各次元と 1 対 1 で対応している。照合関数は、目的に応じて欠損情報を無視したり、特徴的な部位に重みづけをしたり、容易かつ柔軟に対応できる。本提案手法は、正解付けされた 2 次元姿勢データセットを用いた行動検索と、実映像を用いた行動検索の両方において、既存の最新手法を上回る精度を達成した。

文　　献

[1] Tomi Räty. Survey on contemporary remote surveillance systems for public safety. *IEEE Trans. Syst. Man Cybern. Part C*, 40(5):493–515, 2010.

[2] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):288–303, 2010.

[3] Sreemananth Sadanand and Jason J. Corso. Action bank: A high-level representation of activity in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1234–1241. IEEE Computer Society, 2012.

[4] Yigithan Dedeoglu, B. Ugur Töreyin, Ugur Güdükbay, and A. Enis Çetin. Silhouette-based method for object classification and human action recognition in video. In Thomas S. Huang, Nicu Sebe, Michael S. Lew, Vladimir Pavlovic, Mathias Kölsch, Aphrodite Galata, and Branislav Kisacanin, editors, *Computer Vision in Human-Computer Interaction, ECCV 2006 Workshop on HCI, Graz, Austria, May 13, 2006, Proceedings*, volume 3979 of *Lecture Notes*

in Computer Science, pages 64–77. Springer, 2006.

[5] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008.

[6] Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition from video using feature covariance matrices. *IEEE Trans. Image Process.*, 22(6):2479–2494, 2013.

[7] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. Action recognition by learning deep multi-granular spatio-temporal video representation. In John R. Kender, John R. Smith, Jiebo Luo, Susanne Boll, and Winston H. Hsu, editors, *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*, pages 159–166. ACM, 2016.

[8] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Spatiotemporal pyramid network for video action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2097–2106. IEEE Computer Society, 2017.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 568–576, 2014.

[10] Mahdyar Ravanbakhsh, Hossein Mousavi, Mohammad Rastegari, Vittorio Murino, and Larry S. Davis. Action recognition with image based CNN features. *CoRR*, abs/1512.03980, 2015.

[11] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. Simple yet efficient real-time pose-based action recognition. In *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, Auckland, New Zealand, October 27-30, 2019*, pages 581–588. IEEE, 2019.

[12] Federico Angelini, Zeyu Fu, Yang Long, Ling Shao, and Syed Mohsen Naqvi. Actionxpose: A novel 2d multi-view pose-based algorithm for real-time human action recognition. *CoRR*, abs/1810.12126, 2018.

[13] P V.V. Kishore, P Siva Kameswari, K Niharika, M Tanuja, M Bindu, D Anil Kumar, E Kiran Kumar, and M Teja Kiran. Spatial joint features for 3d human skeletal action recognition system using spatial graph kernels. *International Journal of Engineering & Technology*, 7(1.1):489–493, 2017.

[14] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7912–7921. Computer Vision Foundation / IEEE, 2019.

[15] Kalpit C. Thakkar and P. J. Narayanan. Part-based graph convolutional network for action recognition. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 270. BMVA Press, 2018.

[16] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Ahmed Sohel, and Farid Boussaïd. A new representation of skeleton sequences for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4570–4579. IEEE Computer Society, 2017.

[17] Eamonn J. Keogh, Themis Palpanas, Victor B. Zordan, Dimitrios Gunopulos, and Marc Cardle. Indexing large human-motion databases. In Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer, editors, *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 - September 3 2004*, pages 780–791. Morgan Kaufmann, 2004.

[18] X. Zhao, Myung Geol Choi, and Taku Komura. Character-object interaction retrieval using the interaction bisector surface. *Comput. Graph. Forum*, 36(2):119–129, 2017.

[19] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, 2004.

[20] Jun Tang, Ling Shao, and Xiantong Zhen. Human action retrieval via efficient feature matching. In *10th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2013, Krakow, Poland, August 27-30, 2013*, pages 306–311. IEEE Computer Society, 2013.

[21] Arridhana Ciptadi, Matthew S. Goodwin, and James M. Rehg. Movement pattern histogram for action recognition and retrieval. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II*, volume 8690 of *Lecture Notes in Computer Science*, pages 695–710. Springer, 2014.

[22] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. Pose embeddings: A deep architecture for learning to match human poses. *CoRR*, abs/1507.00302, 2015.

[23] Nataraj Jammalamadaka, Andrew Zisserman, Marcin Eichner, Vittorio Ferrari, and C. V. Jawahar. Video retrieval by mimicking poses. In Horace Ho-Shing Ip and Yong Rui, editors, *International Conference on Multimedia Retrieval, ICMR '12, Hong Kong, China, June 5-8, 2012*, page 34. ACM, 2012.

[24] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 53–70. Springer, 2020.

[25] Stuart James, Manuel J. Fonseca, and John P. Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In Mohan S. Kankanhalli, Stefan M. Rüger, R. Manmatha, Joemon M. Jose, and Keith van Rijsbergen, editors, *International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014*, page 313. ACM, 2014.

[26] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, pages 20–27. IEEE Computer Society, 2012.

[27] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7024–7033. Computer Vision Foundation / IEEE Computer Society, 2018.

[28] Girum G. Demisse, Konstantinos Papadopoulos, Djamila Aouada, and Björn E. Ottersten. Pose encoding for robust skeleton-based action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 188–194. Computer Vision Foundation / IEEE Computer Society, 2018.

[29] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4620–4628. Computer Vision Foundation / IEEE, 2019.

[30] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 53. BMVA Press, 2018.

[31] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3980–

3989. IEEE Computer Society, 2017.

[32] Ankan Bansal. *Detecting and Recognizing Humans, Objects, and their Interactions*. PhD thesis, University of Maryland, College Park, MD, USA, 2020.

[33] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In Gang Hua and Hervé Jégou, editors, *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, volume 9914 of *Lecture Notes in Computer Science*, pages 627–642, 2016.

[34] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 472–487. Springer, 2018.

[35] Yadong Pan, Ryo Kawai, Noboru Yoshida, Hiroo Ikeda, and Shoji Nishimura. Training physical and geometrical mid-points for multi-person pose estimation and human detection under congestion and low resolution. *SN Comput. Sci.*, 1(4):208, 2020.

[36] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Real-time multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1302–1310. IEEE Computer Society, 2017.

[37] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2277–2287, 2017.

[38] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1293–1301. IEEE Computer Society, 2017.

[39] Georgia Gkioxari, Bharath Hariharan, Ross B. Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3582–3589. IEEE Computer Society, 2014.

[40] Petr Mandl and M. Rosario Romera Ayllón. On adaptive control of markov processes. *Kybernetika*, 23(2):89–103, 1987.