# 骨格情報の推移に着目した特徴量に基づく スポーツ動画における細粒度な動作分類

佐藤荘一朗† 青野 雅樹††

† 豊橋技術科学大学大学院工学研究科 情報・知能工学専攻 〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系 〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
E-mail: †s-sato@kde.cs.tut.ac.jp, ††masaki.aono.ss@tut.jp

**あらまし** 近年，深層学習を用いて動画内に映っている人物の行動認識を行う研究が盛んに行われている．今日に至るまで提案された動画認識モデルは「大まかな動作分類」に焦点を当てており，体操競技における「1 回転ひねり」と「2 回転ひねり」の動作分類を行う場面を代表とする「細かい動作分類」に焦点を当てていないのが現状である．そこで本研究では，スポーツ動画における細粒度な動作分類の精度向上を目指した手法の提案を目的とする．具体的には，動画内における動作主体の骨格情報の推移に着目した Pose 特徴量を抽出する手法及び RGB フレーム，Optical Flow を代表とする既存の動作特徴量に加えて抽出した Pose 特徴量を入力とする学習モデルを提案する．評価実験には FineGym データセットを使用し，既存の動作特徴量を使用したベースラインと提案手法の性能比較を行った．その結果，提案手法はベースラインの精度を上回ることが確認された．

**キーワード** 動画，深層学習，細粒度動作分類，骨格情報，姿勢推定

## 1 はじめに

近年，深層学習モデルを用いて動画内に映っている人物の行動認識を行う研究が盛んに行われている．このような研究は，スポーツ選手の行動分析及び教育用動画の作成など，様々なところで応用できると考えられる．今日に至るまで，Two Stream ConvNets [1], C3D [2], I3D [3] といった畳み込みニューラルネットワーク (CNN) による学習方法が導入された動画認識向け深層学習モデルが提案されてきた．このような最先端の動画認識モデルの開発は，UCF101 [4]，Kinetics [3] などの動画データセットを用いた評価実験及びベンチマークによって推進されてきた．これらの動画データセットは，スポーツの種類 (例：バスケットボール，サッカー) あるいは人間の行動の種類 (例：ピアノを弾いている，自転車に乗っている) などに基づいた分類ラベルが付与されている特徴を持っている．しかし，既存の動画認識モデル及びその開発の基盤となった動画データセットは,「大まかな動作分類」(図 1) に焦点を当てており，体操競技における「1 回転ひねり」と「2 回転ひねり」の動作分類を行う場面を代表とする「細かい動作分類」(図 2) に焦点を当てていない.「細かい動作分類」を考慮した動画認識に関する研究は，スポーツ選手が披露した技を自動で判別し，その結果に基づいた自動採点を行う場面などに応用できると考えられるが，現時点では研究があまり進んでいないのが現状である．そこで本研究では，スポーツ動画における細粒度な動作分類の精度向上を目指した手法の提案を目的とする．

多くの従来手法では，動画における色に関する情報に基づいた RGB フレーム及び動きに関する情報に基づいた Optical Flow を使用する場合が一般的である．これに対し，我々はスポーツ動画内における動作主体を抽出し，その骨格情報の推移の活用に着目した．従来手法で使用されている動作特徴量とは違う判断材料の導入により，細粒度な動作分類の精度向上が期待できる．動画から人間の骨格情報を取得する方法として，姿勢推定モデルの活用が挙げられる．姿勢推定モデルは，画像内もしくは動画内に映っている全ての人間に対する骨格情報を取得する特徴を持っている．スポーツ動画には選手，観戦者と


図 1 実例レベルの動作分類の例 (UCF101 [4] より引用)


図 2 細粒度な動作分類の例 (体操競技における披露した技の分類 [5])

いった複数の人物が映っている場合が多いが，動作分類の対象は選手のような動作主体のみである．そのため，動作分類に併せて動作主体の抽出も行う必要があると考えられるが，これ自体も困難な問題である．そこで本研究では，姿勢推定モデルの推論結果を活用し，動作主体の Pose 特徴量を抽出する方法及び RGB フレーム，Optical Flow を代表とする既存の動作特徴量に動作主体の Pose 特徴量を付加した独自の学習モデルによる細粒度な動作分類手法を提案する．既存の動作特徴量に動作主体の Pose 特徴量を付加する背景として，動画から得られる複数の特徴量を同時に入力する動画認識モデルを使用した場合，1 種類の特徴量を入力する場合よりも動作分類の精度が高いという結果が示されている [1], [3]．この知見を踏まえ，抽出した Pose 特徴量と既存の動作特徴量を同時に入力する深層学習モデルの構成が細粒度な動作分類に対して有効ではないかと考えた．既存の動作特徴量を使用したベースラインと動作主体を抽出し，付加した Pose 特徴量を導入した提案手法の性能比較を行うための評価実験を実施し，提案手法の有効性を検証する．

## 2 関 連 研 究

本節では，関連研究について述べる．最初に，深層学習を使用した動画認識に関する研究として CNN を使用した動画認識手法及び動画データセットについて述べる．その後，深層学習を使用した姿勢推定手法に関する研究について述べる．

### 2.1 CNN を使用した動画認識

CNN を使用した動画認識手法として，2D CNN を使用した手法及び 3D CNN を使用した手法が挙げられる．

2D CNN を使用した手法は，RGB フレームを CNN へ入力し，空間情報の畳み込みを行うことで入力動画に関する特徴量を抽出するというものである．この手法は，画像認識と同様の方法であるが，RGB フレームの時間的な変化に関する情報を考慮した学習を行うことができない問題点が存在する．そこで Simonyan らは，RGB フレームの空間情報に関する特徴量を抽出するネットワーク (Spatial Stream ConvNet) と，RGB フレームの時間的な変化に関する特徴量を抽出するネットワーク (Temporal Stream ConvNet) を組み合わせた Two Stream ConvNets [1] を提案した．Temporal Stream ConvNet に入力するデータは，Optical Flow と呼ばれる画像間の変化を表現した画像を使用している．複数枚の RGB フレームにおける $x$ 方向及び $y$ 方向の変化量を画像の各チャンネル成分として扱うことで，1 枚の画像という形で Optical Flow を生成する．Two Stream ConvNets を使用した学習及び推論では，Spatial stream ConvNet へ RGB フレームを入力し，それに対応する Optical Flow を Temporal Stream ConvNet へ入力する．これにより，2 種類の分類結果が出力されるが，それらを統合することで最終的な分類結果が得られる．

3D CNN を使用した手法は，2D CNN で行われていた 2 次元畳み込み処理を時間方向に拡張した 3 次元畳み込み処理を行うことで，入力動画に関する特徴量を抽出するというものである．畳み込み処理を 3 次元に拡張することで，空間方向の情報と時間方向の情報を考慮した学習及び推論を可能にしている．3D CNN の先駆け的な手法として，Tran らは画像認識用に提案された VGG-11 の構造をベースとした C3D [2] を提案した．C3D では，VGG-11 の 2 次元畳み込み層を 3 次元畳み込み層に置き換えることで，動画認識を可能としたモデルを実現させている．Carreira らは，Inception-v1 の構造をベースとした I3D [3] を提案した．I3D では，Inception-v1 の 2 次元畳み込み層を 3 次元畳み込み層に置き換えている．これは C3D と同様のアイデアであるが，C3D よりもモデルパラメータ数を抑えつつ，深いモデル構造を実現させている．これにより，ImageNet などの大規模な画像データセットによる事前学習を行ったパラメータの活用による高精度な動画認識を実現させている．

### 2.2 動画データセット

動画認識の研究において使用されてきた代表的な動画データセットとして，UCF101 [4]，Kinetics [3] などが挙げられる．特に Kinetics は，306,245 本の動画及び人間の行動に関する 400 種類のラベルで構成されている大規模な動画データセットで，動画認識モデルの性能評価，事前学習済みモデルの作成などの用途で使用されている [3]．

一方，スポーツ選手が披露した技といった細粒度なラベルが付与されている動画データセットが存在する．Li らは，48 クラスの飛び込み競技動画データセット Diving48 [6] を作成した．Diving48 では，飛び込み競技において選手が披露した技の名前に基づいたラベルが付与されている．Martin らは，21 クラスの卓球動画データセット TT-Stroke21 [7] を作成した．TT-Stroke21 では，卓球のサーブやレシーブといったストロークの違いに基づいたラベルが付与されている．Shao らは，体操競技動画データセット FineGym [5] を作成した．FineGym は，体操競技において選手が披露した技の名前に基づいたラベルが付与されている．また，クラス分類の粒度に応じて 3 種類のラベルが付与されている特徴を持っている．

### 2.3 姿勢推定モデル

近年，深層学習を用いた姿勢推定を行う様々な手法が提案されている．姿勢推定を行う手法はトップダウン・アプローチ及びボトムアップ・アプローチの 2 種類に大別される．トップダウン・アプローチとは，最初に画像内における人物全体の範囲を検出した後，その範囲内において関節の位置を推定する手法である．ボトムアップ・アプローチとは，最初に画像内における関節の位置を推定した後，関節の推定位置同士をグルーピングすることで姿勢推定を行う手法である．以下では，トップダウン・アプローチによる姿勢推定手法について述べる．

Toshev らは，CNN モデルの 1 つである AlexNet を使用し，入力画像から人間の関節の位置を回帰により推定する DeepPose [8] を提案した．Chen らは，Feature Pyramid Network をベースとした GlobalNet によるマルチスケールな特徴量抽出を行い，抽出された特徴量を RefineNet により統合することで骨格座標の推定を行う Cascaded Pyramid Network (CPN) [9]

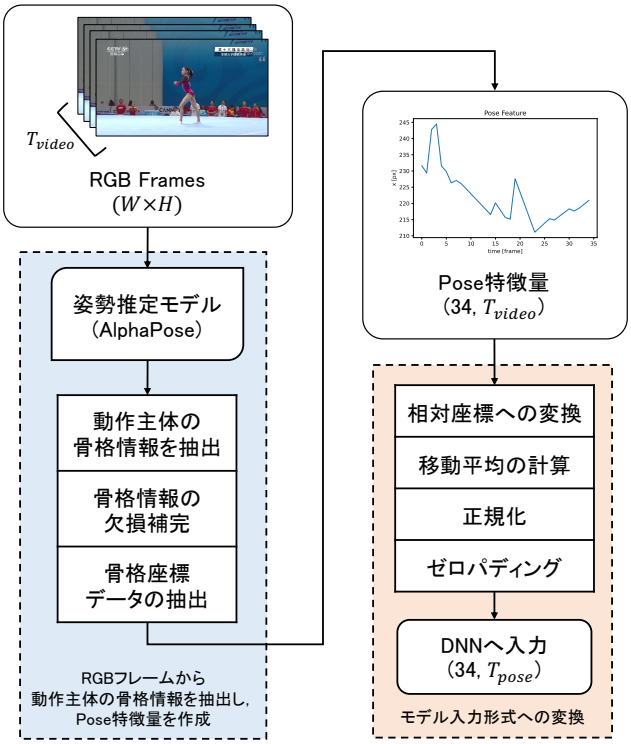



図 3　動作主体の Pose 特徴量抽出の流れ

を提案した．Fang らは，人間が密集しているような混雑シーン画像に対して正確な姿勢推定を行う AlphaPose [10] を提案した．AlphaPose では，最初に入力画像に対して物体検出モデルを使用した人物領域検出を行う．その後，検出された人物領域 (Bounding Box) に対して SPPE(単一人物姿勢推定) を使用することで関節の位置を推定する．

## 3　動作主体の Pose 特徴量抽出

本節では，動作主体の骨格情報の推移に着目した Pose 特徴量を抽出する方法について述べる．動作主体の Pose 特徴量を抽出する流れを図 3 に示す．ここで，入力動画の長さを $T_{video}$，学習モデルへ入力する際の時間方向の次元数を $T_{pose}$ と定義する．また，Pose 特徴量抽出における入力データは $W \times H$ の解像度を持つ $T_{video}$ 枚の RGB フレームである．

$T_{video}$ 枚の RGB フレームに対応する Pose 特徴量抽出は，図 3 における青色の部分で示す流れに基づいて行う．最初に，$T_{video}$ 枚の RGB フレームを AlphaPose [10] へ入力し，RGB フレーム内に映っている人物に対する骨格情報を取得する．骨格情報を取得した後，Bounding Box の位置，大きさ，面積に基づいた動作主体の骨格情報を抽出する．動作主体の骨格情報を抽出した後，認識漏れ等により骨格情報の推移が途切れている場合の考慮を目的として，骨格情報の欠損補完を行う．最後に，骨格情報から 17 種類の骨格座標データを抽出する．以上の操作により抽出された $(34, T_{video})$ の次元数を持つ特徴量を Pose 特徴量として扱う．

Pose 特徴量を学習モデルへ入力する際は，図 3 における橙色の部分で示すように，相対座標への変換，移動平均の計算，正規化，ゼロパディングを行う．学習及び推論で使用する動画

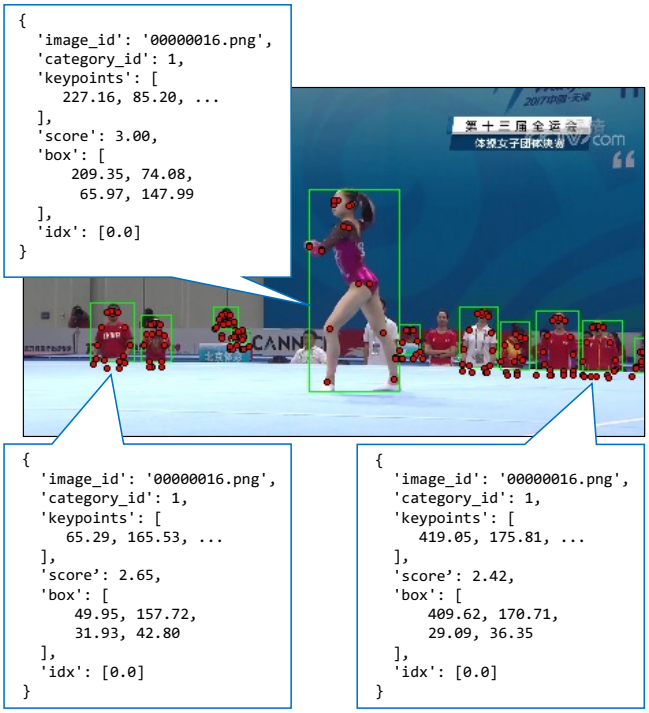



図 4　AlphaPose 推論結果の一例 (選手の他に観戦者も検出している)

は，20 フレーム程度の短い動画から数百フレームにも及ぶ長い動画が混在しているため，Pose 特徴量を学習モデルへ入力させる際に時間方向の次元数を揃える必要がある．そこで，学習モデルへ入力する直前にゼロパディングを行うことで，Pose 特徴量の時間方向の次元数を使用する動画の中で 1 番長いものに揃える．これにより，学習モデルに入力する際の Pose 特徴量の次元数は $(34, T_{pose})$ となる．

### 3.1　AlphaPose による推論

本研究では，トップダウン・アプローチに基づいた姿勢推定モデルの 1 つである AlphaPose [10] を使用する．AlphaPose は，RGB 画像を入力すると画像内に映っている人物の領域及び骨格座標を推定することができる．推論結果の例を図 4 に示す．AlphaPose から得られる推論結果は，入力画像の名前，17 種類の骨格座標データ，Bounding Box(人物領域) に関するデータ，推論結果に関する信頼スコアの 4 種類である．人物領域に関するデータは，画像内における Bounding Box の位置及び大きさに関するデータが取得される．骨格座標データは，画像内に映っている人物の手首，肘，肩，膝といった合計 17 種類の骨格座標が取得される．

### 3.2　動作主体の骨格情報抽出

RGB フレーム内に映っている人物に対する骨格情報を取得した後，動作主体の骨格情報を抽出する．AlphaPose は 1 枚の RGB フレームに対して複数人検出した場合，検出した人数分の推論結果を取得する特徴を持っている．図 4 で示した推論結果の例では，入力画像の名前は共通しているものの，17 種類の骨格座標データ等は異なるものが複数存在している状態である．Pose 特徴量作成では選手のような動作主体の推論結果が必要であるが，それ以外の人物 (例：審判，観客など) の推論結果は

図 5　骨格情報の推移が途切れている例 (赤枠で囲った RGB フレームでは，Bounding Box 及び骨格座標データのプロットが抜けている．)

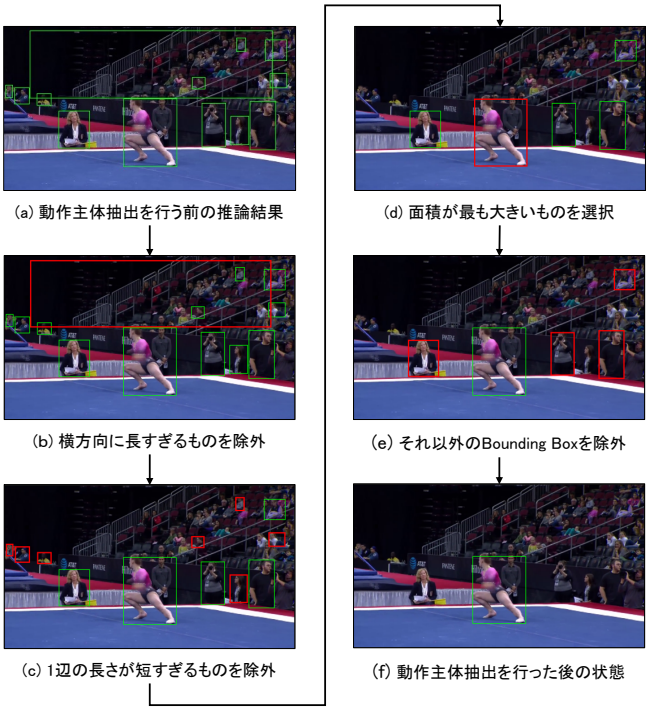



図 6　動作主体の骨格情報を抽出する流れ

不要である．以上の理由から，動作主体の骨格情報を抽出する．

具体的には，RGB フレーム内で検出された Bounding Box の位置，大きさ，面積に基づいた抽出を行う．動作主体の骨格情報を抽出する流れを図 6 に示す．最初に，横方向に長すぎる Bounding Box 及び 1 辺の長さが短すぎる Bounding Box を除外する．その後，残った Bounding Box から面積が最も大きいものを選択し，それ以外の Bounding Box を除外する．事前の観察から，人物領域の面積が最も大きければ，その人物に焦点が当たっていると考え導入している．最終的に残った 1 つの Bounding Box に対応する骨格情報を Pose 特徴量作成に使用する．以上の操作を入力した全ての RGB フレームに対して行い，RGB フレームと動作主体の骨格情報が 1 対 1 で対応するようにしている．

### 3.3　骨格情報の欠損補完

動作主体の骨格情報を抽出した後，骨格情報の欠損補完を行う．AlphaPose に限らず，現状の姿勢推定モデルは人間が激しい動きをした際に認識漏れを起こすことが多い傾向が見られる．特に，体操競技では短い時間の中で身体を 2〜3 回転させる技を披露する場面が多く，その際に認識漏れを起こすことが多い．その例を図 5 に示す．図 5 では，RGB フレームから動作主体の骨格情報を抽出した結果をプロットしている．赤枠で囲った RGB フレームでは，Bounding Box 及び骨格座標データのプロットが抜けていることが確認できる．これは，赤枠で囲った RGB フレームにおいて，対応する動作主体の骨格情報が存在せず，骨格情報の推移が途中で途切れていることを意味している．以上の理由から，欠損区間の前後の推論結果を使用し，骨格情報の欠損補完を行う方法を導入している．

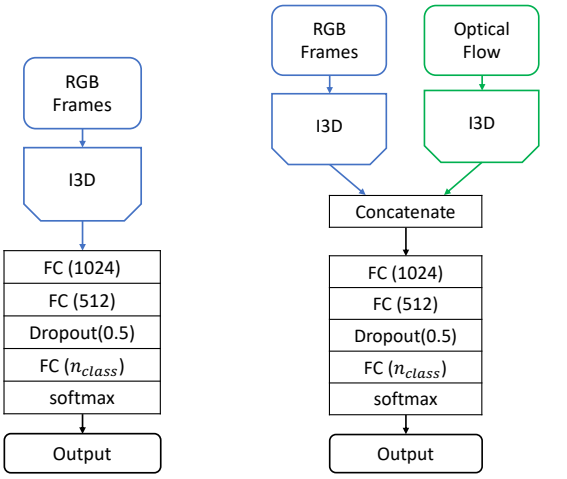
具体的には，欠損区間の直前の推論結果をそのまま欠損区間にも代入することで欠損補完を行う方法もしくは欠損前後の推論結果に基づいた線形補間を行うことで欠損補完を行う方法のいずれかを導入する．本研究では，前者を FPD(Fill by Previous Data)，後者を FLI(Fill by Linear Interpolation) と定義する．

## 4　Pose 特徴量の導入に基づいた学習モデル

本研究では，RGB フレーム，Optical Flow を代表とする既存の動作特徴量に加えて骨格情報の推移を表現した Pose 特徴量を入力とする学習モデルを導入する．具体的には，次に述べる 3 つの学習モデルを導入する．1 つ目は，RGB フレームと Pose 特徴量を入力とし，I3D 及び InceptionTime の出力を行列積演算により結合する学習モデルである．2 つ目は，RGB フレームと Pose 特徴量を入力とし，I3D 及び InceptionTime の出力を連結操作により結合する学習モデルである．3 つ目は，RGB フレーム，Optical Flow ならびに Pose 特徴量を入力とし，2 つの I3D 及び InceptionTime の出力を連結操作により結合する学習モデルである．これらの学習モデルは以降，提案モデル 1，提案モデル 2，提案モデル 3 と呼称する．本節では，提案モデルとの比較で使用するベースラインモデルについて述べた後，提案モデルの概要及び提案モデルに導入している I3D ならびに InceptionTime について述べる．

### 4.1　ベースラインモデル

本研究におけるベースラインモデルを図 7 に示す．ベースラインモデルは，RGB フレームのみを入力するモデル (図 7a) 及び RGB フレームと Optical Flow を入力するモデル (図 7b) の 2 種類を使用する．これらのベースラインモデルは以降，前者をベースラインモデル 1，後者をベースラインモデル 2 と呼称する．ベースラインモデル 1 は，RGB フレームを入力に使用し，I3D により 2048 次元の特徴量を抽出する．抽出された 2048 次元の特徴量は，2 層の FC 層，Dropout 層，出力層を経由した後，入力された RGB フレームに対する分類結果を出力する．ベースラインモデル 2 は，2 つの I3D を使用し，入力さ




(a) ベースラインモデル 1 (RGB フレームのみ入力)　(b) ベースラインモデル 2 (RGB フレームと Optical Flow を入力)

図 7　ベースラインモデル

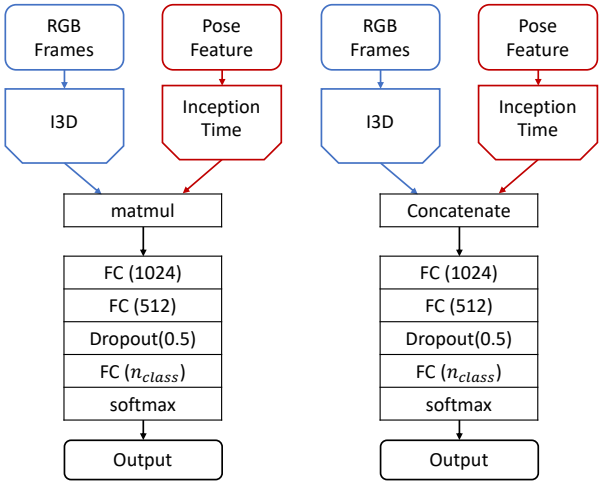



図 8　提案モデル 1




図 9　提案モデル 2

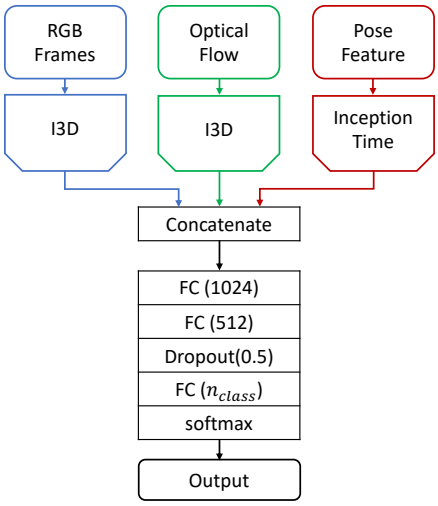



図 10　提案モデル 3

れた RGB フレーム，Optical Flow からそれぞれ 2048 次元の特徴量を抽出する．その後，抽出された 2 つの特徴量に対して連結操作を行い，4096 次元の結合特徴量を作成する．作成した結合特徴量は，2 層の FC 層，Dropout 層，出力層を経由した後，入力された RGB フレーム及び Optical Flow に対する分類結果を出力する．ベースラインモデル 1, 2 における FC 層の出力次元数，Dropout 率ならびに出力層の次元数は，共通の設定とした．FC 層の出力次元数はそれぞれ 1024 次元，512 次元とした．Dropout 層における Dropout 率は 0.5 とした．出力層の次元数は，動画データセットの分類クラス数 $n_{class}$ と同じ次元数とし，活性化関数には softmax 関数を使用した．

### 4.2　提案モデル 1

提案モデル 1 は，I3D 及び InceptionTime の出力を行列積演算 (ここでは，matmul と表記) により結合したモデルである．提案モデル 1 のモデル図を図 8 に示す．提案モデル 1 の matmul では，次元数の異なる 2 つの特徴量を使用した行列積演算を行う．その結果から，結合特徴量を作成する．matmul により作成した結合特徴量は，2 層の FC 層，Dropout 層，出力層を経由した後，学習モデルの入力に対する分類結果を出力する．FC 層の出力次元数はそれぞれ 1024 次元，512 次元とした．Dropout 層における Dropout 率は 0.5 とした．出力層の次元数は，動画データセットの分類クラス数 $n_{class}$ と同じ次元数とし，活性化関数には softmax 関数を使用した．

### 4.3　提案モデル 2

提案モデル 2 は，I3D 及び InceptionTime の出力を連結操作により結合したモデルである．提案モデル 2 のモデル図を図 9 に示す．提案モデル 2 の Concatenate では，2 つの特徴量を連結する操作を行う．これにより，2304 次元の結合特徴量を作成する．結合特徴量の作成後から分類結果の出力までの流れ，FC 層及び出力層の出力次元数ならびに Dropout 率は，提案モデル 1 と同様の設定とした．

### 4.4　提案モデル 3

提案モデル 3 は，2 つの I3D 及び InceptionTime の出力を連結操作により結合したモデルである．提案モデル 3 のモデル図を図 10 に示す．提案モデル 2 と異なる点は，モデルの入力に RGB フレーム，Optical Flow ならびに Pose 特徴量の 3 種類を使用している点である．提案モデル 3 の Concatenate では，3 つの特徴量を連結する操作を行う．これにより，4352 次元の結合特徴量が作成する．結合特徴量の作成後から分類結果の出力までの流れ，FC 層及び出力層の出力次元数ならびに Dropout 率は，提案モデル 1, 2 と同様の設定とした．

### 4.5　I3D

本研究では，ResNet-50 [11] の構造をベースにした I3D [3] を使用する．ResNet-50 ベースの I3D のモデル図を図 11 に示す．I3D は，複数枚の RGB フレームもしくは Optical Flow を入力し，空間方向と時間方向を考慮した 3 次元畳み込み及びプーリングを行うことで，入力に対応する 2048 次元の特徴量を抽出する．

### 4.6　InceptionTime

3 節で述べた Pose 特徴量は 17 種類の骨格座標データの時間的な推移を表現したものであるため，時系列データとして扱

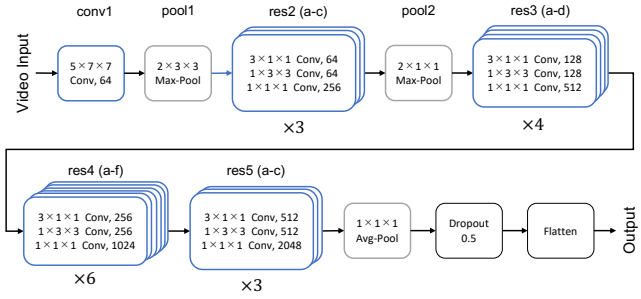
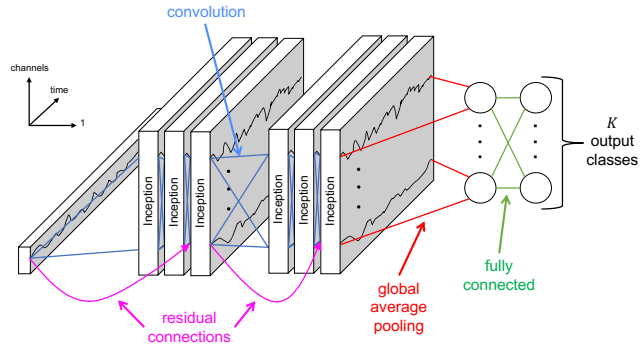
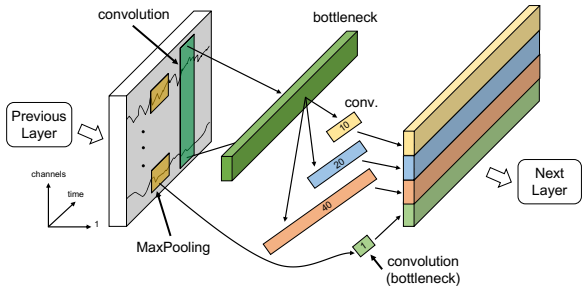



図 11　ResNet-50 の構造をベースとした I3D




(a) Inception Network




(b) Inception Module

図 12　InceptionTime ( [12] を参考に作成)

うことができる．したがって，時系列データから動作分類のための特徴量を抽出するモデルの導入が最適だと考えた．そこで本研究では，Fawaz らが提案した時系列データ分類モデルである InceptionTime [12] を Pose 特徴量の入力直後に導入した．InceptionTime のモデル図を図 12 に示す．InceptionTime は 1 次元畳み込み処理を導入した CNN ベースのモデルで，入力された時系列データにおける数値の変化・パターン等に基づいた特徴量抽出を行う．本研究で使用する InceptionTime は，入力された Pose 特徴量に対応する 256 次元の特徴量を抽出する．

## 5　評価実験

提案手法の有効性を検証するために評価実験を行う．本節では，最初に評価実験で使用するデータセット，評価指標ならびに実験条件について述べる．その後，評価実験の結果及び推論結果の比較について述べる．

### 5.1　データセット

本研究では，Shao らの研究で作成された FineGym データセット [5] を使用した．このデータセットは，体操競技に関する動画で構成されており，各動画に対して 3 種類のラベルが付与されている．付与されている 3 種類のラベルは Event，Set，Element と定義されている．Event とは，動画に対応する体操競技の名前に関するラベルで，Vault(跳馬)，Uneven Bars(段違い平行棒)，Balance Beam(平均台)，Floor Exercises(床運動) の 4 種類の中から 1 個のラベルが付与されている．Set 及び Element は，動画の中で選手が披露した技に基づいたラベルが付与されている．Set は，‘FX_front_salto’，‘BB_leap_jump_hop’ のように，技の大分類に基づいたラベルが付与されている．一方で Element は，‘salto forward stretched with 2 twist’，‘switch leap with 0.5 turn’ のように，技の正式名称に基づいたラベルが付与されている．Element ラベルの名前は，FIG（国際体操連盟）が策定している体操競技採点規則 (Code of Points) [13] に基づいて定義されている．FineGym データセットで定義されている Element ラベルの種類は全部で 530 種類であるが，その中から 99 種類もしくは 288 種類のラベルに関する動画データをピックアップしたサブセットを本研究の評価実験で使用した．99 種類のラベルをピックアップしたサブセットを Gym99，288 種類のラベルをピックアップしたサブセットを Gym288 と定義する．FineGym データセットの内訳を表 1 に示す．

表 1　FineGym データセットの内訳

| Subset | $n_{class}$ * | $T_{pose}$ ** | 学習データ数 | テストデータ数 |
|---|---|---|---|---|
| Gym99 | 99 | 888 | 26,320 | 8,521 |
| Gym288 | 288 | 937 | 29,334 | 9,646 |

* $n_{class}$：サブセットの分類クラス数
** $T_{pose}$：学習モデルへ入力する際の時間方向の次元数

### 5.2　評価指標

評価指標は，Top-1 Accuracy 及び Mean Class Accuracy を使用する．Top-1 Accuracy とは，深層学習モデルの予測結果において，推定確率が最も高かったクラスが正解のクラスである割合を表したものである．Mean Class Accuracy とは，分類クラス単位で Accuracy を算出し，その結果を分類クラス数により平均したものである．分類クラスに該当するデータのみを使用して Accuracy を算出するため，これにより算出される結果は Recall と同じ意味である．クラス $n$ の Recall は式 1 で，分類クラス数 $N$ の Mean Class Accuracy は式 2 で示される．

$$\mathrm{Recall}_n = \frac{TP_n}{TP_n + FN_n} \tag{1}$$

$$\mathrm{MeanClassAccuracy} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Recall}_n \tag{2}$$

### 5.3　実験条件

評価実験では，学習モデルの構成及び骨格情報の欠損補完手法に基づいた実験条件を用意し，性能比較を行う．学習モデルの構成は，4 節で述べたベースライン及び提案モデルを性能比較の対象とする．ベースラインモデル 2 及び提案モデル 3 で使用する Optical Flow は，時間的に連続する 2 枚の RGB フレームに対して TV-$L^1$ [14] を適用し，生成したものを使用する．提案モデル 1, 2, 3 のいずれかを使用する場合，そ

表 2　実験結果

| Method | Gym288 | | Gym99 | |
|---|---|---|---|---|
| | Mean | Top-1 | Mean | Top-1 |
| ベースラインモデル 1 | 25.9 | 59.5 | 54.7 | 66.3 |
| ベースラインモデル 2 | 23.1 | 55.7 | 53.0 | 64.5 |
| 提案モデル 1 + FPD | 29.4 | 62.5 | 62.7 | 71.1 |
| 提案モデル 1 + FLI | 30.3 | 61.9 | 59.9 | 68.9 |
| 提案モデル 2 + FPD | 34.9 | 68.1 | 65.8 | 74.6 |
| 提案モデル 2 + FLI | **38.8** | **69.7** | **68.9** | **76.8** |
| 提案モデル 3 + FPD | 36.3 | 67.6 | 65.2 | 73.9 |
| 提案モデル 3 + FLI | 38.0 | 69.6 | 68.2 | 76.3 |

の入力で使用する Pose 特徴量に対し，3.3 節で述べた骨格情報の欠損補完手法である FPD もしくは FLI のいずれかを適用させる．学習時のエポック数は 60，バッチサイズは 32 とする．損失関数は Cross Entropy Loss を使用する．最適化手法は Adam [15] を使用し，学習率の初期値を 0.001 とする．この条件で学習を開始し，40epoch 経過した後，学習率を 0.001 から 0.0001 に変更する．Pose 特徴量作成における RGB フレームの入力解像度 $W \times H$ は 455×255 とする．学習モデルへの入力に使用する RGB フレームの前処理として，Center Crop，Horizontal Flip，Normalize を行う．ここで，Center Crop の解像度は 224×224，Horizontal Flip を行う確率は 50% (学習時のみ) とする．Horizontal Flip を行った場合，入力データに対応する Pose 特徴量に対し，$x$ 方向の骨格座標データを反転させる．

### 5.4 実験結果

実験結果を表 2 に示す．抽出した動作主体の Pose 特徴量を導入した提案手法は RGB フレームのみを使用したベースラインよりも高い精度という結果となった．特に，2 種類の特徴量を連結操作で結合したモデル (提案モデル 2) を使用し，線形補間による骨格情報の欠損補完を行った条件 (FLI) が最も高い精度であった．

モデルの違いによる結果の比較を行った場合，全体的にベースラインモデル，提案モデル 1，提案モデル 3，提案モデル 2 の順で精度が高くなる傾向が見られた．これは，Pose 特徴量の導入が細粒度な動作分類における精度向上に寄与していることと，特徴量の結合方法によって効果が異なることを示している．提案モデル 1 よりも提案モデル 2 の方が高い精度となった要因として，結合特徴量のデータ構造が挙げられる．提案モデル 1 では，行列積演算により結合特徴量を作成している．そのため，I3D の出力及び InceptionTime の出力に基づいた数値データから結合特徴量の数値データを求める演算を行っている．一方，提案モデル 2 では連結操作により結合特徴量を作成している．そのため，結合特徴量の数値データを求める演算を行っておらず，結合特徴量の中身は I3D の出力に関する部分と InceptionTime の出力に関する部分が明確に分かれている構造となっている．以上のことから，複数の特徴量を使用した演算の結果を扱うよりも，複数の特徴量を連結して結合元の数値データをそのまま扱う方が，細粒度な動作分類において高い効果が見込めると考えられる．

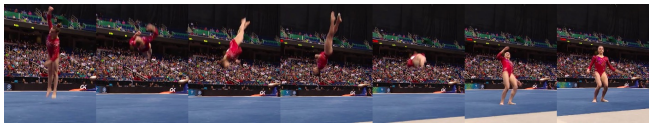
Optical Flow の導入による結果の比較を行った場合，全体的に Optical Flow を導入しない条件の方が，精度が高くなる傾向が見られた．これは，細粒度な動作分類において Optical Flow は精度向上に寄与していないことを示している．その要因として，導入する動作特徴量において考慮される部分の違いが挙げられる．Optical Flow は時間的に連続する 2 枚の RGB フレームにおける物体の変化をベクトルで表現したものであるため，動作主体の変化の他に，動作主体以外の変化及び背景の変化も考慮されていると考えられる．一方，Pose 特徴量は動作主体の骨格情報に関する時間的な推移を表現したものであるため，背景などの変化は考慮されず，動作主体の骨格情報のみ考慮されていると考えられる．細粒度な動作分類が求められる場面では，背景などの変化が少ないケースが多く，動作主体の変化のみ考慮すれば十分であると考えられる．以上のことから，Optical Flow を導入しても精度向上に寄与しなかったと考えられる．

骨格情報の欠損補完手法の違いによる結果の比較を行った場合，提案モデル 2，3 を使用する条件において，欠損直前のデータを使用した欠損補完を行う手法 (FPD) よりも線形補間で欠損補完を行う手法 (FLI) の方が高い精度であった．これは，骨格情報の欠損補完手法の違いによって Pose 特徴量を導入した際の効果が異なることを示している．その一例として，推論結果が欠損していた時間が長いかつ欠損区間の前後における推論結果の差が大きいケースが挙げられる．このケースにおいて FPD を導入した場合では，欠損補完される推論結果と実際に行われた選手の動きの間における乖離が大きくなる可能性が考えられる．一方，FLI を導入した場合では，欠損直前の推論結果だけでなく欠損直後の推論結果も考慮された欠損補完を行うため，FPD よりも乖離が小さくなると考えられる．したがって，欠損補完を行う際の考慮する推論結果の個数の差が，Pose 特徴量を導入した際の効果が異なった要因だと考えられる．

### 5.5 推論結果の比較

FineGym データセットのテストデータを使用し，ベースライン及び提案手法の推論結果を比較する．推論結果を比較した例を図 13，図 14 に示す．ここで，推論結果の比較に使用する提案手法は，表 2 において最も高い精度が確認された提案モデル 2 + FLI の実験条件に基づいた手法とする．

図 13 は，床運動の salto という技に関連するテストデータを使用し，推論結果を比較した例を示している．図 13 より，ベースラインでは 1 回転したと誤った判定をしているのに対し，提案手法では 2 回転したと正しく判定していることが確認できる．図 14 は，段違い平行棒の circle という技に関連するテストデータを使用し，推論結果を比較した例を示している．図 14 より，動画の最後で行われている「半回転して掴み直す」(with 0.5 turn to handstand) という動作をベースラインでは見逃しているのに対し，提案手法では細粒度な動作の違いを正確に捉えていることが確認できる．いずれの例においても，提案手法の導入により体操選手が披露した技における細かい違いを正確

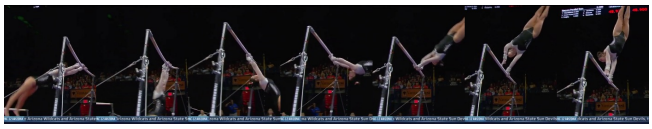


正解ラベル: salto forward stretched with 2 twist

推論結果

- ベースラインモデル1: salto forward stretched with 1 twist
- 提案モデル2 + FLI : salto forward stretched with 2 twist

図 13 推論結果例 1



正解ラベル: giant circle backward with 0.5 turn to handstand

推論結果

- ベースラインモデル1: giant circle backward
- 提案モデル2 + FLI : giant circle backward with 0.5 turn to handstand

図 14 推論結果例 2

に捉えられるようになったため，細粒度な動作分類における精度向上に寄与したといえる．

## 6 おわりに

本研究では，動作主体の骨格情報の推移に着目した Pose 特徴量を抽出する手法及び RGB フレーム，Optical Flow を代表とする既存の動作特徴量に加えて抽出した Pose 特徴量を入力とする学習モデルを提案した．評価実験を実施し，既存の動作特徴量を使用したベースラインと Pose 特徴量を導入した提案手法の性能比較を行った．その結果，提案手法はベースラインの精度を上回ることが確認された．また，ベースライン及び提案手法の推論結果を比較したところ，提案手法の導入により体操選手が披露した技における細かい違いを正確に捉えている例が確認された．

今後の課題として，提案手法の中で最も効果のあった要素に関する調査 (Ablation Study)，細粒度なラベルが付与されている他の動画データセットを使用した実験などが挙げられる．

## 文 献


[1] 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*, pp. 568–576, 2014.

[2] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 4489–4497. IEEE Computer Society, 2015.

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 4724–4733. IEEE Computer Society, 2017.

[4] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, Vol. abs/1212.0402, , 2012.

[5] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2613–2622. Computer Vision Foundation / IEEE, 2020.

[6] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: towards action recognition without representation bias. 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*, Vol. 11210 of *Lecture Notes in Computer Science*, pp. 520–535. Springer, 2018.

[7] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. Fine grained sport action recognition with twin spatio-temporal convolutional neural networks. *Multim. Tools Appl.*, Vol. 79, No. 27-28, pp. 20429–20447, 2020.

[8] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 1653–1660. IEEE Computer Society, 2014.

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7103–7112. Computer Vision Foundation / IEEE Computer Society, 2018.

[10] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: regional multi-person pose estimation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2353–2362. IEEE Computer Society, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

[12] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Min. Knowl. Discov.*, Vol. 34, No. 6, pp. 1936–1962, 2020.

[13] Fdration Internationale de Gymnastique (FIG). 2017-2020 CODE OF POINTS for Womens Artistic Gymnastics, 2017. https://eugymnastics.files.wordpress.com/2017/04/cop_wag_2017-2020_ici-e1.pdf.

[14] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-$L^1$ optical flow. In Fred A. Hamprecht, Christoph Schnörr, and Bernd Jähne, editors, *Pattern Recognition, 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007, Proceedings*, Vol. 4713 of *Lecture Notes in Computer Science*, pp. 214–223. Springer, 2007.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.