Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30

深層学習モデルを用いたフィギュアスケートにおけるステップシークエンスのエレメント認識

岩田 あきほ[1,a)]　川島 寛乃[2]　大越 匡[2]　中澤 仁[1]

概要：フィギュアスケートでは，全ての演技の採点は一つ一つの技の出来栄えの判定結果を複雑に組み合わせて行う．そのため採点に審判の主観が含まれる可能性が問題視されている．そこで本研究では，採点の適正化を支援する取り組みのはじめの段階として深層学習モデルを使って技（エレメント）の認識を行う．本論文ではフィギュアスケートの演技構成要素の一つであるステップシークエンスに着目し，動画からエレメントの判別を行うモデル SkateNet を提案する．また，独自に収集したフィギュアスケートの動画よりスケートデータセットを作成する．実験ではフレームごとにエレメントの認識を行い精度で評価する．その結果，背景などの余分な情報が含まれていると元動画からエレメントを推定し識別することは難しいことから，そのようなノイズをできるだけ除去することが重要であることがわかった．また，より詳細な下半身の関節座標の使用により精度が向上するであろうことが示された．

Element Recognition of Step Sequences in Figure Skating Using Deep Learning Model

Akiho Iwata[1,a)]　Hirono Kawashima[2]　Tadashi Okoshi[2]　Jin Nakazawa[1]

Abstract:
In figure skating, all the performances are scored by a complex combination of judging results of each element. Therefore, the possibility of subjectivity of judges is considered to be a problem in scoring. In this study, as the first step of supporting the improvement of scoring, we use a deep learning model to recognize the elements of a performance. In this paper, we focus on the step sequence, which is one of the components of figure skating, and propose a model, SkateNet, that can identify the elements from the video clips. We also make a skate dataset from figure skating movies we collected. In our experiments, we recognize elements in each frame and evaluate the accuracy of the model. As a result, we found that it is important to remove the extra information such as background as much as possible because it is difficult to estimate and identify elements from the original video. It is also found that more detailed coordinates of the lower body joints would improve the accuracy of the estimation.

1. はじめに

さまざまなスポーツにおいて，競技の勝敗を定めるためにルールに基づいた公平な採点は重要である．特に技の出来栄えや演技の評価が計測できず審判の判断に依存する競技においては，審判の判定の公平性が常に議論されている．

フィギュアスケートにおいては，公正な採点の実現にむけた議論が活発であり，これまで幾度もルールの策定や改正が行われてきた．フィギュアスケートの演技の得点は技を評価する技術点，音楽などの構成・表現力を評価する演技構成点の合計から，演技中の転倒やルール違反の減点をすることで算出される．技術点は技の難易度によって定められた基礎点と技の出来栄え評価に基づく出来栄え点（GOE）の合計から算出されており，演技に含まれるジャンプやスピン，ステップシークエンスなどの技（エレメン

[1] 慶應義塾大学環境情報学部
Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa, 252-0882, Japan
[2] 慶應義塾大学大学院政策・メディア研究科
Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa, 252-0882, Japan
[a)] t18092ai@sfc.keio.ac.jp

Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30

図 1 ステップシークエンスの 12 個のエレメントの一例．作成したデータセットより．

ト）の正確な認識が必要とされる．公式試合では正確なエレメント認識を円滑に行うため 13 人の審判がそれぞれ，競技大会の進行と総監督（1 人），技の種類と難易度の判定（3 人），技の出来栄えの評価と演技構成点の判定（9 人）の役割を担っている．

この中で，技の出来栄えの評価と演技構成点の判定は審判にとって負担が大きく，審判の判定に選手のネームバリューや国籍といった主観が含まれてしまう可能性が指摘されている．審判の負担が大きい要因としては，以下の 3 つの理由がある．第一に，演技における技（エレメント）は広いスケートリンクを使って，数分にわたって高速に行われる．その間連続するエレメントの細かい動作を一つ一つ認識し，出来栄えを評価するためには，高度な識別の技術を要する．第二に，演技の構成や選曲も選手ごとに大きく異なる．そのため，常に異なるエレメントの組み合わせを評価しなければならない．第三に，スコアの算出までの時間は限られている．ジャンプのような配点の高い重要な技においては回転不足の判定など適宜スローモーション動画が用いられることもあるが，ステップシークエンスのような細かい動作のエレメントが連続して行われる部分では，細かい一つ一つのエレメント全てを確認するだけの時間を確保できない．

このように負担の大きい動作認識の問題を解決するために，体操やバスケットボールなどの競技では，人工知能技術を用いた動作認識の自動化が検討されている．特に近年では深層学習の発展に伴い，画像やセンシングデータから人の行動認識（Action Recognition）を行う研究が盛んである．行動認識を用いてエレメントの認識を自動化することで，審判の負担を減らし，さらに選手の動作の正確な認識を実現することが期待されている．しかし，これまでフィギュアスケートにおいては，エレメントの自動認識は実現していない．その一因として，フィギュアスケートでは曲に合わせた多彩な表現が求められ，同じエレメントでも動作や長さが異なり，動作の認識の難易度が高いことが挙げられる．

そこで本研究では，フィギュアスケートにおいてエレメント認識の自動化を目指した取り組みの第一段階として，特に一つ一つの動作が細かく判定の公平性が問題視されるステップシークエンス部に着目し，ステップシークエンスのエレメント認識に取り組む．本研究ではエレメント認識のために深層学習を用いた SkateNet を提案し，ステップシークエンスの動画から各エレメントを認識する．またモデルの学習および評価のために，収集した演技動画を画像に切り出し，フレームごとにエレメントのラベルを付与した独自のデータセットを作成する．実験ではフレームごとのエレメント認識の精度を用いて SkateNet の性能を評価し，結果を分析する．

2. ステップシークエンスの構成

フィギュアスケートは，ジャンプ，スピン，ステップシークエンスなどの動きを組み合わせて音楽にのせて滑走する競技であり，その中でもステップシークエンスは基礎的なスケーティング技術力および表現力が反映される重要な技である．しかし，高速で連続して行われるステップシークエンスのエレメントを瞬時に認識し，出来栄えの評価を行うことは難しく，また一つ一つのエレメント全てを確認するだけの時間を確保できないことから判定の公平性が問題視されている．

ステップシークエンスは，**図 1** の 12 個のエレメントのいずれかを組みわせ構成される．例えば「ツイヅル」とは

Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30

片足で 3 回転以上回転し続ける技であり，「ループ」は片足で円を描くように 1 周する技である．それぞれのエレメントでは右足，左足，前向き，後ろ向き，アウトサイド，インサイドを組み合わせた合計 8 パターンの動きがある．ただし，1 つの演技の中に必ず 12 のエレメント全てが含まれているわけではなく，同じエレメントを複数回行うこともある．後述のデータセット においては，12 のエレメントの中では「カウンター」が行われる頻度が最も高かった．エレメントごとの長さは異なり，「ツイヅル」や「チェンジエッジ」は約 2 秒と比較的長く，「チョクトウ」は約 1 秒未満と短い．ステップシークエンスのエレメントは本来，滑走することによって氷上に描かれた図（トレース）から識別される．そのため，すべての選手の下半身は図を描くためにエレメントごとに似たような動きをする．一方で，上半身は表現力や音楽の解釈における演技構成点を上げるために，手を上げたり回したりするなど選手によって異なった動きをする．そのためステップシークエンスのエレメントの識別には下半身の動作が重要になる．

3. 関連研究

3.1 スポーツの採点支援

さまざまなスポーツの競技種目において，正確で公平な採点を目指し，採点の補助や自動化に取り組んだ研究が進められている．藤原らは，国際体操連盟および日本体操協会との共創により，3D センシング技術を用いた体操競技の採点支援に取り組んでいる [10]．Shao らは体操競技の演技を対象とした行動認識に向けて，一連の基本的な動作とそれぞれに含まれるサブ動作を階層的に細かく定義したデータセットを作成した [6]．

フィギュアスケートにおける取り組みとしては，動画全体から点数を予測することを目指した研究例が存在する．Xu らは Self-Attentive LSTM と Multi-scale Convolutional Skip LSTM と呼ばれる 2 つの補完的なコンポーネントを含むディープアーキテクチャを提案し，フィギュアスケートの演技の動画全体から自動採点を行った [9]．しかし，本研究のように細部の動作の認識に取り組み，実現している研究はまだ存在しない．

3.2 行動認識（Action Recognition）

動画から人の動きを認識する行動認識（action recognition）の分野では，多くのデータセットの作成やモデルの提案が行われている．Karen らは空間的ネットワークと時間的ネットワークを組み込んだ Two-Stream Convolutional Networks を用い，動画中の行動認識を行った [7]．Ji らは監視ビデオにおける人間の行動の自動認識のために，空間次元と時間次元の特徴を抽出し，複数の隣接するフレームに符号化された運動情報を取り込む 3D CNN を開発した [4]．Stroud らは推論の際に物体やカメラの移動によって生じる隣接フレーム間の物体の動きがわかる optical flow[3] を必要とせず，Two-Stream よりも優れた性能を発揮する distilled 3D CNN（D3D）と呼ばれる手法を提案した [8]．

4. SkateNet

本研究では，ステップシークエンスのエレメント認識モデルとして，深層学習を用いた SkateNet を提案する．画像の連続データである動画からエレメントを認識するためには，瞬間的な身体の配置を捉えるための画像的特徴と，動作の時系列的な特徴の双方を同時に考慮する必要がある．そこで，SkateNet では画像特徴量抽出と時系列処理を行う 2 つのニューラルネットを組み合わせ，エレメント認識を行う．SkateNet の全体図を**図 2** に示す．SkateNet では，動画から任意の長さの画像に切り出したフレームの系列を入力とし，フレームごとにエレメントの予測結果を出力する．まず，入力画像を画像特徴量抽出部に入力し，画像の畳み込み結果または姿勢推定結果の関節座標に変換する．ここで得られた結果を画像特徴量とする．次にフレームの画像ごとに得られた画像特徴量を時系列処理部に入力する．時系列処理部では，入力ごとに該当するエレメントを推定し，そのラベルを出力する．

4.1 画像の特徴量抽出

本研究では画像特徴量として，単純に画像を畳み込むことで得られる特徴量ベクトル（image-feature）の他に姿勢推定結果の関節座標（joint-feature）の使用を試みる．

4.1.1 Image-feature

まず，入力画像を畳み込みニューラルネット（CNN）で畳み込む方法について説明する．動画は画像の連続データであり処理が重くなるため，本研究では CNN の軽量化モデルである MobileNetV2[5] を用い，畳み込み処理を行った．入力画像を直接畳み込む方法他に，入力画像に姿勢推定の結果を描画してから MovileNetV2 を用いて畳み込み処理をする方法も行った．姿勢推定には OpenPose[1] を用いる．姿勢推定結果を入力画像に描画した結果の一例を**図 3** に示す．フィギュアスケートの動画から切り出した画像には，選手以外にも，フェンス，観客，コーチ，審判などの様々な情報が背景として含まれている．これらの背景情報はエレメントとは関わりがないものであるが，image-feature では画像全体を入力とするため，選手の動作以外の情報も含まれる．

4.1.2 Joint-feature

演技の動画においては，選手の動作以外の背景情報は，余分な情報としてエレメントの識別に影響を与えてしまう可能性がある．そこで，本研究では image-feature の他に，姿勢推定を用いて関節座標を推定し，その結果を joint-feature として用いる．関節座標を用いることで選手の動作のみに

Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30

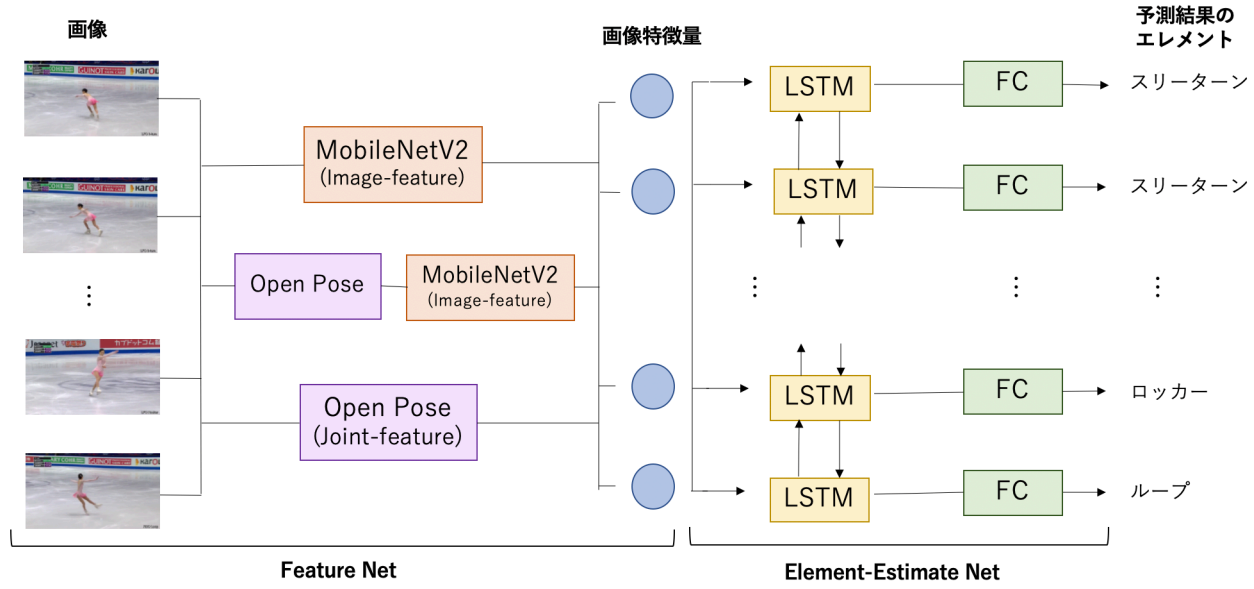


図 2 SkateNet:画像特徴量抽出と時系列処理を行う 2 つのニューラルネットを組み合わせた深層学習モデル.

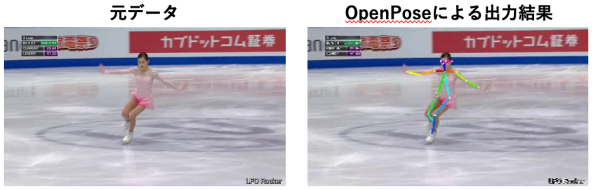


図 3 姿勢推定結果を入力画像に描画した一例.

焦点を当てることが可能である．姿勢推定には OpenPose を用い，画像から 18 個の関節座標の推定を行う．

得られた関節座標は，以下の 3 種類の方法で用いる．

- OpenPose によって検出された 18 箇所の関節の全てを用いる
- 18 箇所の関節のうち鼻，肩，手首などの 12 箇所を上半身の関節座標として用いる
- 18 箇所の関節のうち膝，足首などの 6 箇所を下半身の関節座標として用いる

OpenPose によって出力される 18 箇所の関節部位と，上半身・下半身の区分について**表 1** に示す．

撮影位置が遠いことから人物が小さく映り，人物自体を認識できなかったフレームはデータとしては使用しない．1 フレームに対し 18 箇所の関節座標が出力されるようにしている．しかし，撮影方向や回転する動きなどで選手の関節部位が重なり関節が認識できないことや，フレームの解像度が低いことから手先や足先などの細部が鮮明に映っていないため検出できなかった関節座標が存在する．それは動画 1 本（21，438 フレーム）に対し最大 32 %の割合で起こった．そこで今回は検出できなかった関節座標を（x，y）＝（0，0）に置き換える．

表 1 OpenPose による 18 箇所の関節部位と上半身・下半身の区分.

関節番号	関節部位	振り分け
1	鼻	上半身
2	心臓	上半身
3	右肩	上半身
4	右肘	上半身
5	右手首	上半身
6	左肩	上半身
7	左肘	上半身
8	左手首	上半身
9	右腰	下半身
10	右膝	下半身
11	右足首	下半身
12	左腰	下半身
13	左膝	下半身
14	左足首	下半身
15	右目	上半身
16	左目	上半身
17	右耳	上半身
18	左耳	上半身

4.2 時系列処理

SkateNet の時系列処理部では，得られた特徴量をゲート付き RNN の 1 つである LSTM（Long-short term memory）[2] に入力する．本研究では単層で双方向の LSTM を用いる．LSTM のそれぞれの出力を fully-connected layer（FC）に通し，12 種類のエレメントの分類または「その他」クラスを含める 13 クラス分類を行う．

5. 実験

本研究では独自に作成したデータセットを使用して SkateNet の学習，評価を行う．評価指標にはフレームの画

Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30

表 2 データセットに含まれるエレメントごとの数とフレーム数.

エレメント	エレメント数	フレーム数
ブラケット	72	1,392
エッジの変更	28	639
シャッセ	7	180
チョクトウ	73	1,512
カウンター	108	2,678
クロスロール	14	513
ループ	76	2,277
モホーク	27	499
ロッカー	99	2,061
スリーターン	76	1,352
トウステップ	28	665
ツイヅル	84	3,266
その他	-	28,754
合計	-	45,778

像ごとのエレメント分類結果の精度を用い，前処理方法，モデル構造，学習方法の違いによる結果の違いについて比較し分析する．

5.1 データセット

本研究では独自にフィギュアスケートの映像を収集し，ステップシークエンスの各フレームにエレメントラベルを付与したデータセットを作成した．データセットの作成手順は以下の通りである：

（1）動画の収集 ウェブ上から冬期オリンピック，世界選手権大会などの女子選手のショートプログラム（2分 40 秒）およびフリースケーティング（4 分または 3 分 30 秒）の動画を 40 本収集した．

（2）ステップシークエンス部の切り出し 収集した動画からステップシークエンスが映っている箇所を切り出した．ステップシークエンス部の動画の平均の長さは約 40 秒であった．

（3）連続画像（フレーム）への切り出し ステップシークエンス部の動画を連続画像（30fps）に変換した．

（4）アノテーション 切り出したフレームごとに，12 種類のエレメントまたは「その他」の計 13 クラスのラベルを付与した．動画 1 本に対し，「その他」クラスのフレームが半分近くをしめていた．そのため「その他」クラスが他のエレメントに比べ極端にフレーム数が多くなった．

データセットに含まれるエレメントごとのエレメント数とフレーム数は**表 2** の通りである．

5.2 データ処理

本研究では Data Augmentation として，データ数を増やすためにリサイズ，画像反転処理，色調補正処理，および複数の開始点からのフレーム分割（split）を行う．フレーム分割では，連続画像に対して開始点を 10 個ずつずらしながら指定したフレーム数で分割する．この結果得られたフレームに対し画像ごとにデータセットのラベルを付与する．今回の実験では 12 個のエレメントのみを含むデータと「その他」クラスを含むデータとで，それぞれ 3:1 のフレーム数で訓練データ，テストデータに分割した．それぞれのフレーム分割における訓練データ，テストデータのフレーム数は**表 3** の通りである．

表 3 訓練データ，テストデータに含まれるフレーム数.

Seq Length	クラス数	訓練データ	テストデータ
分割なし	13	94,308	46,977
	12	38,394	14,799
50	13	479,700	180,150
	12	187,800	41,100
100	13	861,600	398,100
	12	358,800	41,400
300	13	2,142,900	916,200
	12	397,800	118,800
500	13	2,610,000	1,288,500
	12	46,500	18,000

5.3 評価指標

本研究ではエレメント認識を以下の精度で評価する：

$$\mathbf{Accuracy} = \frac{N_{\mathbf{correct}}}{N} * 100 \tag{1}$$

N は全フレーム数，$N_{\mathbf{correct}}$ は正解したフレーム数である．

5.4 実験設定

本研究では以下の 4 つの設定下で実験を行う．

実験 1：分割幅の比較

分割なし，50，100，300，500 フレームごとに動画を分割した場合の計 5 パターンでそれぞれ精度評価を行う．画像特徴量には image-feature を用いる．

実験 2：その他クラスの有無

「その他」クラスを除いた 12 のエレメントのみをデータセットとし，実験 1 と同等の工程を行う．

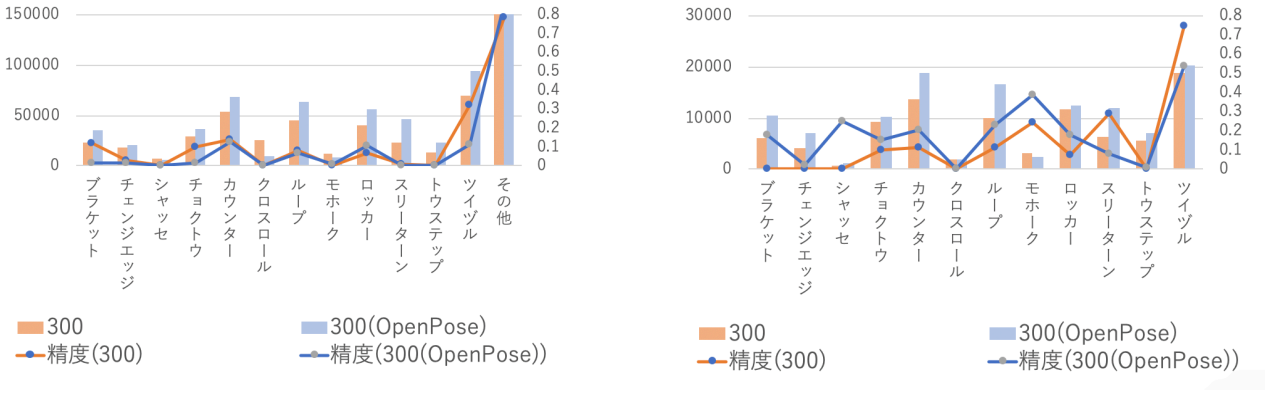
実験 3:image-feature における姿勢推定結果の描画の有無

実験 1 によって出力された精度の高いフレーム数において．OpenPose を用い姿勢推定結果をフレーム上に描画し，実験 2 と同等の工程を行う．

実験 4:image-feature と joint-feature の比較

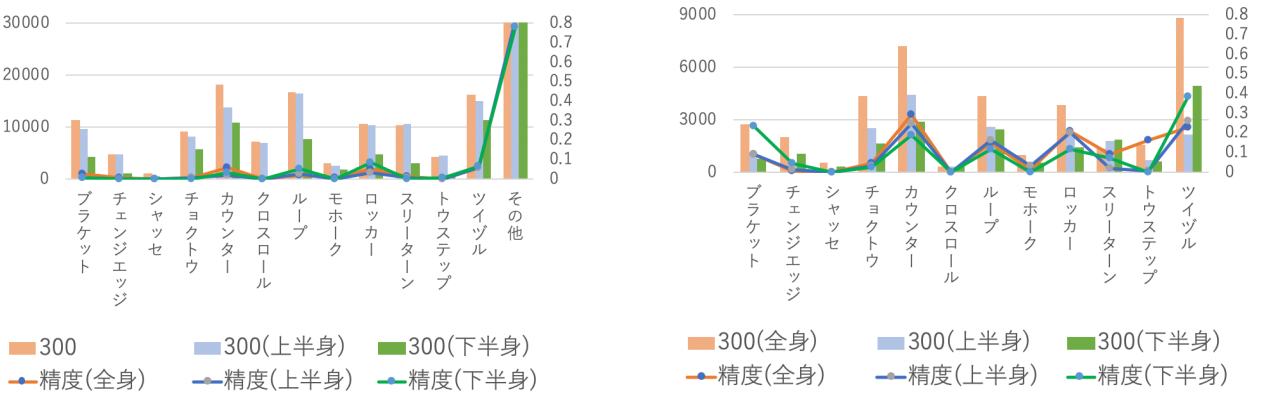
実験 1 によって出力された精度の高いフレーム数において全身の関節座標，上半身の関節座標，下半身の関節座標の 3 種類を joint-feature として用いる．

Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30



[1] 13 クラス分類　　[2] 12 クラス分類

図 4　image-feature におけるエレメントごとの精度．グラフの横軸はエレメント名，主軸はフレーム数，2 軸が精度を表す．



[1] 13 クラス分類　　[2] 12 クラス分類

図 5　joint-feature におけるエレメントごとの精度．グラフの横軸はエレメント名，主軸はフレーム数，2 軸が精度を表す．

表 4　13 クラス分類における精度．

入力	Seq. Length	Batch Size	精度（%）
フレーム	分類なし	1	2.8
フレーム	50	32	18.5
フレーム	100	16	34.0
フレーム	300	8	51.9
フレーム	500	4	51.3
フレーム（OpenPose）	300	8	55.4
関節座標（全身）	300	256	51.8
関節座標（上半身）	300	256	51.4
関節座標（下半身）	300	256	53.4

表 5　12 クラス分類における精度．

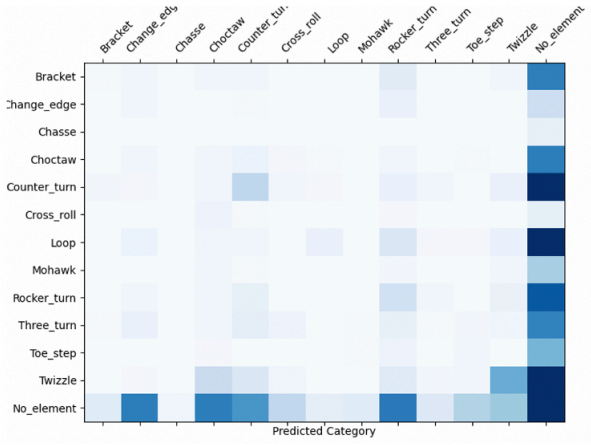
入力	Seq. Length	Batch Size	精度（%）
フレーム	分類なし	1	12.1
フレーム	50	32	16.9
フレーム	100	16	18.8
フレーム	300	8	23.2
フレーム	500	4	6.9
フレーム（OpenPose）	300	8	22.1
関節座標（全身）	300	64	17.1
関節座標（上半身）	300	64	14.6
関節座標（下半身）	300	64	17.4

5.5　結果

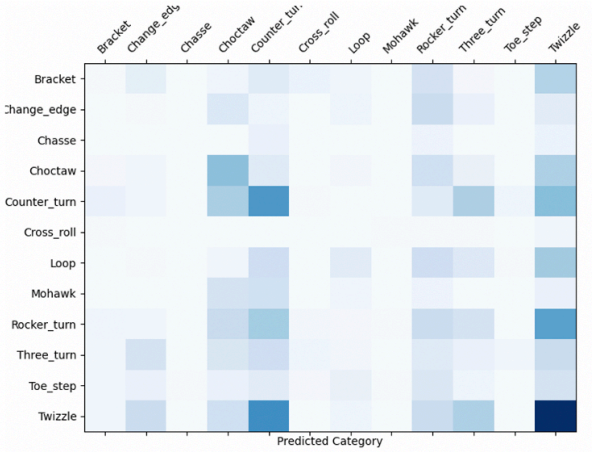
実験 1 の結果では，動画を 300 フレームで分割したものが最も精度が高く，13 クラス分類では 51.9 %であった（**表 4**）．

実験 2 の「その他」クラスを除いた 12 クラス分類でも 23.2 %と，300 フレームで分割したものが最も精度が高くなった（**表 5**）．エレメントごとの精度では 13 クラス分類より 12 クラス分類の方が精度が高くなる傾向がある．その中で，13 クラス分類では「その他」が，12 クラス分類では「ツイゾル」が最も精度が高くなった．つまり，データ数が多ければ多いほどエレメントの精度も高くなる．しかし 13 クラス分類の場合，データ数が比較的多い場合でも「その他」以外の 12 のエレメントのほとんどは極端に精度が低くなった（**図 4**）．横軸が予測したエレメント，縦軸が正解のエレメントを表し，色が濃いほどフレーム数が多いことを示す混合行列からは 2 つの結果が得られた．1 つ目は，13 クラス分類では 12 のエレメントの多くは「その他」クラスに分類されている．これは「その他」の精度は高

Vol.2020-MBL-96 No.27
Vol.2020-UBI-67 No.27
Vol.2020-CDS-28 No.27
Vol.2020-ASD-19 No.27
2020/9/30

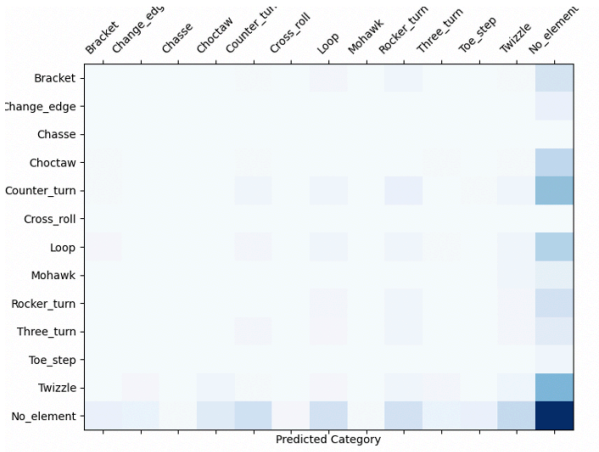


[1] 13 クラス分類

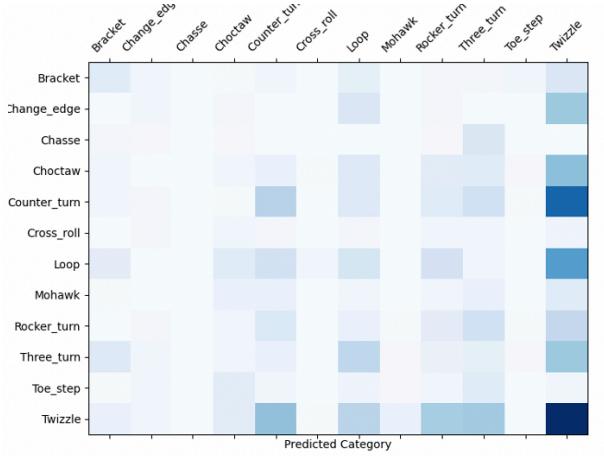


[2] 12 クラス分類

図 6 image-feature における混同行列．横軸は予測したエレメント，縦軸は正解のエレメントを表す．色が濃いほどフレーム数が多いことを示す．



[1] 13 クラス分類



[2] 12 クラス分類

図 7 joint-feature（下半身）における混同行列．横軸は予測したエレメント，縦軸は正解のエレメントを表す．色が濃いほどフレーム数が多いことを示す．

いが 12 のエレメントの精度が極端に低いことと関連があると言える（**図 6**[1]）．2 つ目は 12 クラス分類では「ロッカー」が「カウンター」と認識されている傾向がある．これはどちらも片足で半回転する動きであり，その違いは回転の方向（右回転または左回転）である．そのため，この 2 種類は人の目で見た場合でも識別が難しいと言われている（図 6[2]）．

一方**実験 3** で姿勢推定結果を描画したフレームにおける image-feature を用いた場合，13 クラス分類では精度が 55.4 %まで上がったが，12 クラス分類では 22.1 %と精度が下がった（表 4，表 5）．エレメントごとの精度では，特に 12 クラス分類の場合動画から切り出したフレームよりも，姿勢推定結果を描画したフレームの方が精度が高いものが多かった（図 4[2]）．これらのことから，姿勢推定結果の描画が，識別する際にプラスになるエレメントとマイナスになるエレメントに分かれることがわかる．

実験 4 で関節座標（joint-feature）を用いた場合，image-feature を用いた場合よりも精度は全体的に下がった．下半身の関節座標を用いたものが最も精度が高くなり，13 クラス分類では 53.4 %，12 クラス分類では 17.4 %だった（表 4，表 5）．評価に使用したデータ数に対して各エレメントの精度が最も効率よく高く出たものは下半身の関節座標であった．これらからエレメントの識別には下半身の情報が重要になることがわかる（**図 5**）．また，混合行列からは image-feature を用いた場合に比べ，下半身の関節座標では各エレメントが「ループ」に分類されることが多いと言える（**図 7**[2]）．

6. 考察

本章では SkateNet を用いたエレメント認識の評価の結果から以下の 3 つの観点で考察を行う．

第一にデータ数と精度について，「その他」や「ツイズル」の精度が高かったことから，データ数が多いエレメントは精度が高くなる．また 13 クラス分類の場合「その他」

以外の 12 のエレメントの精度が極端に低かったことから，データ数の少ないエレメントがデータ数の多いエレメントに誤って予測されている．このようにエレメントごとのデータ数の不均衡が精度に大きく影響している．

第二に画像の特徴量抽出について，image-feature を用いる場合，選手以外にフェンスや観客といった背景の情報が加味されてしまっている．元動画からエレメントを推定するにはそのようなノイズをできるだけ除去することが重要である．joint-feature において OpenPose を用いる場合の問題点として正確な姿勢推定が行われていないことが挙げられる．これはフレームの解像度が低いことにより手先や足先など選手の細部が鮮明に映っていないことや，フレーム内に映る観客や審判，コーチといった選手以外の人物に反応してしまっていることが原因である．また，撮影方向や回転などで選手の関節部位が重なることで検出できなかった関節座標を（x，y）=（0，0）に置き換えたことにより，座標の正確性が落ち，データにばらつきができてしまったことが精度に影響していると考える．

第三に，技の識別の難易度について述べる．エレメントの識別は「1 回転」と「1 周」のように回る動作の種類や回転の向きの違いといった細部にわたる．そのため「ロッカー」や「カウンター」のように人の目でも識別が難しいと言われている技が存在する．また，「ループ」のように片足で前向きまたは後ろ向きに 1 周する動きの場合，image-feature では 1 回転と 1 周の違いがわかりやすいが， joint-feature では情報量が少なくなるためその違いを見極めることが難しいと考える．そのためそれぞれの違いがわかるようにするためにはより詳細な情報が必要になる．実験の結果からエレメントを識別するには下半身の情報が重要になることがわかった．このことから，全身の情報を用いる場合は下半身に重きを置く，または下半身のみに焦点を置き，「爪先」「かかと」「膝裏」のようにより詳細な関節座標を用いる．それによって回転の向きや動作の種類も識別できるようになり，精度が上がると考える．

今後は，エレメントごとのデータ数のばらつきをなくし，データ数を増やす．また姿勢推定の質を向上させるため，フレームの解像度を上げ，バウンディングボックスを使用するなどして選手一人に OpenPose が反応するよう改善する．また，より詳細な下半身の関節座標を検出し，関節部位が重なることで検出できない関節座標は，その関節に近い各関節座標の平均に置き換える．

7. まとめ

フィギュアスケートにおいて，高速で複雑なエレメント認識を要する採点の正確性および公平性が問題視されている．本研究ではこの問題に対処するためにステップシークエンス部に着目し，エレメント認識を行う深層学習モデル SkateNet を提案した．また，ステップシークエンスの動画からフレームとエレメント名を対応させた独自のデータセットを作成し，エレメントの推定と精度評価の実験を行った．その結果，背景などの余分な情報が加味されている点で元動画からエレメントを推定し識別することは難しいことから，そのようなノイズをできるだけ除去することが重要であることがわかった．また，エレメントの識別では詳細な下半身の関節座標の使用により精度が向上するであろうことが示された．今後の課題として挙げられるのはデータ数を増やし，姿勢推定の質を向上させること，改善したデータセットからエレメントの識別を行い出来栄えの評価を行うシステムを作成することである．

謝辞

本研究は，JST，CREST，JPMJCR19A4 の支援を受けたものである．

参考文献

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[3] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, Vol. 281, pp. 319–331. International Society for Optics and Photonics, 1981.

[4] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35, No. 1, pp. 221–231, 2012.

[5] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[6] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2020.

[7] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pp. 568–576, 2014.

[8] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 625–634, 2020.

[9] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[10] 藤原英則, 伊藤健一. Ict による体操競技の採点支援と 3d センシング技術の目指す世界, 2018.