多視点で撮影された動画を用いた技能判定

佐藤荘一朗† 青野 雅樹††

† 豊橋技術科学大学 情報・知能工学課程 〒 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, ††aono@tut.jp

あらまし 近年，Youtube 等で作業風景を撮影した動画が公開されている．同じ作業風景を撮影した動画を比較することで，動画に映っている作業者の技能レベルを判定することが可能である．しかし，公開されている動画の数は非常に多く，様々な技能レベルが混在している．そのため，膨大な数の動画の中から高い技能レベルを持つ動画を見つけるのは困難である．また，公開されている動画の中には複数の視点から撮影された動画も存在し，それらの動画から得られる特徴量を組み合わせれば，技能判定の精度向上が期待できると考えられる．そこで本研究では，多視点で撮影された動画を用いた技能判定を行う手法を提案し，技能判定の精度向上を目指す．実験には Epic-Skills 2018 Dataset を使用し，その中から作業者の目線を撮影した動画及び作業場所を真上から撮影した動画を有する Drawing，ChopstickUsing データセットを使用した．実験の結果，提案手法はベースラインよりも技能判定精度が向上したことが確認できた．

キーワード 動画，深層学習，動画特徴量，技能判定

1 はじめに

本節では，研究背景，関連研究，ならびに，本研究の目的について述べる．

1.1 研究背景

近年，Youtube 等で作業風景を撮影した動画が公開されている．その例として，絵を描く様子を撮影した動画，料理する様子を撮影した動画等が挙げられる．同じ作業風景を撮影した動画を比較することで，動画に映っている作業者の技能レベルを判定することが可能である．しかし，公開されている動画の数は非常に多く，様々な技能レベルが混在している．そのため，膨大な数の動画の中から高い技能レベルを持つ動画を見つけるのは困難である．そこで本研究では，作業風景を撮影した動画の技能レベルを判定する手法を提案する．

1.2 関連研究

関連研究として，深層学習を使用して動画に映っている人の技能レベルを判定する研究が行われている．Paritosh らは，オリンピックの競技 (飛込競技，跳馬，フィギュアスケート) における選手の技能レベルを判定する手法を提案した [1]．C3D [2] を使用して動画から時空間特徴量を抽出し，それを SVR(サポートベクター回帰) もしくは LSTM へ入力することで選手の技能レベルを判定する手法である．Kim らは，Siamese Network [3] により，技能レベルが高い動画及び技能レベルが低い動画との類似度を推定する方式で技能判定を行う手法を提案した [4]．動画フレームから抽出された視覚的特徴量及び LSTM を使用したエンコードを行ったものを動画特徴量として，それを LSTM を導入した Siamese Network へ入力することで技能判定を行う．Hazel らは，2 つの動画を比較し，技能レベルが高い動画はどちらの動画であるかを判定する手法を提案した [5]．この手法では，Temporal Segment Network (TSN) [6] に基づいた Two-Stream Convolutional Neural Network (2S-CNN) [7] により，空間情報及び時間情報から技能レベルを表すスコアを取得し，それらを統合することで動画の技能レベルを表すスコアを取得する．技能レベルが高い動画はどちらの動画であるかという判定は，動画の技能レベルを表すスコアを基に行われる．また，Hazel らは，長い動画から技能判定に関係する部分をピックアップするために Attention 機構を導入し，技能判定の学習を行う手法も提案している [8]．

1.3 目的

1.2 節で述べた関連研究では 1 視点を有する動画から得られる特徴量を使用して技能判定を行っているが，Youtube 等で公開されている動画の中には同じタイミングで作業風景を複数の視点から撮影された動画も存在する．その場合，複数の視点から撮影された動画から得られる特徴量を組み合わせれば，1 視点分の特徴量のみを使用した場合よりも技能判定の精度向上が期待できると考えられる．また，サッカー，バスケットボールといったスポーツの試合では，試合の様子を複数のカメラで撮影する場合が多いため，スポーツ選手の技能レベル判定にも応用できると考えられる．そこで本研究では，多視点で撮影された動画を用いた技能判定を行う手法を提案し，技能判定の精度を向上させることを目指す．

1.4 本論文の構成

本論文の構成は以下の通りである．2 節では，本研究で使用するデータセットついて述べる．3 節では，特徴量抽出の手法について述べる．4 節では，提案手法と比較するために用意したベースラインについて述べる．5 節では，本研究で提案する

手法について述べる．6 節では，提案した手法の性能を評価するために行った実験について述べる．7 節では，まとめと今後の課題について述べる．

2　データセット

本研究では，動画を用いた技能判定に関するデータセットを使用する．2.1 節では，本研究で使用するデータセットについて述べる．2.2 節では，2 視点から撮影された動画を有する Drawing 及び ChopstickUsing のデータセットの特徴について述べる．2.3 節では，4 分割交差検証による学習を行う場合におけるデータセットの分割について述べる．

2.1　EPIC-Skills 2018 Dataset

本研究では，[5] の研究で使用された EPIC-Skills 2018 Dataset を使用する．このデータセットは 4 種類の作業風景 (Surgery，DoughRolling，Drawing，ChopstickUsing) を撮影した動画及び技能判定に関する注釈で構成されている．EPIC-Skills 2018 Dataset の動画数及び技能判定の際に比較する動画のペア数を表 1 に示す．

表 1　EPIC-Skills 2018 Dataset の構成

Task	動画数	ペア数
Surgery(KnotTying)	36	596
Surgery(NeedlePassing)	28	362
Surgery(Suturing)	39	701
DoughRolling	33	181
ChopstickUsing	40	536
Drawing(Hand)	20	129
Drawing(Sonic)	20	118

本研究では，2 視点から撮影された動画を有する Drawing 及び ChopstickUsing のデータセットを使用する．これらのデータセットには，作業者の目線を撮影した動画及び作業場所を真上から撮影した動画を有する．本論文では，作業者の目線を撮影した動画を目線動画，作業場所を真上から撮影した動画を固定視点動画と定義する．目線動画の例を図 1 に，固定視点動画の例を図 2 に示す．

2.2　データセットの特徴

本節では，2 視点から撮影された動画を有する Drawing 及び ChopstickUsing のデータセットの特徴について述べる．Drawing は，予め描かれているイラストを見ながら模写する作業風景を撮影した動画のデータセットである．ここで，模写するイラストは人間の手及びキャラクターの顔の 2 種類である．表 1 において，人間の手を模写する動画は Drawing(Hand) に分類され，キャラクターの顔を模写する動画は Drawing(Sonic) に分類される．ChopstickUsing は，箸を使用して 4 個の豆を片方のプラスチック製の容器からもう片方の容器へ移す作業風景を撮影した動画のデータセットである．

図 1　目線動画の例 (動画中，容器等の位置が動的に変化する)

図 2　固定視点動画の例 (動画中，容器等の位置が変化しない)

2.3　データセットの分割

本研究における評価実験では，4 分割交差検証による学習を行う．Epic-Skills 2018 Dataset は，4 分割交差検証に対応した訓練データ及び検証データの分割が予め行われている．4 分割交差検証に対応する訓練データ及び検証データの内訳を表 2 に示す．ここで，分割した訓練データ及び検証データのグループを split1，split2，split3，split4 とする．

3　特徴量抽出

動画から空間的特徴量 (RGB 特徴量) 及び時間的特徴量 (Optical Flow 特徴量) を抽出するために，3 次元畳み込みニューラルネットワーク (3D Convolutional Neural Network：3D-CNN) を使用する．本研究では，Kinetics [9] データセットで事前学習済みの Inflated 3D ConvNets(I3D) [10] を使用する．

I3D へ入力するデータは動画を FFmpeg [11] を使用してフレーム単位で切り出した画像を 455×256 の解像度で縮小した画像を生成し，抽出する特徴量に応じて前処理を行ったものを使用する．本論文では，前述の方法で生成した画像をリサイズ画像，RGB 特徴量を抽出する際に I3D へ入力するデータを RGB Frame，Optical Flow 特徴量を抽出する際に I3D へ入力するデータを Optical Flow Frame と定義する．

RGB 特徴量を抽出する場合，リサイズ画像の中心を 224×224 の解像度でクロップした画像を生成する．これを RGB Frame とする．Optical Flow 特徴量を抽出する場合，時間的に連続する 2 枚のリサイズ画像を使用して Optical Flow を生成する．Optical Flow とは，2 枚以上の時間的に連続する画像の中で共通して写っている部分等から動作及びパターンが移動する方向を推定し，ベクトルにしたものである．Optical Flow

表 2　実験で使用するデータセットの構成

	split1		split2		split3		split4	
Task	Train	Valid	Train	Valid	Train	Valid	Train	Valid
ChopstickUsing	314	222	280	256	299	237	299	237
Drawing(Hand)	62	67	76	53	79	50	63	66
Drawing(Sonic)	65	53	64	54	69	49	64	54

の生成アルゴリズムは，Lucas-Kanade 法 [12]，$TV-L^1$ [13]，Franeback 法 [14] など多数存在する．本研究では，$TV-L^1$ を用いて Optical Flow を生成する．時間的に連続する 2 枚のリサイズ画像を使用して Optical Flow を生成した後，中心を 224×224 のサイズでクロップする．これを Optical Flow Frame とする．

特徴量抽出を行う際，RGB Frame 及び Optical Flow Frame を 20 分割し，その分割単位毎に I3D へ入力する．抽出する特徴量として，I3D の最後から 2 番目の層である AveragePooling3D 層の出力を使用し，それを Flatten 層及び全結合層により次元数を 1024 次元に変換した特徴量を使用する．したがって，1 つの動画から抽出される特徴量の次元数は 20×1024 次元となる．

4　ベースライン

本研究では，提案手法との性能を比較するためにベースラインを用意する．4.1 節では，ベースラインとする学習モデルについて述べる．4.2 では，学習モデルの中で構成されている MLP について述べる．

4.1　学習モデル

本研究では，図 3 に示す学習モデルをベースラインとする．ベースラインは，固定視点動画から得られる動画特徴量のみを使用した学習モデルである．学習モデルの入力は技能判定に使用する 2 つの固定視点動画であり，それぞれ固定視点動画 A 及び固定視点動画 B とする．

技能判定の流れは，最初に I3D を使用してそれぞれの固定視点動画から動画特徴量を抽出する．そして，抽出した動画特徴量を多層パーセプトロン (MLP) へ入力する．その後，Subtract 層で 2 つの MLP 出力値の差分を求め，その差分を活性化関数 tanh により $-1 \sim +1$ の範囲に収まるように正規化する．本論文では，RGB 側の活性化関数 tanh の出力を RGB Score，Optical Flow 側の活性化関数 tanh の出力を Flow Score と定義する．最後に，RGB Score 及び Flow Score の平均値を求める．この平均値が固定視点動画 A 及び固定視点動画 B に対して技能判定を行った結果とする．図 3 に示す学習モデルの場合，判定結果の値が正の数 (0 より大きい値) であれば固定視点動画 A の方が技能レベルが高いと判定され，負の数 (0 より小さい値) であれば固定視点動画 B のほうが技能レベルが高いと判定される．

4.2　MLP

MLP の構成を図 4 に示す．MLP の入力は I3D を使用して

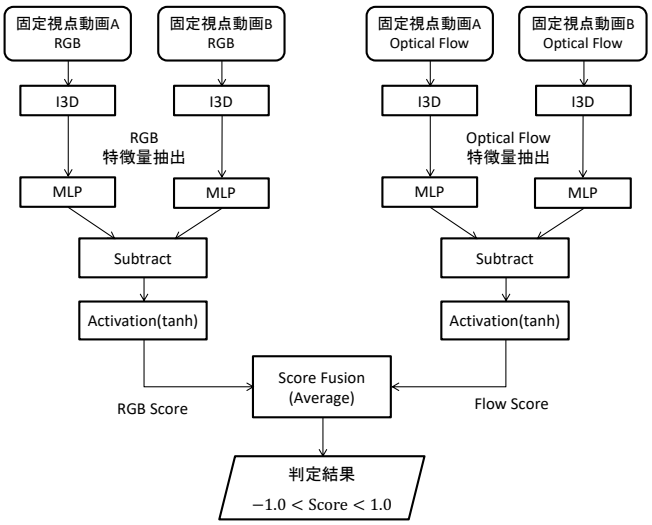


図 3　ベースラインの学習モデル

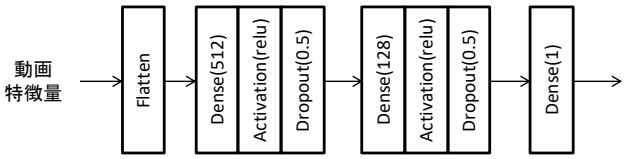


図 4　MLP の構成

抽出された動画特徴量とする．中間層は 2 層導入し，それぞれの次元数は 512 次元，128 次元とする．また，活性化関数は relu 関数を使用し，dropout 率は 0.5 とする．出力層の次元数は 1 次元であるため，MLP から出力される値は実数値となる．

5　提 案 手 法

本研究では，多視点で撮影された動画を用いた技能判定手法を 2 種類提案する．5.1 節では，提案手法 1 について述べる．5.2 節では，提案手法 2 について述べる．

5.1　提案手法 1

提案手法 1 の学習モデルを図 5 に示す．固定視点動画から得られる動画特徴量に加え，目線動画から得られる動画特徴量を使用する点がベースラインと異なる．技能判定に使用する 2 つの固定視点動画を固定視点動画 A 及び固定視点動画 B，技能判定に使用する 2 つの目線動画を目線動画 A 及び目線動画 B とする．これらの動画を学習モデルの入力とする．本論文では，固定視点動画及び目線動画から得られた動画特徴量を連結したものを入力して技能判定を行い，その結果を実数値で出力する部分を Comp-NN Model と定義する．

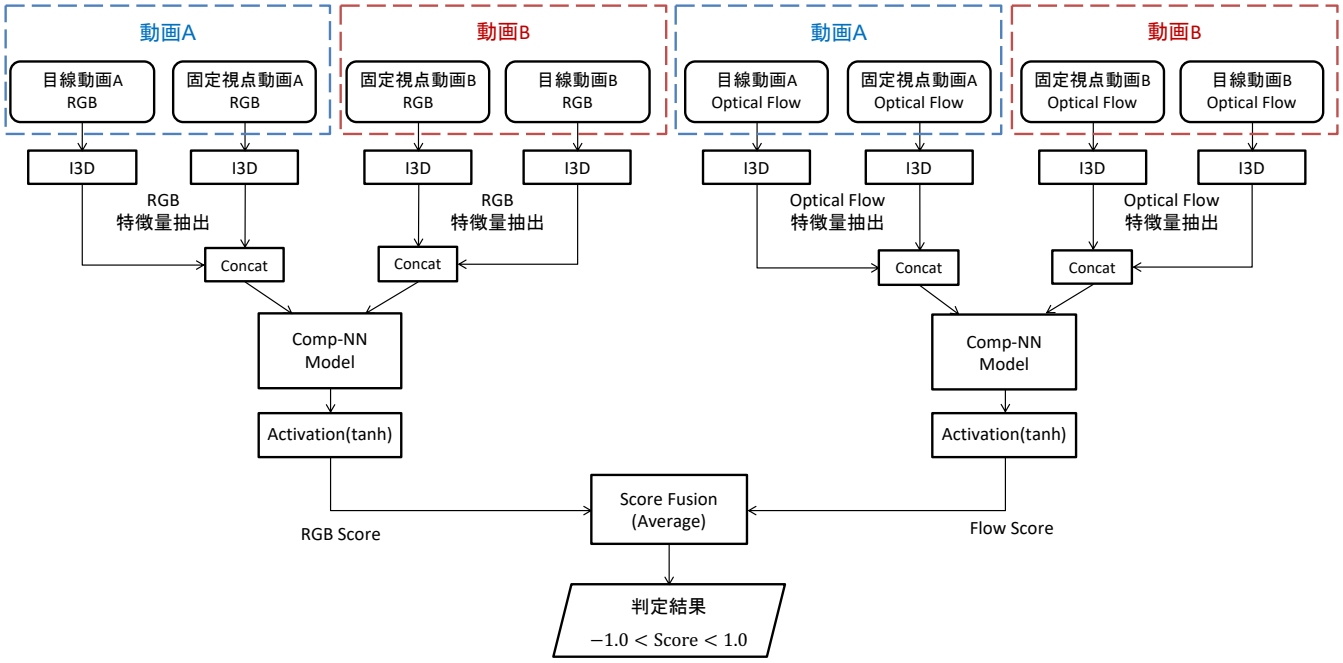


図 5　提案手法 1 の学習モデル

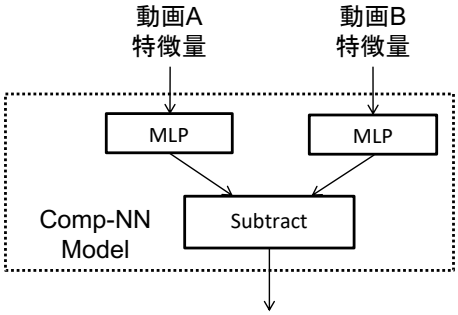


図 6　提案手法 1 の Comp-NN Model

技能判定の流れは，最初に I3D を使用して固定視点動画及び目線動画から動画特徴量を抽出する．その後，抽出した固定視点動画及び目線動画の動画特徴量を連結する．そして，連結した動画特徴量を Comp-NN Model へ入力し，その出力を活性化関数 tanh により $-1 \sim +1$ の範囲に収まるように正規化する．最後に，RGB Score 及び Flow Score の平均値を求める．この平均値が動画 A 及び動画 B に対して技能判定を行った結果となる．図 5 に示す学習モデルの場合，判定結果の値が正の数であれば動画 A の方が技能レベルが高いと判定され，負の数であれば動画 B のほうが技能レベルが高いと判定される．また，MLP は図 4 と同様の構成である．

図 6 に示すように，提案手法 1 の Comp-NN Model は 2 つの MLP 及びそれらの出力の差分を計算する Subtract 層で構成される．固定視点動画から得られる動画特徴量の次元数は 20×1024 次元，目線動画から得られる動画特徴量の次元数は 20×1024 次元であるため，Comp-NN Model へ入力する動画特徴量の次元数は 40×1024 次元となる．

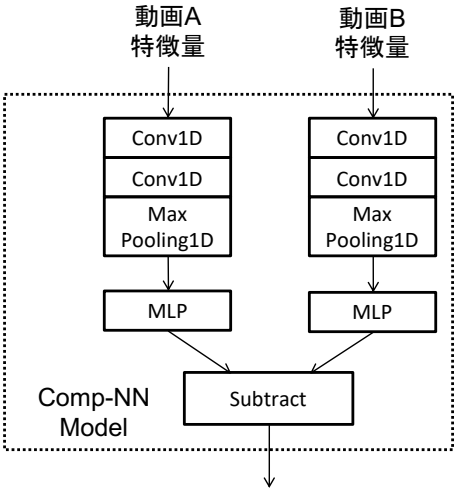


図 7　提案手法 2 の Comp-NN Model

5.2　提案手法 2

提案手法 2 では，I3D へ入力するデータを拡張し，それに対応した Comp-NN Model の構成を提案する．

5.2.1　データ拡張

特徴量抽出の前処理において，リサイズ画像の中心を 224×224 の解像度でクロップした画像の他にリサイズ画像の左側及び右側をそれぞれ 224×224 の解像度でクロップした画像も使用する．これら 3 種類のクロップ画像を I3D への入力データとして扱う．データ拡張を行った際の動画特徴量の次元数は 60×1024 次元となる．

5.2.2　Comp-NN Model の構成

提案手法 2 の Comp-NN Model を図 7 に示す．データ拡張を行うと固定視点動画から得られる動画特徴量の次元数は 60×1024 次元になるため，Comp-NN Model へ入力する動画

特徴量の次元数は 80×1024 次元となる．これにより学習モデルのパラメータ数が増加するため，提案手法 2 の Comp-NN Model では，パラメータ数を削減することを目的として 2 層の 1 次元畳み込み層 (Conv1D) 及び 1 層の 1 次元プーリング層 (MaxPooling1D) を MLP の直前に導入した構成とする．また，提案手法 2 の学習モデルは Comp-NN Model を除いて図 5 と同様の構成である．

6 評価実験

本研究では，提案手法の性能を評価するために，2 節で述べたデータセットを使用して評価実験を行った．6.1 節では，実験設定について述べる．6.2 節では，評価実験で性能を評価する学習モデルについて述べる．6.3 節では，評価実験の結果について述べる．6.4 節では，性能を評価したモデルを使用して技能判定結果を出力した例について述べる．

6.1 実験設定

6.1.1 学習パラメータ

評価実験で設定した学習パラメータを表 3 に示す．

表 3　実験設定

パラメータ	名称・設定値
最適化関数	Adam
学習率	1e-6
学習率減衰	1e-6
バッチサイズ	32
エポック数	100

6.1.2 損失関数

Margin Ranking Loss に基づき，学習モデルの出力に合わせて新たに定義した損失関数を使用した．Margin Ranking Loss は x_1, x_2 を損失関数の入力，正解ラベルを y，マージンを m とすると，以下の式 1 に示す損失関数 L_1 で表される．

$$L_1 = max(0, -y * (x_1 - x_2) + m) \tag{1}$$

また, [5] では以下の式 2 に示す損失関数 L_2 が使用されている．これは，損失関数 L_1(式 1) に $m = 1, y = 1$ を代入したものに相当する．x_1 には技能レベルが高い動画の技能判定スコアが入力され，x_2 には技能レベルが低い動画の技能判定スコアが入力される．

$$L_2 = max(0, 1 - (x_1 - x_2)) \tag{2}$$

これらの損失関数に基づき，以下の式 3 に示す損失関数 L_3 を定義した．ここで，x は損失関数の入力である．損失関数 L_2(式 2) における $x_1 - x_2$ の部分は，学習モデルにおける Subtract 層に相当するため，損失関数への入力は x のみである．

$$L_3 = max(0, 1 - x) \tag{3}$$

6.1.3 評価指標

式 4 で計算される割合を評価指標とした．本論文では，この割合を技能判定精度と定義する．

$$\text{技能判定精度} = \frac{\text{正しく技能判定された動画のペア数}}{\text{検証に使用した動画ペアの数}} \tag{4}$$

6.2 学習モデル

評価実験では，4 節で述べたベースラインの学習モデル及び 5 節で述べた提案手法の学習モデルを構築し，性能評価を行った．また，提案手法の一部を導入した場合における性能評価も行った．性能評価を行う学習モデルを表 4 に示す．ここで，提案手法の構成要素を①，②，③の 3 つに分ける．これらの説明を次に示す．

①：目線動画から得られる特徴量を使用
②：I3D へ入力するデータを拡張
③：1 次元畳み込み層及び 1 次元プーリング層を導入

表 4　性能評価を行う学習モデル

	構成要素		
	①	②	③
ベースライン			
提案手法 1	✓		
提案手法 2	✓	✓	✓
ベースライン + データ拡張		✓	
提案手法 1 + パラメータ削減	✓		✓

6.3 実験結果

実験結果を表 5 に示す．Drawing データセットを使用した場合では提案手法 1 が最も精度が高い結果となり，ChopstickUsing データセットを使用した場合では提案手法 2 が最も精度が高い結果となった．ベースラインの学習モデルにおける入力データを拡張すると，ChopstickUsing データセットを使用した場合では技能判定精度が 1.3%向上し，Drawing データセットを使用した場合では技能判定精度が 0.3%低下したことが確認された．したがって，データ拡張は ChopstickUsing データセット対しては有効であるが，Drawing データセットに対しては有効ではないと考えられる．提案手法 1 の構成に 1 次元畳み込み層及び 1 次元プーリング層を導入すると，ChopstickUsing データセットを使用した場合では技能判定精度は変化せず，Drawing データセットを使用した場合では技能判定精度が 1.9%低下したことが確認された．したがって，1 次元畳み込み層及び 1 次元プーリング層の導入は，学習モデルのパラメータ数を削減する場合には有効であるが，精度向上には有効ではないと考えられる．

6.4 技能判定結果例

性能を評価した学習モデルを使用して，技能判定結果の予測値を出力した．その例を表 6 に示す．表 6 及び後掲する表 7 において，2 つのペア動画 (A, B) に対する判定結果が正の値であった場合，動画 A の方が技能レベルが高いことを示す．また，判定結果が負の値であった場合，動画 B の方が技能レベル

表 5　実験結果

	技能判定精度 [%]	
	ChopstickUsing	Drawing
ベースライン	77.3	84.3
提案手法 1	78.7	**85.4**
提案手法 2	**79.7**	84.4
ベースライン + データ拡張	78.6	84.0
提案手法 1 + パラメータ削減	78.7	83.5

表 6　技能判定結果例

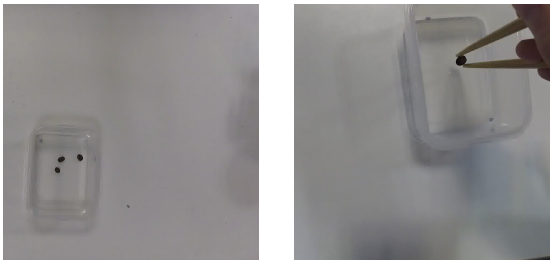
	ベースライン	提案手法 1	提案手法 2
判定結果	-0.266	-0.245	**0.197**

が高いことを示す．実際は動画 A の方が技能レベルが高いため，ベースライン及び提案手法 1 では誤った技能判定が行われていること，提案手法 2 では正しい技能判定が行われていることが確認できる．また，技能レベルが高い動画のクロップ画像を図 8 に，技能レベルが低い動画のクロップ画像を図 9 に示す．技能判定の学習及び予測については，これらのクロップ画像に映っている物体の位置及び連続する動画フレーム間の移動量等から得られた特徴量を使用している．

使用したクロップ画像と技能判定結果の関係に着目すると，ベースラインでは，固定視点動画の中心をクロップした画像のみを使用するため，箸で豆を掴もうとする動作の特徴量が得られていない．これにより，誤った技能判定が行われたと考えられる．これに対し提案手法 2 では，学習に使用する特徴量の種類を増やしたことにより箸で豆を掴もうとする動作の特徴量を得られるようになったため，正しい技能判定が行われたと考えられる．一方，提案手法 1 では，今回の技能判定結果例において誤った技能判定が行われたが，技能判定結果の予測値はベースラインよりも 0.21 増加した．したがって，目線動画から得られる特徴量を追加したことにより，ベースラインよりも技能判定に関する性能が向上したといえる．

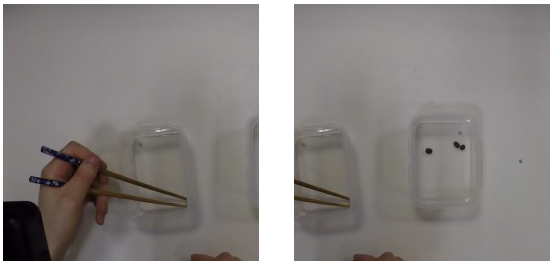
この他に，学習モデルへ入力するデータを反転した場合における技能判定結果の予測値を出力した．具体的には，固定視点動画 A の RGB Frame を入力する部分に固定視点動画 B の RGB Frame を入力し，固定視点動画 B の RGB Frame を入力する部分に固定視点動画 A の RGB Frame を入力した．また，固定視点動画の Optical Flow Frame 及び目線動画の RGB Frame，Optical Flow Frame においても同様の操作を行った．この操作により入力を反転した場合における技能判定結果の例を表 7 に示す．ここで，A1 及び G5 は技能判定の比較を行う動画の名前である．学習モデルへ入力するデータを反転した場合，判定結果の符号は反転するが，絶対値は変化しないことが確認できる．このことから，入力する方向に関係なく，技能レベルが高い動画はどちらの動画であるかを判定することが可能であるといえる．

(a) 固定視点動画 (左側)　(b) 固定視点動画 (中心)

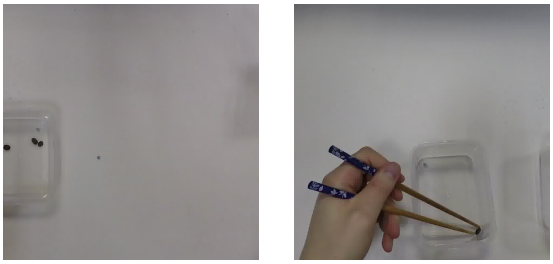


(c) 固定視点動画 (右側)　(d) 目線動画 (中心)

図 8　技能レベルが高い動画のクロップ画像



(a) 固定視点動画 (左側)　(b) 固定視点動画 (中心)



(c) 固定視点動画 (右側)　(d) 目線動画 (中心)

図 9　技能レベルが低い動画のクロップ画像

表 7　入力を反転した結果の例

入力		判定結果		
入力 A	入力 B	ベースライン	提案手法 1	提案手法 2
A1	G5	0.966	0.993	0.997
G5	A1	-0.966	-0.993	-0.997

7　おわりに

本研究では，多視点で撮影された動画を用いた技能判定を行う手法を提案した．具体的には，固定視点動画及び目線動画から得られる特徴量を用いて，それらを組み合わせた特徴量による技能判定を行う手法，I3D へ入力するデータを拡張し，それに対応した学習モデルを構成する手法を提案した．実験の結果，

提案した 2 つの手法はベースラインよりも精度が向上したことが確認できた．また，使用するデータセットによって最も精度が高くなる提案手法が異なることも確認できた．今後の課題として，他のデータセットを使用した場合における最適な技能判定手法の提案及びデータセットの拡張などが挙げられる．

謝　　辞

本研究の一部は，科研費基盤 (B) (課題番号 17H01746) の支援を受けて遂行した．

文　　献

[1] Paritosh Parmar and Brendan Tran Morris. Learning to Score Olympic Events. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015.

[3] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2, pp. 1735–1742. IEEE, 2006.

[4] Seong Tae Kim and Yong Man Ro. Evaluationnet: Can human skill be evaluated by deep networks? *arXiv preprint arXiv:1705.11077*, 2017.

[5] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, pp. 20–36. Springer, 2016.

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

[8] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The Pros and Cons: Rank-Aware Temporal Attention for Skill Determination in Long Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[10] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] FFmpeg. `https://www.ffmpeg.org/`.

[12] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[13] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pp. 214–223. Springer, 2007.

[14] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pp. 363–370. Springer, 2003.