骨格データと画像情報に基づく適応重み付き融合を用いた人間行動認識の精度改善

Koizumi, Kento / 小泉, 健人

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
17
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2022-03-24
(URL)
https://doi.org/10.15002/00025259

骨格データと画像情報に基づく 適応重み付き融合を用いた人間行動認識の精度改善

Improving Accuracy of Human Activity Recognition Using Adaptive Weighted Fusion Based on Skeleton Data and Image Information

小泉　健人
Kento Koizumi
法政大学大学院情報科学研究科情報科学専攻
E-mail: kento.koizumi.4q@stu.hosei.ac.jp

Abstract

Human Activity Recognition is to make a computer recognize human activity using data collected from video camera or devices such as Kinect. There has been active research on activity recognition models using skeleton data, which is less sensitive to background information and angle, and the overall accuracy of the models is improving year by year. However, some activities such as "reading" and "writing" are difficult to recognize, and the accuracy has not improved much. Therefore, in this research, for activities that are difficult to recognize with skeleton data alone, the image information including appearance information and objects used by the subject are fused. In addition, unlike existing fusion methods, we proposed a method that fuses the output of each deep learning model with different weights depending on the output results. The weights are predicted by a machine learning model that has been previously trained with weights calculated by the newton method. The experimental results show that the accuracy of difficult-to-recognize activities can be improved by fusing image information using the proposed fusion method. It is also shown that the overall accuracy of the proposed fusion method in this research is higher than the existing fusion methods.

1. まえがき

映像や Kinect 等のデバイスから取得した情報を用いて人間の行動を認識させることを人間行動認識と呼ばれている[1]．防犯カメラを用いた不審な行動の検出等での活用が期待されている[1, 2]．近年では GPU 性能の向上や大規模なデータセット登場により，深層学習を用いた行動認識モデルの研究が増えている．近年では RGB 映像に比べ背景情報の影響を受けない，RGB 映像を入力させて学習させる場合に比べ計算量コストが低くなる骨格データを用いた行動認識研究が行われている[1, 3]．入力に使用される骨格データは，kinect 等のデバイスから取得された大規模データセットを使用し行動認識モデルの全体精度の向上を目的とした研究がなされている．骨格データを用いた行動認識モデルの全体精度は年々向上しているが，認識が難しく精度が低い行動が存在する．例えば reading や writing といった手を使う行動の精度が他の行動と比べて低く，原因として骨格データはノイズが多いことや RGB 映像と比べ色や形といった情報が不足している可能性がある．したがって本研究では解決方法の 1 つとして骨格データに加え，被験者が行動に使用している物や外見情報を含む画像情報を RGB 映像から抽出して融合させる．また既存研究で使用されている融合手法[5, 6]では精度が低下してしまう等の問題がある為，各深層学習モデルの出力結果に応じて重みを変えて融合させる手法を提案する．行動予測段階で骨格データのみでは認識が難しい場合に画像情報を融合させることで認識に必要な情報を補い合うことが可能となり精度向上が期待できる．映像データから抽出した画像情報と本研究で提案した融合手法を用いることで骨格データのみ使用した場合に比べ，認識が難しい行動の精度が改善されることを確認した．また提案融合手法を用いることで従来の融合手法に比べ全体精度が向上することが判明した．

2. 関連研究

2.1. 骨格データを用いた行動認識

骨格データを用いた行動認識モデルの研究では，行動認識モデルの学習用データセットとして ROSE 研究室 Shahroudy らが作成した大規模データセット NTURGB+D[4]が用いられている研究が多い．骨格データを用いた行動認識の学習用モデルとして RNN，LSTM といった時系列を考慮できる深層学習モデルが用いられることが多かったが，近年では畳み込み処理時に周りの関節情報を考慮できる GCN を用いた行動認識モデルの研究がなされている[7]．またそれぞれの骨格データの座標 x,y,z を 0 から 255 に正規化し CNN に学習させるモデルや，角度，位置のパラメータを学習モデルに組み込むことにより最適化させる CNN モデルの研究[8]もされており従来の RNN，LSTM を用いた行動認識モデルに比べ高い全体精度をとなった．しかし reading，writing，typing on keyboard 等の一部の行動の精度が改善されていない．色や形の情報が少なく骨格データのみでは判別が難しいと考えられる．

Supervisor: Prof. Runhe Huang

2.2. 既存融合手法

既存の行動認識モデルの研究では複数データや特徴量を学習させる際に同時に学習させ特徴量を融合させる手法と別々に学習させテスト時に融合させる手法が存在する．前者の手法として，Rahmani らの研究[6]では Concatenate Layer とよばれる層でそれぞれの深層学習モデルから出力された特徴量を融合，Boissiere らの研究[9]では 2 つの深層学習の出力結果を Concatenate Layer で結合し MLP で融合していく Optimized Network を用いている．骨格データのみの手法に比べ精度は改善しているが，モデルによっては Concatenate Layer で特徴量を融合することにより過学習が発生することや，学習の際にメモリ等のリソースを多く消費してしまい学習できないことが問題となっている．別々に学習させテスト時に融合させる手法としては，各深層学習モデルの出力結果を足し合わせ平均化する Average Fusion と呼ばれる手法が使われている[5]．平均値を取ることで全体精度の向上が見られる行動認識モデルも存在したが，クラス別の精度では大きく精度が低下してしまう行動も存在した．Zhang らの研究[8]では各モデルの出力に対して重み付け融合を行い行動認識させるモデルとなっている．全体精度が高くなる重みを探索させることで平均値融合より高い精度が出る結果となった．Average Fusion や Weighted Fusion はすべての行動に対して同じ重みを用いて融合を行っているが，各モデルの予測結果に応じて適切な重み付けを行い融合させることでさらに精度が向上する可能性がある．

3. 提案手法

本研究では人物の外見や物などを含む画像情報を融合させ，さらに骨格情報の予測結果に応じて骨格データと画像情報の予測結果の重みを変える手法を提案する．

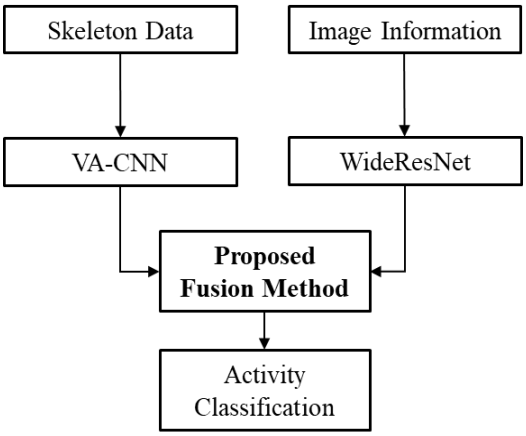

図 1　画像情報との融合モデル全体図

3.1. 提案融合手法について

本研究で提案する融合手法について説明する．事前に機械学習モデルに対して各深層学習モデルの行動予測結果を入力し，目的変数として重みを学習させる．重みは最急降下法よりも収束が早いニュートン法[10]と呼ばれる手法を用いて各モデルの行動予測結果の重みを loss 関数が 0 になるように計算させる．ニュートン法を用いることで行動予測結果ごとに正しく行動認識できるような重みの計算を可能となる．ニュートン法の計算式を(1), (2)に示す．Eは骨格データを用いた行動認識モデルと画像を用いた行動認識モデルの loss 関数，$W = (w_{1,k}, w_{2,k})$は重みの値，ηは値を動かす幅を決めるパラメータを示している．pは出力層の値が正解データの出力だった場合のみ 1 を出力しそれ以外は 0 を出力する関数である．y_1は骨格データを用いた深層学習モデルの出力，y_2は画像を用いた深層学習モデルの出力となっている．kは反復回数を表す．(1)式を用いて重みを計算していき，$|w_{1,k+1}\text{-}w_{1,k}|$が 1e-2 未満かつ$|w_{2,k+1}\text{-}w_{2,k}|$が 1e-2 未満，もしくは loss 関数である(2)式が 0 以下となった場合，計算を終了させる．

$$W_{k+1} = W_k - \eta \frac{\partial E}{\partial W} \tag{1}$$

$$E = -\sum_{class} p(class) \log\left(w_{1,k} * y_1 + w_{2,k} * y_2\right) \tag{2}$$

各深層学習モデルから出力されたi番目の行動予測結果$y_{1,i}, y_{2,i}$をもとに学習させた機械学習に loss が低くなる重み$w_{1,i}, w_{2,i}$を予測させ，2 つの深層学習モデルの出力結果に対して重み付き融合させたものをもとに最終的な予測結果を出力させる．

$$y_i^{fusion} = w_{1,i} * y_{1,i} + w_{2,i} * y_{2,i} \tag{3}$$

正しく行動認識できるような重みを機械学習に予測結果と一緒に学習させることで，機械学習からは各深層学習モデルの行動予測結果に応じて正しく認識できる重みの予測が出力される．いくつか認識が難しい行動に関しては画像情報を学習させたモデルの精度が高い為，予測結果で認識が難しいとなった場合，画像情報の重みが大きくなることが予想される．これにより今まで正しく認識できていなかった精度の低い行動について精度改善が期待できる．提案融合手法で使用する機械学習モデルは，非線形の回帰を行えるニューラルネットワーク，ランダムフォレスト，サポートベクター回帰(SVR)の 3 つの機械学習モデルを用いて，全体精度比較を行い全体精度が一番高い機械学習モデルを本研究で使用する．

3.2. 行動認識モデル

本研究で使用する行動認識モデルの全体モデル図を図 1 に示す．骨格データは従来使われてきた LSTM Model に比べ学習が早く全体精度も向上している Zhang らが提案した VA-CNN モデル[8]を使用する．まず既存研究[8]に従い kinect で取得されたtフレーム目の関節番号jの骨格データ$v_{t,j}$に対して最初に RGB 画像に変換する．各フレームの骨格データに含まれる 3 次元座標x, y, zを 0 から 255 に(4)式を用いて正規化し，縦を時系列順，横を骨格データ

の関節番号順に並べられた 3 チャンネル分行列を生成し VA-CNN モデル[8]に学習させる．c_{min}, c_{max}はデータセットに含まれる骨格データの最大座標と最小座標を示す．

$$v'_{t,j} = \left(255 \times \frac{v_{t,j} - c_{min}}{c_{max} - c_{min}}\right) \quad (4)$$

画像情報は ResNet の学習時間と精度を改善させた WideResNet[11]呼ばれる CNN モデルに学習させる．学習させた各深層学習モデルに骨格データと画像データを入力すると，出力結果として 60 クラス分の予測結果が出力される．予測結果に対して本研究で提案した融合手法を用いて予測結果を融合させ，最終的な行動予測結果を出力させる．

4. 実験

4.1. 実験概要

前実験として 3 つの機械学習モデル，最急降下法とニュートン法の 2 つの重みの計算手法，計 6 パターンの全体精度を調査する．全体精度が一番高いモデルを本研究の提案手法で使用する機械学習モデル，目的変数である重みの計算手法として使用する．次に融合モデルの全体精度と NTU RGB+D に含まれる行動の中で認識率が特に低い行動の精度を調査し，骨格データ学習モデルや画像情報を用いた学習モデルとの比較を行う．また既存の研究で使用されている融合手法を本研究のモデルに実装した場合の全体比較と認識が難しい行動の精度比較を行い本手法の有効性について検証を行う．

4.2. 実験設定

最初に骨格データと画像情報の学習させるモデルのハイパーパラメータについて説明する．VA-CNN と WideResNet50-2 の各ハイパーパラメータは既存研究に従い設定を行う．Epochs 数は 100，学習率は 1e-4 とし，5 回精度が低下しない場合学習率を 0.1 倍させる．loss 関数はクロスエントロピー誤差関数を用いる．学習の際に使用する最適化アルゴリズムは Adam を用いる．学習時に loss が 10 回連続で低下しない場合，精度低下を防ぐ為に早期停止を行う．検証用データは学習用データセットの 5%とする．次に融合手法で使用する各機械学習モデルのハイパーパラメータについて説明する．各機械学習モデルのハイパーパラメータは，Grid Search を用いて誤差が低くなる値を用いる． SVR のカーネルは RBF，正則化係数 C は 1，不感損失関数は 0.001 とする．ランダムフォレストの max_depth は 3， n_estimators は 100 とする．ニューラルネットワークの Epochs 数は 100，学習率は 1e-3 とし，損失関数は平均二乗誤差を用いる．ニューラルネットワークは 3 層モデルを使用する．計算ニューラルネットワークの学習の際に使用する最適化アルゴリズムは Adam を用いる．学習に使用する特徴量は MIC[14]と呼ばれる非線形な相関を解析できる手法を用いて特徴量選択を行う．

4.3. 学習用データセット

本研究では Shahroudy らが開発した NTU RGB+D[4]と呼ばれるデータセットを用いる．データセットの概要を表 1 に示す．データセットには人間の関節の位置を表す骨格データに加え RGB 映像， 深度映像，IR 映像があり，サンプル数は 56880 となっている．各データは Microsoft が開発したモーションキャプチャーkinect で取得されており，データセットは 2016 年に作成されている． データセットには drinking water や eat meal/snack，reading や writing 等の日常の行動，kicking something，cheer up 等の運動，sneezing, staggering, falling down 等の健康やけがに関する行動，kicking，pushing，pushing 等の不審な行動が含まれている．年齢が 10~35 歳の 40 人の被験者を対象に撮影されており，カメラの角度に関しては正面と横，斜め 45 度の 3 視点から撮影されたものがそれぞれ含まれている．関節は全部で 25 関節あり，骨格データに関しては x, y, zの 3 軸分の座標となっている．

表 1　NTU RGB+D データセットの概要

Number of samples	56880
Number of classes	60
Number of testers	40
Number of Joints	25
Date of Year	2016

4.4. データの前処理

実験で使用するデータの前処理について説明する．画像データは Wang ら[12]の研究を参考に映像データから数フレーム切り出し学習に使用する．切り出すフレーム数は既存研究で一番精度が高い 5 フレームとする．フレームの抽出方法として既存研究を参考に RGB 映像データから 1 つの動画データの総時間を 5 等分しそれぞれ 1 フレームずつ合計 5 フレーム分画像を切り出す．画像には背景情報が含まれているが，背景を含めた全体の画像を使用すると行動認識時に背景情報の影響を受けてしまう問題[3]が判明している為，Meta AI(旧 Facebook AI)が開発した detectron2[13]と呼ばれるソフトウェアに含まれる人物検出を使用し，被験者の画像領域を抽出して学習に使用する．

4.5. 評価方法

評価方法は，NTURGB＋D データセットの論文で推奨されている Cross Subject (CS)，Cross View (CV)と呼ばれる評価方法を用いる．CS は 40 人の被験者のうち 20 人を学習用のグループ，残りの 20 人をテスト用として精度を評価する方法である．角度は正面と左右，斜めから撮影されたものが含まれるが CS での表ではすべての角度から撮影されたデータを使用する．学習させていな被験者の行動に対しても正しく認識できるかモデルの汎用性を調査することが目的である．CV は正面と左右から撮影されたものを訓練データとし，斜め 45 度から撮影されたものをテストデータとして精度の評価を行う．CV での評価は被験者 40 人すべての行動データを使用する．学習さ

せていない角度からの行動に対して正しく認識できるか汎用性を評価する．

5. 実験結果と考察

5.1.前実験の結果と考察

前実験の実験結果を表 2 に示す．Optimized Methods は重みの最適化手法，Machine Learning は重みを学習させた機械学習を示している．CS， CV 共にニュートン法で目的変数である重みを計算させ SVR に重みを事前に学習させ融合させた場合，他の手法と比べ精度が高くなる結果となった．

表 2　各機械学習モデル使用時の行動認識精度比較

Optimized Methods	Machine Learning	CS	CV
Newton Method	**SVR**	**91.9%**	**95.9%**
	Random Forest	91.7%	95.5%
	Neural Network	91.5%	95.5%
Gradient Descent	SVR	91.8%	95.7%
	Random Forest	91.7%	95.7%
	Neural Network	91.4%	95.6%

ニュートン法を使用して重みを計算した中で SVR が一番高い精度となった．最急降下法に比べニュートン法の収束が早く，loss が低くなる重みを計算できていると考えられる．また SVR は最適化すべきハイパーパラメータの数がランダムフォレストやニューラルネットワークに比べ少なく，精度が高くなるパラメータを見つけられやすい為，他の手法に比べ精度が高くなったのではないかと推測される．表 2 の実験結果から本研究では融合手法で重みを計算する際に SVR，重みの計算手法としてニュートン法を使用する．

5.2. 提案手法を用いたモデルの精度結果と考察

骨格データを VA-CNN に学習させたモデルと画像情報を WideResNet50-2 に学習させたモデル，提案融合手法を用いたマルチモーダルな行動認識モデルの全体精度を表 3 に示す． CS では提案手法の全体精度が 91.9%となり他の 2 つの既存モデルに比べ全体精度が向上した．CV でも同様に提案手法の全体精度が 95.9%で向上した．次に NTU RGB+D データセットの中で特に精度が低かった 3 つの行動の精度調査を行った．

表 3　行動認識モデル全体精度比較

	CS	CV
Skeleton Data Only [8]	87.9%	93.7%
Image Information Only [11]	63.5%	61.2%
Skeleton+Image (Proposed Fusion)	**91.9%**	**95.9%**

次に NTU RGB+D データセットの中で特に精度が低かった 3 つの行動の精度調査を行った．認識が難しい行動精度比較結果を CS，CV 順に図 2，図 3 に示す．CS では骨格データで認識が難しかった reading，writing，typing on keyboard はそれぞれ 66.7%，72.4%，90.5%となり骨格データを学習させたモデルや画像情報のみを学習させたモデルに比べ精度が大幅に改善される結果となった．CV についても同様に reading に関しては 78.4%，writing は 80.0%，typing on keyboard は 90.2%と他の学習モデルに比べ精度改善が見られた．

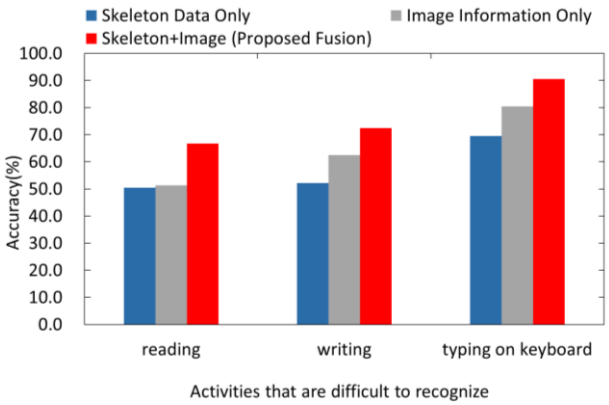

図 2　認識が難しい行動精度比較(CS)

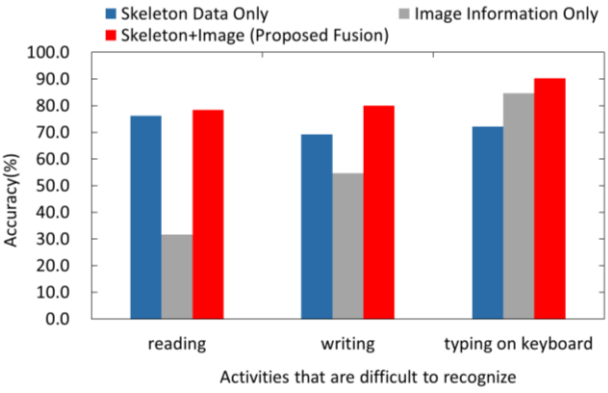

図 3　認識が難しい行動精度比較(CV)

表 3 から骨格データだけでなく被験者の画像情報を加えることで CS，CV 共に全体精度が向上することが判明した．CS が 4.0%上昇と CV よりも精度上昇率が高く，骨格データだけでは学習させていない被験者の動きの認識に限界があったが，画像情報を加えることで被験者の外見情報や行動に使用している物の情報が加わり，学習させていない未知の被験者の行動に対しても汎化性能が高まったと推測される．図 2，図 3 から reading や writing といった認識が難しい行動に関しても精度が向上することが判明した．特に reading，writing は手先以外の動きがほとんど無く姿勢も似ており，被験者によっては本を読む際に立ったりしたり手の伸ばし方も異なる為，骨格データだけではとくにこの 2 つの区別が難しいことが既存研究から判明しており，画像情報を融合することで reading や writing に使用するオブジェクトの情報が認識しやすくしているのではないかと考えられる．

5.3. 既存融合手法との精度比較と考察

既存行動認識で使用されている融合手法を本研究で実装したモデルに使用した場合の全体精度比較と，クラス別の精度比較を CS，CV 評価順に行った．まず CS，CV の全体精度比較結果を表 4 に示す．Concatenate Fusion は，CS が 75.1%，CV が 79.6%と骨格データを学習させたモデルに比べ全体精度が大幅に低下した．Optimized Network に関しても同様に CS が 77.7%，CV が 83.4%で全体精度が低下する結果となった．Average Fusion は CS では 90.7%と上昇が見られたが CV は 93.2%と全体精度が低下した．Weighted Fusion では CS が，91.7%，CV は 95.7%の精度となり骨格データのみを学習させたモデルに比べ全体精度が向上した．提案融合手法に関しては CS の全体精度 91.9%，CV の全体精度は 95.9%と他の融合手法に比べ一番精度が向上する結果となった．

表 4　既存融合手法との全体精度比較

	CS	CV
Concatenate Fusion [6]	75.1%	79.6%
Optimized Network [9]	77.7%	83.4%
Average Fusion [5]	90.7%	93.2%
Weighted Fusion [8]	91.7%	95.7%
Proposed Fusion	**91.9%**	**95.9%**

次にクラス別の精度について調査した． NTU RGB+D データセットの中で特に認識が難しい行動の精度を CS，CV 順に図 4，図 5 に示す．学習モデルを同時に学習させ融合させる Concatenate Fusion については reading，writing に関して CS，CV 共に低い精度となっており骨格データを学習させたモデルよりも精度が低下している．同様に Optimized Network でも CS，CV で 3 種類の行動の精度が低下している．別々に学習させ融合させる融合手法である Average Fusion，Weighted Fusion，提案融合手法ではどの手法もほぼ全ての行動が精度向上した．本研究で提案した融合手法は CS では reading，typing on keyboard に関してそれぞれ 66.7%，90.5%と Average Fusion，Weighted Fusion とほぼ変わらない精度を得た．writing に関しては Average Fusion の精度が 76.5%で高い結果となった． CV では reading，writing に関しては提案融合手法がそれぞれ 78.4%，80.0%と Average Fusion，Weighted Fusion と比べ精度が高くなる結果を得た．typing on keyboard に関しては Average Fusion が 96.2%と一番高い結果となった．

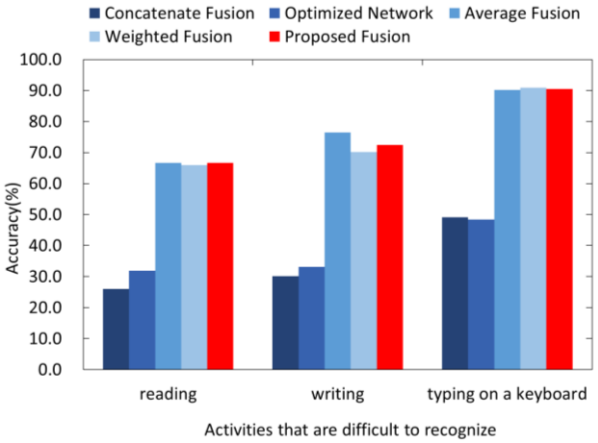

図 4　既存融合手法とのクラス別精度比較 (CS)

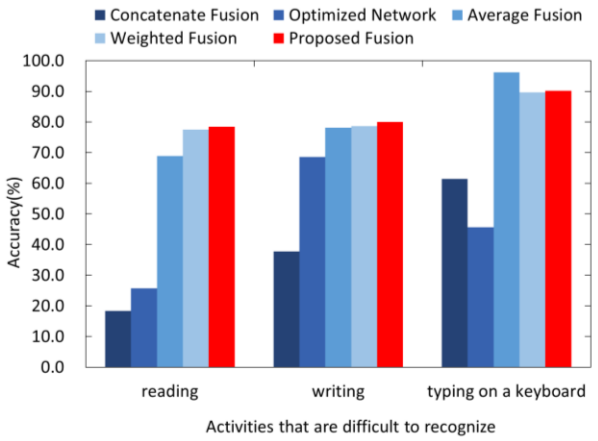

図 5　既存融合手法とのクラス別精度比較 (CV)

表 4 から Concatenate Fusion や Optimized Network といった学習を同時に行い，特徴量を融合させる融合手法では全体精度が CS，CV 共に大幅に低下しており，原因の 1 つとして過学習になっていると考えられる．行動認識で使用される深層学習モデルは，骨格データを学習させている VA-CNN の場合は View Adaptive Subnetwork と呼ばれる前処理を最適化するネットワークが繋がっておりパラメータ数が多い．WideResNet50-2 も中間層のパラメータ数が多く複雑なモデルとなっている為，より過学習を起こしやすいと推測される．そのため各モデルを同時に学習させる場合，各融合する 2 つの深層学習モデルをネットワークの少ないものにする必要があると考えられる．各モデルを別々に学習させ出力層を融合させ判断を行う融合手法では全体的に精度向上が見られた．Average Fusion は，CS は精度向上が見られたが CV では精度が低下する結果となった．Average Fusion は各モデルの重みを 1 にして出力結果を判別する手法な為，モデルのごとの精度や各モデルの精度が考慮されておらず，全体精度が低下したと推測される．Weighted Fusion を用いることで全体精度が向上，さらに各モデルの認識しやすい行動と認識しにくい行動を考慮した提案融合手法では Weighted Fusion よりも全体精度が向上することが判明した．提案手法で述べたように，全ての行動予測結果に対して同じ重みを用いるのではなく，各深層学習モデルの行動予測結果に応じて適切な重み付けを行うことでより正しく認識しやすくできていると考えられる．図 4，図 5 より認識が難しい行動に関しても一部の行動を除く行動に関して Average Fusion や Weighted Fusion を上回る結果となった．CV の typing on keyboard では Average Fusion を用いたモデルの精度が一番高い結果となった．表 3 の全体精度比較や図 2 図 3 のクラス別精度比較結果より，画像情報モデルの精度が高い結果ではない．画像情報を学習させているモデルの改良が必要と考えられる．提案融合手法に関して一部の行動が適切な重み付けが出来ていない可能性がある．融合手法に用いている機械学習モデルのハイパーパラメータ調整や，学習時に目的変数の重みを計算させている手法を他のアルゴリズムを用いて精度を検証していく必要がある．

6. おわりに

本研究では骨格データのみでは認識が難しい行動の精度改善を目的として，骨格データだけでなく画像情報を融合させ，さらに適応重み付き融合手法を提案した．NTU RGB+D データセットを用いて実験を行った結果，表 3 より提案手法を用いて画像情報を融合させ，行動認識させることで，全体精度が向上することが判明した．図 2，図 3 のクラス別精度結果より従来の研究で精度が低かった行動の精度改善が見られた．また表 4 より既存の融合手法に比べ全体精度が向上することが判明した．パラメータを変更することや重みの計算方法を他のアルゴリズムを使用して精度の調査を行っていくことである．また認識が難しい行動のさらなる精度改善の為に，他の深層学習モデルを使用して画像情報を学習させたモデルの精度改善が必要と考える．加えて本研究では NTU RGB+D を使用したが，今後は他のデータセットを使用した際の精度検証を行っていく．

文　献

[1] Z. Sun, J. Liu, Q. Ke and H. Rahmani, “Human Action Recognition from Various Data Modalities: A Review,” *arXiv preprint arXiv: 2012.11866*, 2020.

[2] F. Rezazadegan, S. Shirazi, B. Upcroft and M. Milford, “Action recognition: From static datasets to moving robots,” *Proc. 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3185-3191, 2017.

[3] Y. He, S. Shirakabe, Y. Satoh and H. Kataoka, “Human Action Recognition without Human,” *Proc. European Conference on Computer Vision Workshops (ECCV)*, pp. 11-17, 2016.

[4] A. Shahroudy, J. Liu, T. -T. Ng and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Action Analysis,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010-1019, 2016.

[5] H. Wang and L. Wang. “Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 499-508, 2017.

[6] F. Baradel, C. Wolf and J. Mille, “Human action recognition: Pose-based attention draws focus to hands,” *Proc. IEEE International Conference on Computer Vision Workshops (ICCV)*, p. 604-613, 2017.

[7] S. Yan, Y. Xiong and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” *Proc. AAAI Conference on Artificial Intelligence*, pp. 7444-7452, 2018.

[8] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue and N. Zheng, “View adaptive neural networks for high performance skeleton-based human action recognition,” *IEEE Trans. pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1963-1978, 2019.

[9] A. M. de Boissiere and R. Noumeir, “Infrared and 3d skeleton feature fusion for rgb-d action recognition,” *arXiv preprint arXiv:2002.12886*, 2020.

[10] H. H. Tan and K. H. Lim, “Review of second-order optimization techniques in artificial neural networks backpropagation,” *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1, pp. 012003, 2019.

[11] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.

[12] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. Van Gool, “Temporal segment networks for action recognition in videos,” *IEEE Trans. pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740-2755, 2018.

[13] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo and R. Girshick, “Detectron2,” https://github.com/facebookresearch/detectron2, 2019.

[14] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, “Detecting novel associations in large data sets,” *Science*, vol. 334, no. 6062, pp. 1518-1524, 2011.