様式6

申請先学部 工 学部 採択番号 No. 1

# 令和2年度大阪大学未来基金「学部学生による自主研究奨励事業」研究成果報告書

| | | | | | |
|---|---|---|---|---|---|
| ふりがな<br>氏名 | わだ てつや<br>和田 哲也 | 学部<br>学科 | 工学部<br>電子情報工学科 | 学年 | 3年 |
| ふりがな<br>共同<br>研究者氏名 | たかしま かずき<br>高嶋 和毅 | 学部<br>学科 | 工学部<br>電子情報工学科 | 学年 | 3年 |
| | しんざき よしお<br>新崎 義峰 | | 工学部<br>電子情報工学科 | | 3年 |
| | チャラン バブプリータ プラタプ<br>Charan Bhavpreeta Pratap | | 工学部<br>電子情報工学科 | | 3年 |
| アドバイザー教員<br>氏名 | やぎ やすし<br>八木 康史 | 所属 | 産業科学研究所 | | |
| 研究課題名 | ディープラーニングを用いた空手の型競技の上手さ判定 | | | | |
| 研究成果の概要 | 研究目的、研究計画、研究方法、研究経過、研究成果等について記述すること。必要に応じて用紙を追加してもよい。(先行する研究を引用する場合は、「阪大生のためのアカデミックライティング入門」に従い、盗作剽窃にならないように引用部分を明示し文末に参考文献リストをつけること。) | | | | |

## 研究目的

空手の型は来年の東京オリンピックの種目であるが、初心者には非常に分かりにくいスポーツである。目の肥えていない素人には「なぜ勝ったのか、負けたのか」「どこで勝負が別れたのか」などが分からず、スポーツとしての盛り上がりに欠けている。そこで、型競技の「上手さ」を可視化して誰でも空手の型競技を楽しめるようにしたいと以前から考えていた。

型の「上手さ」可視化するために、コンピューターが型の「上手さ」を認識できる必要がある。

また、そもそも型とは個別の空手技を決まった順番で繰り出してゆく競技である。そこで本研究では、型全体の上手さを推定するためのステップとして、技の動画を読み込み、その「上手さ」を点数で推定することを目的とする。

## 研究計画

一般的に空手の型は全体で３０秒から９０秒ほどの長さがあり、１～３秒程度の個別の技が連なったものとして考えることができる。そこで、型の上手さを評価するというタスクを「型の動画を技ごとに分割する」「個別の技の上手さを評価する」という２ステップに分けて考える。

### 去年の自主研究で作成済みの内容

・空手の型のデータセット

空手教室の生徒 15 人及び本研究の代表者１人が流派「林拳空手道」の「少林拳１」という型を演じる様子を、選手を取り囲むように 21 方向から撮影した。カメラの解像度は 640×480 である。選手１人につき少林拳１の型を合計４回演じてもらった。様々な上手さのデータを作るために、そのうち２回はわざと手を抜いて演じてもらった。カメラの不具合等によりデータの一部が欠損したため、最終的に得られたデータはのべ 63 人の 17 視点からの映像となった。

### 今年取り組んだこと

・データセットのラベル付け
機会学習を行うためには、撮影した型の映像について、それぞれどの程度の上手さかを表すラベルを付与する必要がある。そこで、流派「少林拳空手道」の師範1名に、データセットのラベル付けを行ってもらった。
・C3D モデルで上手さを認識
上記のデータセットをニューラルネットワークに学習させ、その後、型に含まれる個別の技の上手さを推定した。型の動画を技ごとに分割する作業は将来的に取り組むものとし、本研究ではすでに技ごとに分割された動画に対して上手さを推定することに取り組んだ。近年、行動認識では、動きに関する特徴と物体や見た目に関する情報を同時に特徴化することができる3次元の畳み込みニューラルネットワーク(3DCNN)による手法が注目されている。その中でも、シンプルなアーキテクチャで理解しやすいモデルとして本研究ではC3D[1]を利用した。

**今後取り組みたいこと**
本研究は型競技の上手さを認識するためのステップとして取り組んだ。今後、技の区切りの認識や、Grad-CAM[2]などの CNN に対する可視化技術を利用して上手さの可視化を行うことを検討している。また、今回取り組んだ内容についても認識精度をより向上させるために、オプティカルフローを利用した2ストリーム C3D[3]を導入することを検討している。

**研究方法**
・データセットのラベル付け
流派「少林拳空手道」の師範1名に、データセットのラベル付けを行ってもらった。本研究では型の動画を技ごとに分割し、技の動画について上手さを学習・推定した。そのため、型全体の上手さではなく、技ごとに上手さを表すラベルを付与する必要があった。ラベルづけを行う師範には、各選手について全 17 視点からの型全体の映像を見てもらったのち、型に含まれる5つの技についてそれぞれ上手さを表すラベルを付与してもらった。
本来の空手の試合では審判が 7人おり、各審判は「テクニカルパフォーマンス」と「アスレチックパフォーマンス」の2項目についてそれぞれ 10 点満点で小数点一桁までの点数をつける。前者は姿勢・美しさ・正確さなどに関する項目であり、後者は力強さやスピードに関する項目である。それぞれの項目に 7人分の点数が付けられるが、そのうち点数が大きい方から2つ、小さい方から2つを取り除き、残った 3人分の点数を合計する。各項目に 30 点満点で得点がつくが、「テクニカルパフォーマンス」の点数（以下 TCH スコア）は 0.7 倍、「アスレチックパフォーマンス」の点数（以下 ATH スコア）は 0.3 倍し、それら2つの点数を合計して 30 点満点の得点となる。
本研究では実際の空手競技の評価手順にならい、2つの評価項目について各 10 点満点で 0.1 点刻みのラベルをつけた。さらに、各技についての総合点（以下 TOTAL スコア）を TCH スコアの 0.7 倍と ATH スコアの 0.3 倍の和として計算した。この際、ラベルづけする師範には選手が全力で演じているのか、手を抜いて演じているのかはえないことによって、全力か手を抜いているかには関係なく、型の上手さにのみ注目してラベルをつけた。

・C3D モデルで上手さを認識
上記のデータセットを学習させ、型に含まれる個別の技の上手さを推定した。すでに技ごとに分割された動画に対して上手さを推定することに取り組んだ。

C3D とは３次元の畳み込みを行うニューラルネットワークのモデルである。前半の特徴抽出器は８層の畳み込み層と５層のプーリング層の組み合わせで、後半は全結合層である。ネットワーク図を以下に示す。

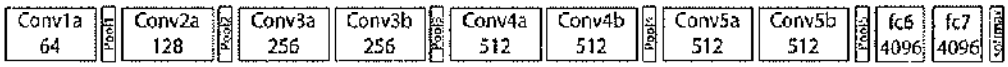



図１　C3D のネットワーク図

本研究で扱う空手型のデータセットは量が少ないため十分に学習できない可能性があった。そこで特徴抽出器のパラメータは Sport1M[4]データセットで学習済みのモデルのパラメータをコピーして使用し、学習の際は後半の全結合層のパラメータのみを更新した。10 点のラベルを 0-1 に正規化したものを真値として、C3D を利用し回帰問題として技の「上手さ」を学習した。データセットの数が少ないため、認識精度を評価する際は 16人の被験者で leave on esubject out の交差検証を行った。

## 経過

研究メンバー４人で作業を分担し、それぞれ得意な内容を担当した。作業内容は、本研究に関連する先行研究の調査や、学習に使用するデータの前処理、C3D の実装作業などである。C3D の実装についてはインターネット上に公開されているプログラムを参考にしたが、入力するデータやプログラムを実行する環境などが原因で様々なエラーが発生し、正常に学習できるようになるまでかなりの労力を要した。

２週間に一度、研究メンバー全員と指導教員の村松先生でミーティングを行い、研究の進め方や実験結果の考察について議論した。

## 研究成果

・最適なモデル構成

まず、本研究の問題設定に最も適したネットワーク構成を検討するために、全結合層について、層の深さや各層の素子数を様々に変更したネットワークを用意した。技２のカメラ１からの映像を用い、被験者 4, 14 をテストデータ、それ以外を訓練データとして学習・推定を行った。その後、各ネットワークの推定値について正解値からの平均二乗誤差を計算した。なお、バッチサイズは使用した計算機が対応できる最大値の 64 とし、エポック数は 100 とした。結果を表１に示す。

表１　様々なネットワーク構成における技の上手さの正解値からの推定値の平均二乗誤差

| fc6 | fc7 | 平均二乗誤差 |
|---|---|---|
| 32 | 32 | 0.67225025 |
| 64 | 32 | 0.45799172 |
| 128 | 32 | 0.59152805 |
| 64 | 64 | 0.44930709 |
| 128 | 64 | 0.49674745 |

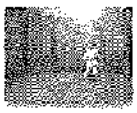
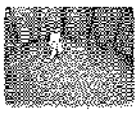
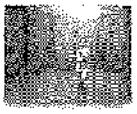
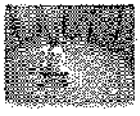
| | | |
|---|---|---|
| 256 | 64 | 0.49126252 |
| 128 | 128 | 0.46261298 |
| 256 | 128 | 0.49477157 |
| 256 | 256 | 0.47834114 |
| 512 | 256 | 0.46333839 |
| 512 | 512 | 0.40679331 |
| 1024 | 512 | 0.43310637 |
| 1024 | 1024 | 0.46810058 |
| 4096 | 4096 | 0.41332422 |

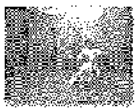
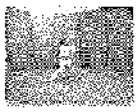
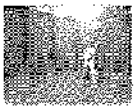
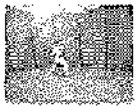
fc6 層、fc7 層の素子数がそれぞれ 512 のネットワークにおいて最も平均二乗誤差が小さくなった。以降ではこのネットワーク構成で学習を行った。

・最適な視点

次に、入力する動画の視点によって推定精度がどの程度変化するか確かめるために、各技について５つの視点の動画で学習・推定を行った。これらの学習・推定においては、技と視点以外の条件は全て一致している。その後、各条件における推定値について、正解値からの平均二乗誤差を計算した。被験者１をテストデータ、それ以外を訓練データとし、ラベルは tch score を使用した。結果を表２に示す。

表２　各技の各視点における推定値の正解値からの平均二乗誤差

| | | カメラ 10 | カメラ 16 | カメラ 18 | カメラ 20 | カメラ 21 |
|---|---|---|---|---|---|---|
| 技 1 | 入力動画の一例 |  |  |  |  |  |
| | 平均二乗誤差 | 0.016188647 | 0.204276282 | 0.150594432 | 0.195018102 | 0.067065429 |
| 技 2 | 入力動画の一例 |  |  |  |  |  |
| | 平均二乗誤差 | 0.23035774 | 0.236131906 | 0.219709996 | 0.27113468 | 0.172383575 |
| 技 3 | 入力動画の一例 |  |  |  |  |  |
| | 平均二乗誤差 | 0.194904524 | 0.258880139 | 0.273030376 | 0.15408194 | 0.173665725 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 技4 | 入力動画の一例 |  |  |  |  |  |
| | 平均二乗誤差 | 0.164972997 | 0.173882608 | 0.167448567 | 0.096982272 | 0.159568469 |
| 技5 | 入力動画の一例 |  |  |  |  |  |
| | 平均二乗誤差 | 0.021621271 | 0.060623025 | 0.08496228 | 0.032133186 | 0.04101096 |

※各技内で最小の値をグレーで表示

視点によって推定誤差が2倍程度変化することがわかった。また、光の加減で選手の姿が真っ白に写っている視点では推定誤差が大きくなる傾向があった。これは、真っ白に写っている場合だと選手の動きに関する情報が少なくなるためだと考えられる。他に、水平から撮影したカメラの方が推定誤差が小さい傾向にあった。これは、より選手が大きく映るためだと考えられる。

・最適なスコアラベル

次に、使用するラベルによって推定精度がどの程度変化するか確かめるために、各技についてTCHスコア、ATHスコア、総合スコアの3種類のスコアを用いて学習・推定を行った。スコアラベルによってばらつきが異なっており、ばらつきが大きいラベルで推定するほど推定値の正解値からの平均二乗誤差は大きくなりやすいため、単純にこの値を比較することはできない。そこで、推定値の正解値からの平均二乗誤差(a)とラベルの平均値を推定値とした時の平均二乗誤差(b)を計算し、aをbで除算した値を比較する。(b)は入力にかかわらず一定の値を出力するモデルである。aがbに対して小さいほど判別性能が良いと考えることができる。結果を表3に示す。

表3　各スコアラベルにおける平均二乗誤差

| | | 平均二乗誤差(a) | 平均値を推定値としたときの平均二乗誤差(b) | a/b |
|---|---|---|---|---|
| 技1 | TCHスコア | 0.016188647 | 0.246067019 | 0.065789584 64 |
| | ATHスコア | 0.013524071 | 0.175605946 | 0.07701374186 |
| | 総合スコア | 0.012795448 | 0.268271202 | 0.04769594319 |
| 技2 | TCHスコア | 0.172383575 | 0.345396825 | 0.4990884754 |
| | ATHスコア | 0.258647038 | 0.215147392 | 1.202185326 |
| | 総合スコア | 0.193212849 | 0.273614412 | 0.7061501176 |
| 技3 | TCHスコア | 0.173665725 | 0.339430587 | 0.5116384075 |
| | ATHスコア | 0.263973215 | 0.245810028 | 1.073891157 |
| | 総合スコア | 0.201473454 | 0.26711479 | 0.7542579514 |

| | | | | |
|---|---|---|---|---|
| 技４ | TCH スコア | 0.096982272 | 0.31766188 | 0.3053003153 |
| | ATH スコア | 0.182636125 | 0.225704208 | 0.8091835192 |
| | 総合スコア | 0.118970571 | 0.210600504 | 0.5649111412 |
| 技５ | TCH スコア | 0.021621271 | 0.232582514 | 0.09296172177 |
| | ATH スコア | 0.039448433 | 0.222222222 | 0.1775179485 |
| | 総合スコア | 0.024415601 | 0.183758025 | 0.1328682165 |

５つの技のうち４つで TCH スコアが最も推定誤差が小さくなった。また全ての技において ATH スコアより TCH スコアの方が推定誤差が小さくなった。総合スコアは TCH スコアと ATH スコアを合算して計算されることを考慮すると、ATH スコアより TCH スコアの方が推定しやすいと考えられる。
RGB の C3D ネットワークはアピアランスに強い影響を受けると言われている。ATH スコアは動きの緩急に関する評価項目であるため、RGB の C3D ネットワークではうまく特徴抽出できなかったと考えられる。一方、物体の動きを表すオプティカルフローを入力とする C3D ネットワークを用いれば ATH スコアの推定誤差が小さくなる可能性がある。

・最適な視点とスコアについて交差検証
以上で求めた、推定誤差が最も小さくなるネットワーク構成・視点・スコアラベルにて、学習・推定を行い、推定値について正解値からの平均二乗誤差を計算し、a を b で除算した値を比較した。なお、学習・推定時は leave one subject out の交差検証を行った。結果を表４に示す。また、技２について正解値と推定値の関係をプロットした図を図２に示す。

表４　最適なネットワーク構成・視点・スコアラベルにおける交差検証の平均二乗誤差

| | カメラ番号 | スコアラベル | 平均二乗誤差 (a) | 平均値を推定値とした時の平均二乗誤差(b) | a/b |
|---|---|---|---|---|---|
| 技１ | カメラ１０ | 総合スコア | 0.148219307 | 0.199414412 | 0.743272792 |
| 技２ | カメラ２１ | TCH スコア | 0.188003927 | 0.345396825 | 0.5443128404 |
| 技３ | カメラ２１ | TCH スコア | 0.282614012 | 0.339430587 | 0.8326120949 |
| 技４ | カメラ２０ | TCH スコア | 0.231145866 | 0.31766188 | 0.7276474793 |
| 技５ | カメラ１０ | TCH スコア | 0.133427753 | 0.232582514 | 0.573679211 |

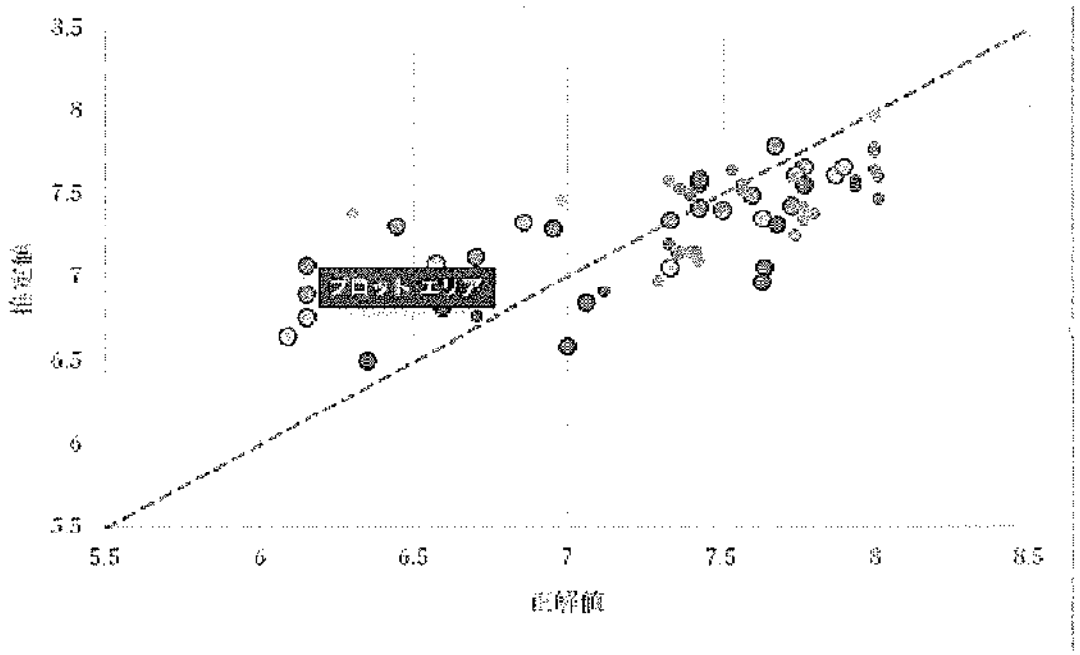



図２　技２における正解値と推定値の関係

a/b の値は技によって 0.55~0.85 程度となった。推定値の誤差はおおよそ 0.3~0.5 点程度となる。図２からもわかるように、モデルは全く学習していないわけではないが、一方で熟練者の知見を得たとまでは言えない。本研究の目的である、初心者にわからないような型の細かな違いを認識するまでは至っていない。

要因の一つとしてあげられるのはデータセットの不足、特に下手な型のデータの不足である。型競技では、決まったルールから外れた動きをすればどのような動きでも下手だと判定される。今回のデータセットに含まれる下手な型は様々な動きをしており、学習データに含まれていない下手な動きがうまく推定できなかったと考えられる。

## 参考文献

[1] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[2] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *The IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618-626.

[3] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[4] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.