Vol.2016-CVIM-203 No.49
2016/9/6

画像分類を用いた一人称行動認識

瀬川 雄太[1,a)]　川本 一彦[1,b)]　岡本 一志[2,c)]

概要：本研究では, 学習済み DCNN(deep convolutional neural network) の fine-tuning を利用した, 画像分類ベースの一人称行動認識を行う手法を提案する. 画像認識問題において state-of-the-art な DCNN の学習や, 学習済み DCNN モデルの fine-tuning のような再学習は魅力的な手法である. しかしながら, 必要となる訓練データを正解ラベルとともに大量収集することはコストが高い. 一方で, 一人称視点行動に付随する物体は映像中に現れやすく, かつ映像を通して見えの変化が小さいため, これを認識するための画像生成は容易に行うことができる. そこで, 行動に付随する物体の画像を人工的に生成し背景画像と合成することで, 一人称行動認識に向けた訓練画像を用意する. 人工訓練データセットを用いて, 学習済み DCNN モデルにおける最終層を fine-tuning し, domain-specific な応用の検証を行う. 実験では, 実際に撮影した一人称視点映像 20 種について, 画像分類ベースの一人称読書行動認識の精度を評価する. 人工訓練データセットに関する検証として, 背景画像選択および人工本画像のための生成処理選択が, 識別精度に与える影響を調査する. 前者の検証において, ImageNet 画像を背景合成した訓練画像を用いて, ImageNet データセットを学習済みの Inception-v3 モデルを fine-tuning した場合に最も精度が高く, F 値にして 91.8% であった. 後者の検証において, 本表面のテクスチャを描画する処理が最も識別精度を高めることがわかり, この処理を付加した訓練画像による評価ではいずれも 85% 以上の F 値を得た.

キーワード：一人称行動認識, 画像分類, 物体認識, 深層学習

1. はじめに

インターネット環境やウェアラブルデバイスの発達に伴い, ADL(activity of daily living) の記録と収集が活発化している. ADL の分析や応用は, QoL(quality of life) 向上のためのシステム構築に役立つ. 特に, ウェアラブルな一人称視点カメラを用いて記録された映像における ADL は, 行動認識の技術によって日々解析されている.

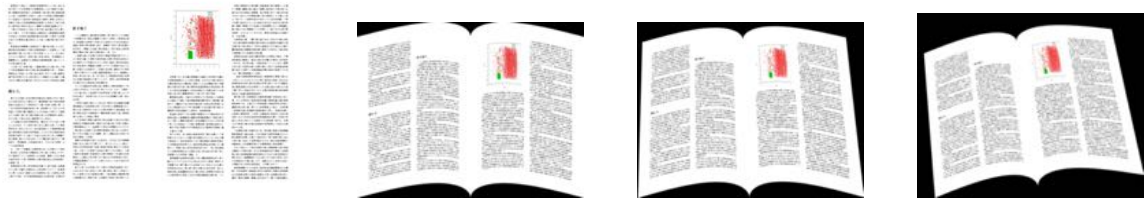
一人称視点における行動認識はポピュラーな話題であり, これまで多くの認識手法が提案されてきた. 中でも, optical flow のようなモーション特徴を映像から抽出して認識に用いる手法が多く存在する [1, 2]. 静止画像ベースの手法として, 特徴点における方向ヒストグラムを用いた特徴記述を行い, これを SVM によって識別する手法がある [3]. よく用いられる画像ベースの手法として, dense-SIFT 記述子による bag-of-features を用いるものがある [4]. 近年においては, 画像認識における state-of-the-art なモデルとして DCNN(deep convolutional neural network) が注目を浴びており, これを用いた一人称行動認識の手法も多く提案されている. Ryoo ら [5] は DCNN の中間層から抽出した特徴量を用いることで, 従来手法よりも高い精度での行動認識を実現している. さらに, DCNN に時空間方向の畳み込みの処理を加えた C3D(convolutional 3D) が提案され, 一人称行動認識に用いられている [6].

DCNN による識別精度を向上させる技術として, 学習済み DCNN モデルに対する fine-tuning がある. Daniel ら [7] は, 手作業で収集したデータセットを用いて学習済み DCNN の fine-tuning を行い, 高精度での一人称行動認識を実現している. DCNN の学習や fine-tuning のような再学習の技術は魅力的であるが, 必要となるラベル付き訓練データの大量収集はコストが高い.

一方で, 一人称行動に付随する物体は一人称視点映像中に現れ, かつ映像を通して見えの変化が少ない. したがって, このような物体を表現する訓練画像は比較的単純な処理のみによって人工生成することができる. そこで本研究では, 人工訓練データセットを DCNN の fine-tuning に用いた, 画像分類ベースの一人称行動認識の手法を提案する.

一人称行動の解析において, PC のようなデジタルな物体を用いる場合は利用履歴を記録することができるため, これを解析に活用できる. 一方で, 本のようなアナログな物

1 千葉大学大学院融合科学研究科
2 電気通信大学 大学院情報理工学研究科
a) segawa@chiba-u.jp
b) kawa@faculty.chiba-u.jp
c) kazushi@uec.ac.jp



(a) 左ページ. (b) 右ページ. (c) ページ結合と歪曲化. (d) 射影変換. (e) 回転.

図 1 人工本画像の生成手順. 決定したレイアウトに従いテクスチャを描画した (a)(b) に対して (c)-(e) の順に処理する.

図 2 人工訓練データセット (dataset A) 中のポジティブサンプル.

図 3 人工訓練データセット (dataset A) 中のネガティブサンプル.

体を用いる場合はそのような手段を取ることができない. そこで本研究では, 一人称読書行動の認識を目的として, 必要となる訓練画像を合成により図 2 のように用意する. まず, 図 1 のような開いた本を表す画像を人工生成する. 次に, 実際の一人称視点画像である図 3 のような背景と合成することで, 図 2 のように開いた本の映った訓練画像を用意する. これらを “Reading” クラスのサンプルとして学習に用いる. さらに, “Others” クラスのサンプルとして, 図 3 ように開いた本を含まない背景のみの画像を用いる. 識別器に対して入力された一人称視点の画像を, 本を含む画像とそうでない画像に分類し, それぞれを “Reading” および “Others” の行動へと対応づけることで, 読書行動の認識へと応用する.

実験では, 人工訓練データセットを, ImageNet データセット [8] を学習済みの state-of-the-art な DCNN モデルである Inception-v3 [9] の fine-tuning に用いることで, domain-specific な応用の検証を行う. fine-tuning の際には, 学習の計算コストを下げるため, Inception-v3 の最終層におけるパラメータのみを再学習する. 比較のための 3 つの従来手法として, 学習済みでない表 2 のような構成の 3 層 DCNN, NN(nearest neighbor), および SVM(support vector machine) を識別器として用いる. NN および SVM の学習には, 従来よく用いられている, dense-SIFT による fisher features [10] を与える.

さらに, 訓練画像合成における背景画像, および人工本画像の生成処理の選択が, 識別に与える影響について 2 つの実験によって検証する. まず, 訓練画像の合成に用いていた実際の一人称視点画像を ImageNet データセットの画像に差し替えることで, 図 4 のおよび 5 のように新たな人工訓練データセットを作成し, これによる識別精度を評価し比較する. 次に, 本画像の人工生成の際に行うレイアウトの描画, エッジの歪曲化, 射影, および回転変換の処理を部分的に削減することで, いくつかの単純化した訓練データセットを図 10 のように新たに 7 種類作成し, これらによる識別精度を比較する.

2. 一人称読書行動認識に向けた訓練画像の人工生成

一人称行動に付随する物体は一人称視点映像中に現れ, かつ映像を通して見えの変化が少ないため, これを表現する画像は比較的単純な処理のみで生成できる. そこで, 一人称読書行動の認識に向けて, 開いた本の画像を人工生成

図 4　ImageNet データセットの画像を用いた人工訓練データセット (dataset B) 中のポジティブサンプル.

図 5　ImageNet データセットの画像を用いた人工訓練データセット (dataset B) 中のネガティブサンプル.

表 1　開いた本画像の生成に向けて設定した主なパラメータ.

処理	パラメータ	値
L	余白	上下左右に 10%
	カラム数	1 または 2
	図の種類	図表および写真
	図の有無	サンプル全体の 80%
	図の大きさ	ページ縦の 12.5%〜50%
	図の配置	トップまたはボトムに 1 つ
	本文の文字	平仮名, 片仮名, および常用漢字
	本文の改行	1 文字描画ごとに 1% の確率で発生
	見出しの大きさ	ページ縦の 10%
	見出しの配置	任意の行に 1 つ
D	歪曲化強度 (α)	0.1〜0.3
P	ホモグラフィ行列	図 7 を参照
R	回転角度	$-10°$〜$+10°$

する.

本画像の人工生成は主に, 図 1 に示すような, レイアウトの描画, エッジの歪曲化, 射影変換, および回転変換の 4 つの処理手順からなる. 以降, これらの処理を L(layout drawing), D(distortion), P(projection), および R(rotation) と呼ぶこととする. これらの処理は, 表 1 に示される生成パラメータに従って行われる.

本表面に描かれたテクスチャはもとより電子的な手順で印字や描画されたものである. そこで, 本表面のテクスチャを表 1 に示す幾つかのパラメータ設定により人工的に生成することを考える. まず, 本のレイアウトとして本文, 見出し, および図の描画領域を決定する. 本文および見出しの領域には, 無作為に決定した平仮名, 片仮名, および常用漢字からなる文字列を, 横書きになるよう水平に並べる. この時, 文字の羅列によるテクスチャを表現することを目的としているので, 文法規則などを考慮せず無作為な順序で並べる. 図の領域には, あらかじめ用意した画像データセットから無作為に選択された図を配置する. 以上の処理によって, 図 1(a), (b) のようなページ画像を得る.

得られた 2 つのページ画像のエッジに, 歪曲化処理を加える. エッジの歪曲化は, 厚みのある本のエッジの形状を表現するためのものである. ページ画像上の画素 (x, y) の y 座標を歪曲後の座標

$$y' = y - f(x) \tag{1}$$

へと変換することで, 歪みを表現したページ画像を生成する. ここで $f(x)$ は, 画像上の水平方向の位置に依存して歪み方を変えるための関数であり, 強度パラメータ α を用いて,

$$f(x) = \alpha(x\sqrt{1-x^2}) \tag{2}$$

と定義している. 図 6 は関数 $f(x)$ の形状を表している. 歪めた 2 つのページ画像を隣り合うように結合することで, 図 1(c) のような画像を得る.

一人称視点での物体の見えを再現するために, 射影変換の処理を行う. ここで, 視点が一人称に固定されていることから小さな見えの変化のみを想定すれば良いため, 射影変換には図 7 に示すような手順で変化の小さいパラメータのみを与えている.

最後に, 回転の処理を加える. 回転も同様に, 一人称視点であることを考慮して, 表 1 に示した小さなパラメータ範囲での変換で表現している.

人工生成した本画像を, 背景画像と合成する. 背景画像には, 図 3 に示すような, あらかじめ実際に撮影した一人称視点画像を用いる. 一人称行動を映した映像中には行動に

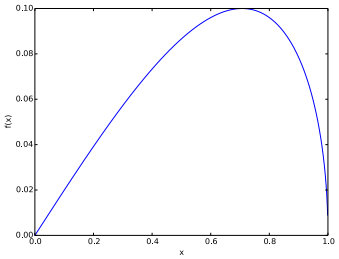


図 6　式 (2) で定義される, 歪曲化処理のための関数 $f(x)$ の形状 ($\alpha = 0.1$). 水平方向の座標に依存して歪みを変えるように設定している.

図 7　射影変換のためのパラメータ決定の手順. 画像の右上および左上の端点 (赤点) を, それぞれ縦横 10% の大きさの領域 (緑枠) 内から無作為に決定した任意の点 (緑点) に変換するようなホモグラフィ行列を推定し, 射影変換を行う.

付随する物体が常に見えていることを想定し, 背景画像上に本画像の 80% 以上が存在するような位置に人工本画像を合成し, 図 2 のような画像を得る.

合成された画像を, “Reading”を表すポジティブクラスのサンプルとして学習に用いる. 一方で, 図 3 に示すような合成における背景画像をまた, “Others”を表すネガティブクラスのサンプルとして学習に用いる. したがって, 訓練データセットにおけるポジティブおよびネガティブサンプルは, 人工生成された開いた本画像が合成されているかどうかの差異のみをもっている. 以降, この訓練データセットを dataset A と呼ぶ.

3. 一人称視点映像を用いた読書行動の認識

画像合成により用意した人工訓練データセットを用いて, 実際に撮影された一人称視点映像中の読書行動を画像分類ベースで認識する. 人工訓練データセットに関する検証として, 2 つの実験を行う. 一つは訓練画像合成に用いる背景選択による識別精度の比較で, もう一つは人工本画像の生成処理の選択による識別精度の比較である.

画像分類のための識別器として, state-of-the-art な DCNN モデルである Inception-v3 を用いる. Inception-v3 は ImageNet データセットを学習済みの公開 DCNN モデルである. 大規模画像データセットを学習済みの DCNN モデルを, 適用する問題に合わせたデータセットの fine-tuning によって転移学習することで, より高精度での画像分類ができることが知られている [11]. そこで本研究でも, 人工訓練データセットを Inception-v3 の fine-tuning に用いることで, domain-specific な応用の検証を行う. また, ImageNet 学習済みモデルの中間層表現は, 他の多くのデータセットに対して高い分類性能を持つ特徴抽出に応用できることについて既に調査されている [12]. そこで, モデル全体ではなく, モデルの出力層直前に位置する全結合層のパラメータのみを fine-tuning することで, 学習に要する計算時間を短縮する.

表 2　Inception-v3 モデルの他に用いた, 中間層を 3 層持つ DCNN モデルの構成.

layer	size/stride	input size
conv1	7×7/1	256×256×1
pool1	2×2/2	256×256×16
norm1	-	128×128×16
conv2	7×7/1	128×128×16
pool2	2×2/2	128×128×16
norm2	-	64×64×16
conv3	7×7/1	64×64×16
pool3	2×2/2	64×64×16
norm3	-	32×32×16
linear1	-	32×32×16
linear2	-	128
softmax	-	16
output	-	2

加えて比較のために, 従来用いられている 3 つの手法でも同様に一人称読書行動の認識を行う. まず, 表 2 に示した 3 層 DCNN モデル全体を学習し用いる. さらに, dense-SIFT による fisher features を, NN および SVM の学習に用いる. dense-SIFT および fisher feature の組み合わせは, DCNN が注目を浴びた ILSVRC(ImageNet Large Scale Visual Recognition Challenge)2012 においても実際に牛久ら [13] によって用いられている, ポピュラーな画像特徴量である.

識別精度の評価のために, 様々な場面での読書行動を含む数分間の一人称視点映像 20 種を撮影した. 評価用データセットとして, これらの映像から毎秒 6 フレームずつ図 8 のような静止画を抽出し, 手作業によって各フレームにラベル付けをして用意した. 20 種の映像それぞれについて算出した F 値の平均によって, 各手法による識別精度を評価する.

3.1 訓練画像の背景選択による識別精度の比較

人工訓練データセットとして, 図 2 および 3 に示すようなポジティブおよびネガティブサンプルをそれぞれ 25,000 ずつ用意した dataset A を学習に用いる. 4 つの手法それぞれによる識別を行い, 表 3 上段のように精度を評価した.

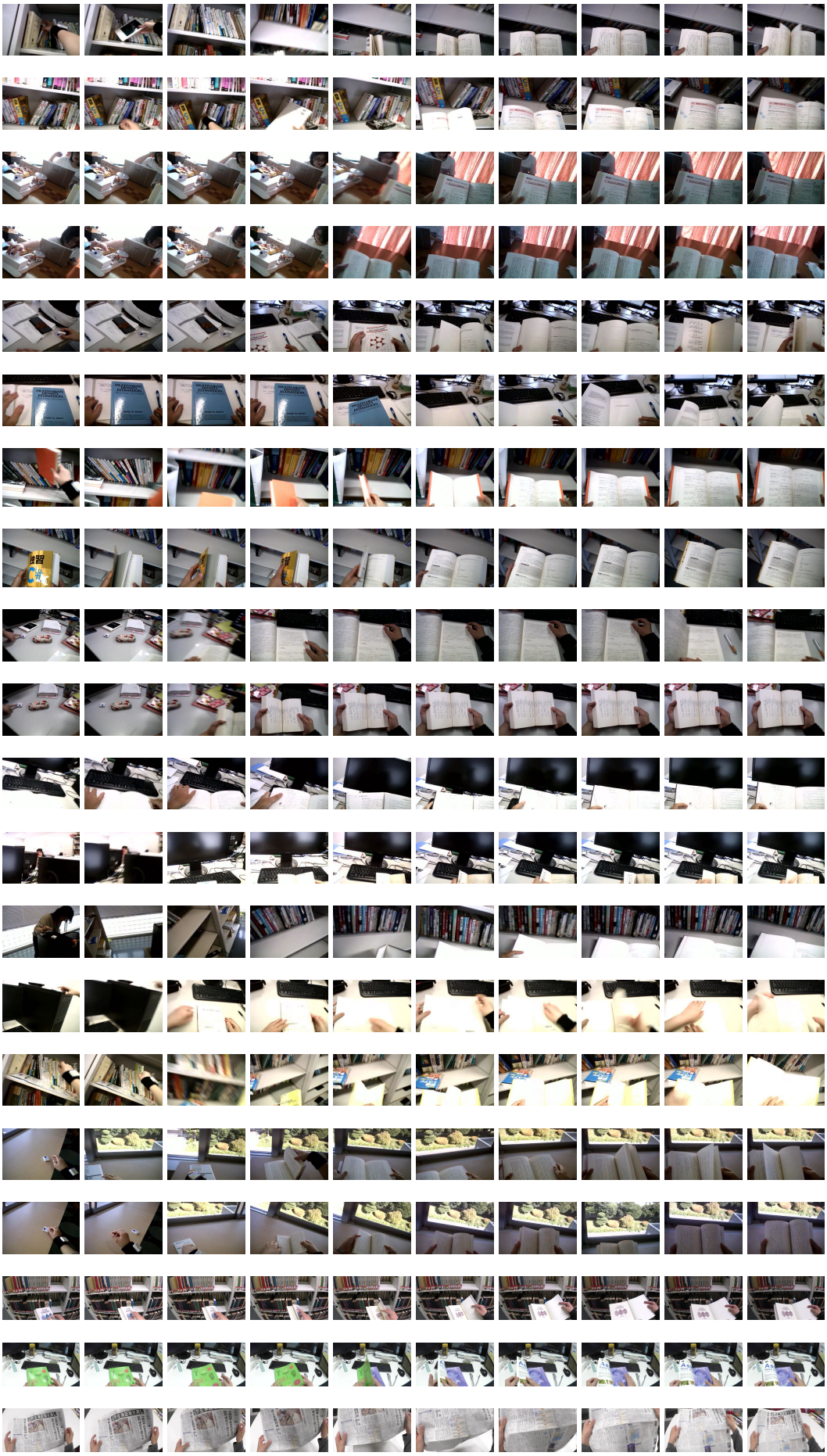
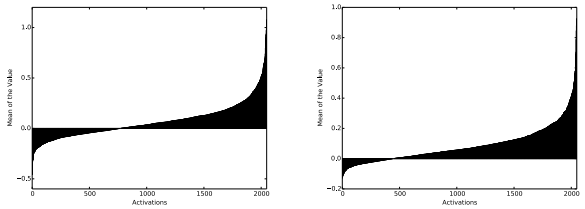


図 8　実際に読書行動を撮影した評価用の一人称視点映像 20 種. 抽出したフレーム全てに, 手作業によって “Reading” または “Others” クラスのラベル付けをしている.

表 3　dataset A および B を用いた 4 つの手法による, 一人称読書行動の識別精度の比較.

F-measure(%)	NN	SVM	3-layer DCNN	Inception-v3
dataset A	74.1	50.7	48.2	70.7
dataset B	55.5	86.2	15.3	**91.8**



(a) dataset A　　(b) dataset B

図 9　1,024 組のポジティブおよびネガティブサンプルに対する 2,048 次元の活性化を抽出し, 各次元について計算した差の平均を表したもの.

表 4　fine-tuning に用いる訓練画像の数を変化させた場合の識別精度.

dataset size	1,000	5,000	25,000
F-measure(%)	91.6	90.1	91.8

Inception-v3 は著しく高い識別精度とはならず, その一方で NN が最も高い精度での識別を行うことができている.

さらに, 訓練画像の背景選択による識別精度を比較するために, dataset A とは別の新たな人工訓練データセットを生成する. これまで用いていた実際の一人称視点画像の代わりに, より収集コストの小さく済む ImageNet データセット中の画像を, 訓練画像合成の際の背景として用いる. 図 4 および 5 は, ImageNet 画像によって新たに生成した人工訓練データセットのポジティブおよびネガティブサンプルの一例である. 以降, この訓練データセットを dataset B と呼ぶ.

dataset B を用いて同様に評価した識別精度を表 3 下段に示す. 実際の一人称行動における画像と無関係の背景を訓練画像に用いたにもかかわらず, Inception-v3 の識別精度は大きく上がり表中で最も高い値となった. これは, あらかじめ ImageNet データセットによって訓練され, それらの画像を十分に識別することのできる Inception-v3 が, ImageNet 画像を背景とした訓練画像のうち前景部分をより強く識別に寄与するように fine-tuning されたと考えることができる. 以上の結果より, ImageNet 画像を訓練画像合成の背景として用いることで収集コストをより小さくすることができるだけでなく, ImageNet 学習済みの Inception-v3 の fine-tuning に用いることでより高精度な画像分類ができることが確認できた.

fine-tuning によって再学習しているのは, モデルの出力層直前に位置する全結合層におけるパラメータのみであり, それ以前に位置する中間層のパラメータに一切の変化を与えていない. しかしながら, fine-tuning した Inception-v3 は高精度で一人称読書行動を識別することができている. よって, Inception-v3 の中間層表現はポジティブおよびネガティブクラスを十分に識別できる特徴を抽出する性能をもとより有していたと考えることができる. この事実を検証するために, ポジティブサンプルとそれに対応するネガティブサンプルの組について, Inception-v3 から得られる中間層特徴を比較した.

まず, 1,024 組の両サンプルを Inception-v3 に入力し, 全結合層直前における全 2,048 次元の活性化を計算し, 中間層特徴のベクトルとして得る. 次に, 得られた両サンプルの特徴ベクトルの差を取り, 1,024 組について平均を得る. 最後に, 計算された差の平均の大きさによって 2,048 次元の要素を並べ替える. 図 9 はこれらの手順によって得られた, 両クラス間における Inception-v3 の活性化の応答差の分布である.

図から, いずれのデータセットにおいても分布はおおよそ似た形状であり, 特定の少数の活性化が大きな差を持っていることが観察できる. このことから, 特徴空間上でこれらに対応する次元において両クラス間で十分に分離しており, 出力へのロジスティック回帰に大きく寄与していると考えることができる. 入力した両サンプルには, 人工生成した本画像が合成されているかそうでないかの差異のみが存在する. したがって, 分布のうち大きな差を持っている少数の活性化は, このような入力画像の差異に対して反応しているものだと考えられる. このことから, このような活性化に対してこのような強い差異をもたらすような人工訓練データを作ることで, より読書行動認識に有効な人工訓練データの作成への応用を検討できる.

Inception-v3 は fine-tuning の過程において, 図 9 に示したような特定の活性化に強く反応するように重みを再学習したと考えることができる. これを別の視点から検証するために, fine-tuning に用いる訓練データ数を変えて識別精度を比較する. これまで用いていた各クラス 25,000 ずつのデータセットの他に, 5,000 および 1,000 のデータセットによる識別精度を評価し表 4 のように比較した.

表より, いずれのデータ数においても, 識別精度は大きく変わらないことが確認できる. fine-tuning されるべき全結合層の重みは, 2,048 次元の活性化を 2 クラスに回帰させるために 4,096+2 だけ存在する. しかしこの実験結果から, 実質的に fine-tuning されるべき次元数は 1,000 以下で十分であったと考えることができる.

3.2　人工本画像の生成処理の選択による識別精度の比較

読書行動の認識に対してより有効な本画像の特徴を調査するために様々な人工本画像を生成し, これらを用いて fine-tuning した Inception-v3 モデルによって識別精度を比較する. これまでの本画像の生成には, レイアウトの描画,

図 **10** dataset B における人工本画像の生成処理を部分的に省略することで単純化し新たに用意した 7 種類の訓練画像の一例. 最上段から順に, None, R, RP, RPD, L, LR, および LRP の処理の組み合わせで生成されたものを表している.

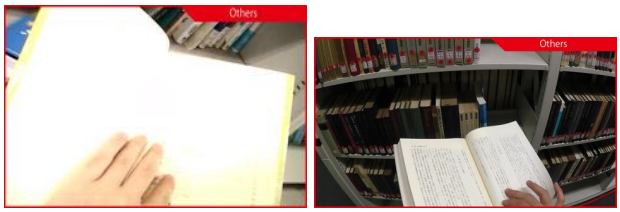
表 **5** 人工本画像の生成処理 L, D, P, および R を部分的に省略し, 単純化した訓練データセットでの識別精度の比較. LRPD は表 4 における dataset size 5,000 そのものである.

Processing	None	R	RP	RPD	L	LR	LRP	LRPD
F-measure(%)	28.5	46.3	56.7	49.5	**85.1**	**86.5**	**89.1**	**90.1**

エッジの歪曲化, 射影変換, および回転変換の 4 つの処理をすべて行っていたが, これらのうち適用する処理を部分的に選択し, その組み合わせを考えることで, 計 8 種類の人工本画像を用意する. 図 10 はこれらの組み合わせによって新たに人工生成した 7 種類の訓練画像の一部である. これらの訓練画像を用いて, ポジティブおよびネガティブサンプルそれぞれ 5,000 ずつのデータセットを用意する.

以上のデータセットを実際に fine-tuning に用いて, 表 5 のように識別精度を比較した. 本画像そのものの形状を変えるための処理 D は, その他の R および P に比べて識別精度に対してあまり影響を与えていないことが確認できる. 一方で, 処理 L を加えているものはそうでないものに比べて高い識別精度であることが確認できる.

処理 L が大きな影響を与えることは, 図 11 に示すような評価データに対する識別結果の側面からも確認できた. 図 11 は処理 LRPD 全てを用いた場合の識別例のうち, 特に悪い識別精度であったものを一部抜粋したものである. 図 11(a) のような撮影環境の光量が多い場合や, 大部分が空白であるページを開いているなどして本表面のテクスチャが映っていない映像の場合に識別精度が低くなることを確認している. 一方でこのような映像に対しては, 処理 L を加えない場合はむしろ, 処理 L を加えた場合よりも識別精度が上がることも確認している. また図 11(b) のように読んでいる本が縦書きの場合にも, 識別精度は低くなっていた. これは, 本実験で生成した人工本画像が全て横書きを想定したレイアウト描画の処理をしていることに起因していると考えられる. 以上のことから, 人工訓練データセットに対する fine-tuning の過程では, 本画像のエッジ特徴のみならず, 本表面のレイアウトによって表現されるテクスチャ特徴を大きく寄与させるようなパラメータ学習が行われていると言える. 特に, 文字列の為す方向性を持った周波数成分が強く識別に関わっていると推測できるため, これを表現するような描画処理を考えることで, 処理 L のより有効かつ効率的な手段を検討することができる.



(a) (b)

図 **11** すべての処理 LRPD を行った訓練画像で fine-tuning した Inception-v3 による識別例. 赤枠の画像は “Others”として認識されたフレームであることを表す. (a) は光量の多い環境で撮影されたため本表面のテクスチャが視認できない状態である. (b) は人工本画像のレイアウトとして与えなかった, 縦書きの本を読んでいる状態である.

4. おわりに

本研究では, 学習済み DCNN に対して fine-tuning を適用し, 画像分類ベースの一人称行動認識の手法を提案している. 一人称読書行動認識への応用のために, 人工的に生成した開いた本の画像と任意の背景画像との合成により訓練画像を用意する.

実験では, ImageNet データセットを学習済みの DCNN モデルである Inception-v3 の fine-tuning に対して人工訓練データセットを用いる domain-specific な応用の検証として, 実際に撮影した一人称読書行動を認識し, その F 値を評価している. 加えて, 従来用いられている 3 層 DCNN, NN, および SVM による 3 つの手法でも同様に識別精度を評価している.

これらの手法を用いて, 人工訓練データセットの, 訓練画像合成に用いる背景選択, および人工本画像の生成処理の選択に関しての, 2 つの側面から検証をしている. まず背景選択の方法として, 実際に撮影した一人称視点画像を用いる場合と, ImageNet データセット中の画像を用いる場合で比較した. 結果として, ImageNet 画像を背景とした人工訓練データセットに対して fine-tuning した Inception-v3 モデルによる識別精度が最も高かった. Inception-v3 モデルが ImageNet を学習済みであったことが寄与していると考え, この事実を今後検証していく必要がある.

さらに, ポジティブサンプルとネガティブサンプルに関して Inception-v3 から抽出される中間層特徴の分布を解析した. 結果として, 特徴空間上における, ある特定の少数の次元に関する平均が大きく分離していることがわかった. したがって, これらの少数の活性化のみが線型分離に寄与しているため, 少数の訓練データ数の fine-tuning でも高い識別性能を獲得できるということを確認した. この事実に対する別の側面からの考察として, fine-tuning に必要な訓練データ数を調査したところ, 1,000 ないし 5,000 程度でも十分であることを確認した.

次に人工本画像の生成処理の選択に関して, 必要な処理を部分的に省略することで単純化した人工訓練画像を 7 種類用意し, 一人称読書行動の識別に対して大きく寄与する処理について調査した. 結果として, 本表面のレイアウトの描画の処理を加えることが, 最も識別性能を向上させていることがわかった. このことから今後は, テクスチャが与える DCNN 特徴への影響を解析することで, より効果的なレイアウト描画の処理の方法について検討する. さらに, これらの訓練画像の人工生成に関する知見から, 物体が付随して起こる読書以外の一人称行動認識への拡張を目指す.

謝辞　本研究は JSPS 科研費 JP25330186,JP16K00231 の助成を受けたものです.

参考文献

[1] Zhan, K., Guizilini, V. and Ramos, F.: Dense motion segmentation for first-person activity recognition, *Control Automation Robotics Vision (ICARCV), 2014 13th International Conference on*, pp. 123–128 (2014).
[2] Fathi, A., Farhadi, A. and Rehg, J. M.: Understanding Egocentric Activities, *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pp. 407–414 (2011).
[3] Behera, A., Chapman, M., Cohn, A. G. and Hogg, D. C.: Egocentric activity recognition using Histograms of Oriented Pairwise Relations, *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, Vol. 2, pp. 22–30 (2014).
[4] Yan, Y., Ricci, E., Liu, G. and Sebe, N.: Egocentric Daily Activity Recognition via Multitask Clustering, *IEEE Transactions on Image Processing*, Vol. 24, No. 10, pp. 2984–2995 (2015).
[5] Ryoo, M. S., Rothrock, B. and Matthies, L.: Pooled motion features for first-person videos, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 896–904 (2015).
[6] Takamine, A., Iwashita, Y. and Kurazume, R.: First-person activity recognition with C3D features from optical flow images, *2015 IEEE/SICE International Symposium on System Integration (SII)*, pp. 619–622 (2015).
[7] Castro, D., Hickson, S., Bettadapura, V., Thomaz, E., Abowd, G., Christensen, H. and Essa, I.: Predicting Daily Activities from Egocentric Images Using Deep Learning, *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ISWC '15, ACM, pp. 75–82 (2015).
[8] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252 (online), DOI: 10.1007/s11263-015-0816-y (2015).
[9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *CoRR*, Vol. abs/1512.00567 (2015).
[10] Perronnin, F., Sánchez, J. and Mensink, T.: Improving the Fisher Kernel for Large-scale Image Classification, *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, Springer-Verlag, pp. 143–156 (2010).
[11] Girshick, R., Donahue, J., Darrell, T. and Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587 (2014).
[12] Jeff Donahue, Yangqing Jia, e.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, Proceedings of The 31st International Conference on Machine Learning, pp. 647–655 (2014).
[13] Ushiku, Y., Harada, T. and Kuniyoshi, Y.: Efficient Image Annotation for Automatic Sentence Generation, *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, ACM, pp. 549–558 (2012).