CH-004

スペクトログラム画像の畳み込み演算による行動認識
Activity recognition by convolution of spectrogram

伊藤 千紘[1)]　酒造 正樹[1)]　前田 英作[1)]
Chihiro Ito　Shuzo Masaki　Eisaku Maeda

概要

時系列センサ情報を利用した行動認識において、recurrent neural network (RNN)、long short-term memory (LSTM)、convolutional neural network (CNN) などの深層学習の活用が試みられており、時系列信号をそのまま扱う場合や周波数解析後に深層学習の入力とする場合などがある。しかし、その多くは、特定のデータセットでの識別性能を評価することに留まっており、CNN の構造、周波数解析の効果、畳み込み演算の効果などに関する検討が不足していた。また、行動認識 (activity recognition) という用語で括られる行動にも実は多様な要素があり、信号の発生機序に着目した分析が必要である。そこで本論文では、行動に伴ってセンシングされた信号を行動が能動的なものか受動的なのかに類別した上で、FFT スペクトラム画像への変換、畳み込みサイズの最適化、CNN 最終段でのセンサ統合を行う識別手法を提案し、その有効性を検証した。その結果、提案手法の従来手法に対する優位性を実証するとともに、受動的な行動に対しては時間方向と周波数方向の両方向の畳み込みが有効であるのに対し、能動的な行動に対しては周波数方向の畳み込みのみが効果を持つことを明らかにした。

1 はじめに

世界的に普及しているスマートフォンやウェアラブル端末などのデバイスには、加速度、GPS など多様なセンサが搭載されており、これらセンサから得られるデータを利用して様々な新しい機能やサービスが検討されている。その代表的なものとしてリストバンド型デバイスを利用した手の動作認識 [1] や、スマートフォンから得られるセンサ情報を利用した行動認識 [2] などがあげられる。これらは広く行動認識 (activity recognition) として一括りに扱われ、信号処理と機械学習の技術を組み合わせた様々な方法が提案されており、近年は深層学習の利用も試みられている。しかしながら、深層学習を利用した行動認識に関する研究を改めて見直したとき、ニューラルネットワークの潜在的な識別能力と学習データの量に頼ったやや強引な結果であることも少なくない。そこで本論文では、以下の 2 つの観点から行動認識の問題を捉え直してみることとした。

第一に、センサ情報は人間の何らかの「動き」を反映したものであるという基本に立ち返り、広義に「行動 (activity)」として扱われている認識対象に対して、センシングされた信号の特性から行動を分類し、検討を行うことが重要である。そうした例として、hand action に関連する持続的な一連の動きを hand activity として定義し、action と activity を切り分けた上で hand activity の認識を試みたもの [1]、その動的特性に着目し、人間の動きを動的 (dynamic) なものと静的 (static) なものに分けたもの [3] などがある。一方、我々は、動き（行動）の能動性に着目し、人間の行動の識別 (今回扱った例では、Walk、Car、Train などの移動手段の識別 [2]）を例として取り上げ、主体である人間による能動的 (active) な行動によって発生した動きと、何らかの外的要因による受動的 (passive) な動きに大別（図 1）して分析することを試みた。

第二に、深層学習に関するアルゴリズム的視点からの検討を行なうことが必要である。スマートフォンが登場する以前は、映像から得られる各動作の hand-crafted な特徴量を利用したヒューリスティックスな手法で解くことが一般的であった。しかし、センシングできる行動の幅が広がったことでより高次元時系列信号を扱うことが増え、従来の手法では利用されていない特徴量を利用できる可能性が生じた。深層学習技術の発展に伴い、従来の特徴抽出や特徴選択の手法をより効率化させるために、RNN や LSTM などの時系列処理に適したネットワークモデルやスペクトログラム画像と CNN を組み合わせる手法などが試みられている。しかし、CNN による標準的な畳み込み演算を盲目的に利用することが多く、性能の評価結果だけに焦点が行きがちで、アルゴリズムの効用に関する分析が不十分であることが多い。

そこで本論文では、上記 2 つの観点から行動認識の問題を見直し、信号の周波数解析と CNN を組み合わせた手法を提案する。行動に伴ってセンシングされた信号を行動が能動的なものか受動的なのかに類別した上で、FFT スペクトラム画像への変換、畳み込みサイズの最適化、CNN 最終段でのセンサ統合を行う識別手法を提案し、その有効性を検証し、提案手法の従来手法に対する優位性を実証した。さらに、受動的な行動に対しては時間方向と周波数方向の両方向の畳み込みが有効であるのに対し、能動的な行動に対しては周波数方向の畳み込みのみが効果を持つことを明らかにした。これらの実験結果は、行動認識における標準的な情報処理手法の確立につながるとともに、いわゆる深層学習における畳み込み演算の意味について新たな視点を投げかけるものである。

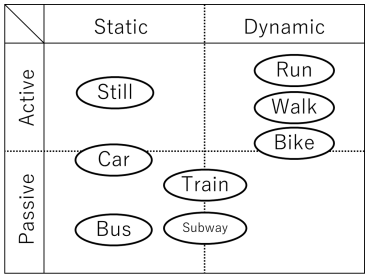


図 1: 動きの動的性質と能動性に基づく行動の分類

1）　東京電機大学 Tokyo Denki University

2 発生要因と動作特性に着目した動作の分類

前節で述べたように、センシングされた信号の生起原因となる人間の動きには、その動的性質から static なものと dynamic なものに分けられることに加えて、デバイス保持者の active な動作・行動によって発生するものと、保持者の意思とは関係のない passive な力が関与することによって発生するものに分けることもできる。Sussex-Huwai Locomotion Dataset[2](以降、SHL データ) は、スマートフォンから得られるセンサ情報に対して、Still、Walk、Run、Bike(自転車)、Car、Bus、Subway、Train の 8 種類のラベルを付与したものであるが、この 8 カテゴリから想定される行動に対して、動的性質 (dynamic, static) と能動性 (active, passive) の 2 つの観点から整理分類したものが図 1 である。Run や Walk といった行動から得られる信号は、dynamic かつ active なものである。一方で、Bus や Car に乗っているという行動から得られる信号は、Bus や Car の物理的運動に起因するものであるという意味で passive なものである。Train や Subway という行動から得られる信号は、乗車しているという状況は、着席中の passive な信号となる場合もあれば、車両内の移動や乗降による dynamic な場合もある。この信号の能動性と識別モデルの関係を明らかにするため、SHL データにおける 8 種の行動のうち、Still、Walk、Run、Bike を人間の active な動きに起因する行動 (以下、A 群)、Car、Bus、Subway、Train を何らかの外的要因による passive な行動 (以下 P 群) として 2 つに分けて各種識別モデルを構築し、分析を行った。

3 周波数解析と畳み込み深層学習

行動認識のための時系列信号の処理に対しては、RNN や LSTM などの深層ニューラルネットワークが用いられることが多い [4]。我々は、FFT 処理後のスペクトログラム画像と 2 次元 CNN を組み合わせる方法を提案し、その優位性をコンペティションにおいて示した [4][5]。一方、時系列データを CNN と組み合わせた方法は、他にも様々な試みがある。周波数解析を使用しない方法としては、センシングされた 3 方向 (xyz) の加速度信号を周波数解析せずに画像に変換し、2 次元 CNN の入力にする方法 [6] や、複数のセンサ軸の原信号を並べて画像として扱う方法 [7] がある。FFT 処理後のスペクトログラム画像と CNN を組み合わせる試みは、複数センサから生成される複数のスペクトログラム画像を統合する方法に関して、異なるいくつかの方法が提案されている。xyz の加速度センサから生成される 3 枚のスペクトログラム画像を RGB 画像として扱う方法 [1]、複数のスペクトログラム画像を時間方向に並べて結合する方法 [5]、周波数方向に結合する方法 [8]、時間方向と周波数方向の両方に結合する方法 [9]、複数センサからの信号系列を並べて画像として、それに対して 2 次元 FFT をかける方法 [7] などがある。さらに畳み込み演算のかけ方にも、通常の物体認識で行われるのと同様の 2 次元 CNN を行う場合 [1][5] と時間方向の 1 次元 CNN を行う方法 [8] などがある。こうした多くの既存研究があるものの、方法間の性能比較や畳み込みの効果、その適切なパラメータなどについては十分な検討がされているとは言えない。

我々は、これまでの予備検討の結果からスペクトログラム画像と CNN の適切な組み合わせが、行動認識において有効であると感触を得ているが、どのような畳み込み処理がどのような場合に有効であるかについて、前述べた信号の特性にも注意を払いながら検証を行った。

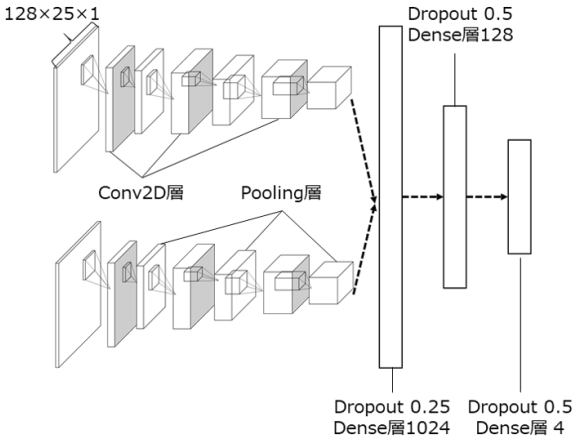


図 2: CNN モデル

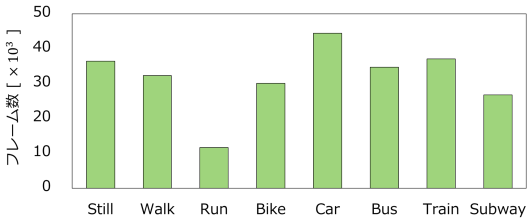


図 3: SHL データの行動別フレーム数（1 フレームは 5 秒、500 点のデータからなる)

4 提案手法

4.1 CNN モデル

行動認識を目的として、周波数解析の結果としてのスペクトログラム画像と CNN を組み合わせた方法を考えるとき、異なるセンサ信号の統合処理と畳み込み演算のかけ方の 2 点が大きなポイントになる。予備検討として、異なるセンサ信号のスペクトログラム画像を結合することで統合し、3 層の畳み込み層と 2 層の全結合層を持つ CNN モデルと組み合わせた [5]。他にも、異なるセンサ信号のスペクトログラム画像に対してそれぞれに畳み込み演算を行い、全結合層前で特徴を結合することで統合した [10]。画像を結合した場合、センサ信号が本来持つ情報とは異なる情報が畳み込まれてしまうため、信号の特性を踏まえたうえでのモデル構築としては優れていないことが分かった。

このような予備検討結果 [5][10] を踏まえて、図 2 に示す多入力 CNN モデルを提案する。この提案手法では、加速度センサ、ジャイロセンサなど各センサ信号に対して FFT 処理を行い、スペクトログラム画像を作成し、CNN モデルの入力とする。各センサのスペクトログラム画像に対して、3 層の畳み込み層で畳み込み演算を行い、特徴を結合した後に 3 層の全結合層で分類を行う。プーリング層は最大プーリングを用いた。各畳み込み層のフィルタは 16, 32, 64 とし、各畳み込み層の入力と同じ長さを出力がもつように入力にパディングを行った。活性化関数には ReLU を用い、最適化関数は Adam

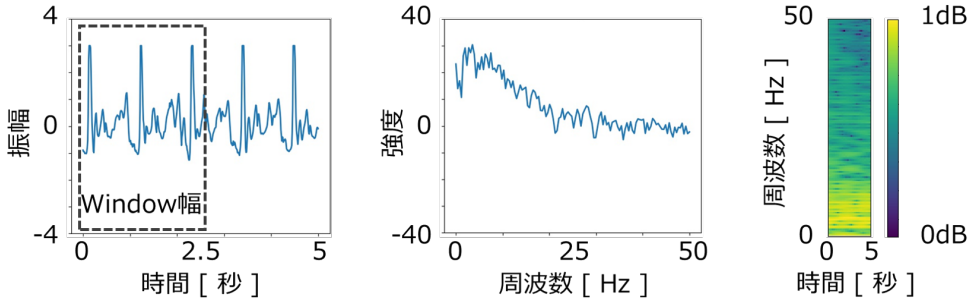


図 4: FFT 処理の流れ。原信号（左）、FFT 処理結果 (中)、CNN への入力となるスペクトログラム画像（右）

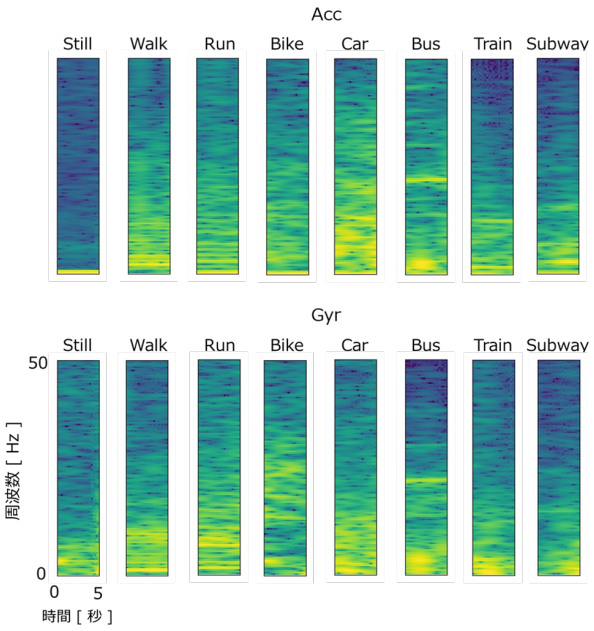


図 5: 各 8 カテゴリのスペクトログラム画像の例（Acc（上）、Gyr（下））

とした。Adam の学習係数は 0.001 とし、損失関数には交差エントロピーを用いた。

ここで、予備検討 [5][10] では畳み込みサイズ (3, 3) としてしていたが、センサ信号の特性に応じた適切な畳み込み演算を分析するために、実験では時間方向と周波数方向の畳み込みサイズを変えて検証した。提案した CNN モデルは、それぞれのセンサ信号に対して畳み込み演算を行い、全結合層前でその特徴を結合するため、基本的なアーキテクチャは変わらずに入力に用いるセンサ信号の数が可変であるため拡張性の高さが利点である。

4.2 入力画像データ

実験には 2018 年のチャレンジに採用された SHL データセット [2][11] を用いた。本データセットには、腰位置に装着したスマートフォン内蔵の加速度、ジャイロなどのセンサデータ（100 Hz）が 60 秒単位として含まれている。なおこの区間内に行動が遷移するものを除外した上で、現実的な応用として 5 秒の認識時間となるようセグメント処理を施した。その結果、500 点を 1 フレームとして、行動別の内訳を図 3 に示す。A 群、P 群のフレーム数はそれぞれ、110,520、143,064 となった。

SHL データセットに含まれるセンサデータは各 3 軸の情報が与えられているが、本論文においては、予備検討 [5] を受け、加速度センサのノルム (Acc) とジャイロセンサの y 軸 (Gyr) の 2 軸の情報を採用する。これは、加速度、ジャイロデータの相関分析を行った結果、相関が低い組み合わせであり、各センサ単軸ごとの識別器を作成した結果、他軸のものよりも高い性能を持ちえることを理由としている。

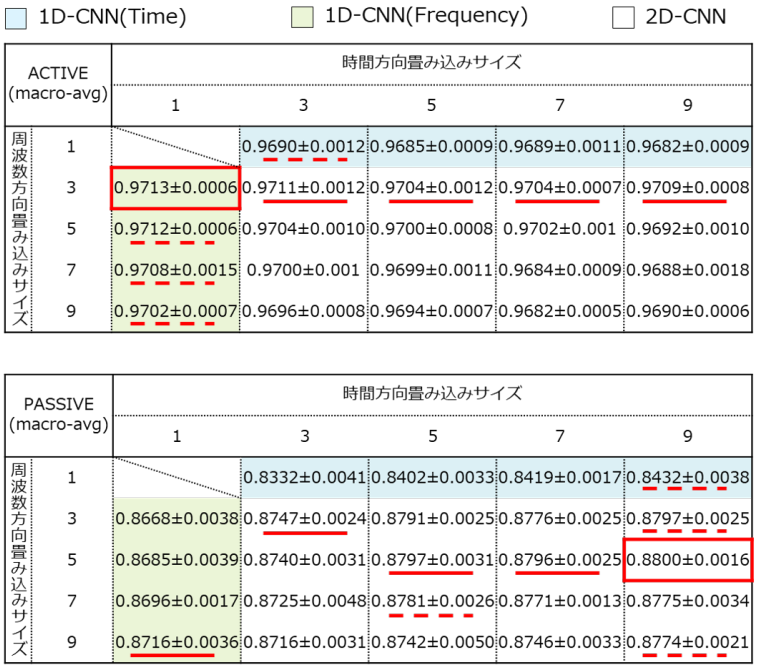
次に、A 群及び P 群の各センサごとに正規化を行い、FFT スペクトログラム画像を作成した。FFT 処理の流れを図 4 に示す。セグメント区間の 5 秒間のデータに対して、ウィンドウ幅 2.56 秒 (256 data points)、オーバラップ 0.1 秒 (10 data points) で FFT を行った。窓関数にはハミング窓を用いた。$10 \times \log_{10}(\sqrt[2]{FFT.real^2 + FFT.imag^2})$ で dB 変換を行い、高さ 128、幅 25 のスペクトログラム画像を作成した。作成したスペクトログラム画像の例を図 5 に示す。ただし、CNN モデルに入力する際はグレースケールとした。

5 実験結果

検証は 5 分割交差検証を用いて行い、検証データに対する loss が最も低いエポックのモデルで評価を行った。詳細な分析のため、時間方向と周波数方向の畳み込みサイズを 1, 3, 5, 7, 9 と変えて分析を行った。分析結果を図 6 に示す。各郡内の最良値を赤線で囲み、周波数方向の最良値を赤実線、時間方向の最良値を赤破線で示した。

A 群では、どの時間方向の畳み込みサイズにおいても周波数方向の畳み込みサイズ 3 の時が最も識別性能が高く（図中の赤実線）、周波数方向の畳み込みサイズ 1 の時と比較して有意に識別性能が向上していた。一方、時間方向の畳み込みは、畳み込みサイズ 1 の場合がわずかではあるが最も性能が高い（赤破線）。即ち、時間方向の畳み込みは効果が見られない。P 群では、時間方向の畳み込み、周波数方向の畳み込みともに有効に働いており、畳み込みサイズ (9,5) の時が最良であった（赤枠）。どの時間方向の畳み込みサイズにおいても周波数方向の畳み込みが有効であり（赤実線）、また、どの周波数畳み込みにおいても時間方向の畳み込みが有効であった（赤破線）。以上の結果から、畳み込みサイズを最適化することによって有意な性能向上を実現できること、A 群と P 群で最適な畳み込みサイズが大きく異なること、A 群では時間方向の畳み込みが効果をもたないことなどが明らかになった。

A 群と P 群での基準モデル（畳み込みサイズ (3,3)）と最良畳み込みモデルにおける識別結果の混同行列を図 7 に示す。A 群の最良畳み込みモデルは、畳み込みサイズ



ACTIVE (macro-avg)

周波数方向畳み込みサイズ ＼ 時間方向畳み込みサイズ	1	3	5	7	9
1		0.9690±0.0012	0.9685±0.0009	0.9689±0.0011	0.9682±0.0009
3	0.9713±0.0006	0.9711±0.0012	0.9704±0.0012	0.9704±0.0007	0.9709±0.0008
5	0.9712±0.0006	0.9704±0.0010	0.9700±0.0008	0.9702±0.001	0.9692±0.0010
7	0.9708±0.0015	0.9700±0.001	0.9699±0.0011	0.9684±0.0009	0.9688±0.0018
9	0.9702±0.0007	0.9696±0.0008	0.9694±0.0007	0.9682±0.0005	0.9690±0.0006

PASSIVE (macro-avg)

周波数方向畳み込みサイズ ＼ 時間方向畳み込みサイズ	1	3	5	7	9
1		0.8332±0.0041	0.8402±0.0033	0.8419±0.0017	0.8432±0.0038
3	0.8668±0.0038	0.8747±0.0024	0.8791±0.0025	0.8776±0.0025	0.8797±0.0025
5	0.8685±0.0039	0.8740±0.0031	0.8797±0.0031	0.8796±0.0025	0.8800±0.0016
7	0.8696±0.0017	0.8725±0.0048	0.8781±0.0026	0.8771±0.0013	0.8775±0.0034
9	0.8716±0.0036	0.8716±0.0031	0.8742±0.0050	0.8746±0.0033	0.8774±0.0021

図 6: 時間方向と周波数方向の畳み込みサイズを変えた際の識別性能の比較（F 値の macro 平均）。Active 群（上段）、Passive 群（下段）。　赤枠：各群内の最良値、赤実線：各群内の周波数方向の最良値、赤破線：各群内の時間方向の最良値。

(3,1) の時であり、F 値 0.9713 である。P 群の最良畳み込みモデルは、畳み込みサイズ (5,9) の時で、F 値 0.8800 である。基準モデルと比較すると Bus、Train、Subway の認識が大幅に改善されている。

時間方向と周波数方向の畳み込みサイズを変えたことで CNN がスペクトログラム画像に対してどこを重視しているかを確認するために、正しく Bike、Subway と予測したスペクトログラム画像 10 枚をそれぞれ Grad-CAM[12] により可視化し、図 8 に示す。畳み込みサイズを変えたことで重視されている部分が変化していることがわかる。

6 考察

スマートフォン等で得られる時系列信号から行動（activity）の認識を行う技術については、これまで様々な研究報告がある。しかしながら、本稿冒頭で述べたように、行動認識として一括りにして議論をしてしまいがちであるが、センシング信号の発生機序に立ち返ってデータの処理アルゴリズムを検討することが重要である。こうした観察と考察は、特徴の抽出・変換に加え、機械学習アルゴリズムの選択、パラメータの最適化を行う際に非常に重要な役割を果たす。

我々は、これまでの検討から、時系列信号を LSTM の入力とする方法や時系列信号を直接信号に変換して CNN の入力とするなどの従来法に比べ、スペクトログラム画像と CNN を組み合わせる方法を提案し、その有効性を示してきた。しかしながら、CNN は本来、画像中の物体認識のために開発されてきた手法であり、スペクトログラム画像を入力としたときの CNN の働きをしているのかについては、明らかではなかった。そこで、本研究では、信号発生の原因となる行動（activity）をその能動性という観点から active 群と passive 群とに分け、CNN の畳み込み演算において時間方向と周波数方向との各畳み込みサイズを変える実験を行った。

その結果、周波数方向の畳み込み演算は A 群と P 群ともに効果があり、最適な畳み込みが存在し、一方、時間方向の畳み込み演算は P 群のみに効果がみられた。信号の能動性に着目したとき、active な動きと passive な動きでは時間方向の畳み込み演算の効果の有無に違いがあり、active な動きでは微細な動きが重要な意味をもち、時間方向の畳み込みサイズを大きくしたことで微細な動きがぼやけてしまっていることが示唆される。また、active な動きは突然動きが変わる場合がある一方で、Subway や Train など passive な動きは車両に乗っていることが多い。したがって、動きの性質上 active な動きはより短時間の変化に着目すべきで、passive な動きはより長い時間で見るべきであることを示唆しており、これは、我々の直感的な感覚とも整合する。

7 おわりに

本論文では、広義に activiy として扱われる認識対象に対して、FFT スペクトラム画像への変換、畳み込みサイズの最適化、CNN 最終段でのセンサ統合を行う識別手法を提案し、その有効性を示した。さらに、行動に伴うセンサ情報には、能動性（Active or Passive）の異なるものがあることを指摘し、その特性によって、認識に最適な畳み込み演算が異なることを示した。周波数方向の畳み込み演算は active と passive ともに有効であるが、時間方向の畳み込み演算は passive においてのみ有効であった。これらの実験結果は、行動認識における標準的な情報処理手法の確立につながるとともに、いわゆる深層学習における畳み込み演算の意味について新たな視点を投げかけるものである。今後、人工生成データなども活用しながら、深層学習における畳み込み演算の役割

Active

(3, 3)

Ground-Truth \ Predicted	Still	Walk	Run	Bike
Still	98.76	0.88	0.00	0.36
Walk	3.85	95.21	0.05	0.88
Run	0.08	1.10	98.34	0.48
Bike	3.11	1.45	0.11	95.32

(3, 1)

Ground-Truth \ Predicted	Still	Walk	Run	Bike
Still	98.74	0.77	0.00	0.49
Walk	3.88	94.93	0.05	1.14
Run	0.09	1.16	98.24	0.51
Bike	2.86	1.17	0.12	95.84

Passive

(3, 3)

Ground-Truth \ Predicted	Car	Bus	Train	Subway
Car	95.71	2.13	0.94	1.22
Bus	3.82	90.39	2.78	3.01
Train	1.47	1.90	84.25	12.38
Subway	2.06	2.66	15.73	79.56

(5, 9)

Ground-Truth \ Predicted	Car	Bus	Train	Subway
Car	95.60	2.21	1.16	1.03
Bus	3.34	90.93	2.84	2.90
Train	1.28	1.98	85.41	11.34
Subway	1.93	2.49	15.55	80.04

図 7: 基準モデル（畳み込みサイズ (3, 3)）と最良畳み込み（畳み込みサイズ (3, 1)（A 群）、(5, 9)（P 群））モデルにおける混同行列の比較。Active 群（上段）、Passive 群（下段）。赤下線：基準モデルより有意に性能が上昇したところ。

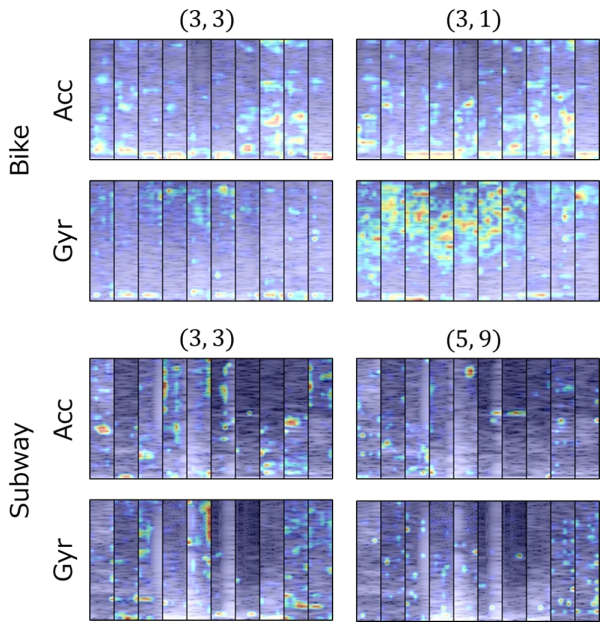


図 8: Grad-CAM による識別寄与領域の可視化。Bike（上 2 段）と Subway(下 2 段）について、2 種のセンサデータで学習を行った CNN モデルを解析したもの。左列：基準モデル（畳み込みサイズ (3,3)）、右列：最良畳み込み（畳み込みサイズ (3,1)、(5,9)）

についてさらなる検討を進めていく予定である。

参考文献

[1] G. Laput and C. Harrison. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ' 19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[2] H. Gjoreski, M. Ciliberto, L. Wang, F. J. Ordonez Morales, S. Mekki, S. Valentin, and D. Roggen. The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices. *IEEE Access*, 6:42592–42604, 2018.

[3] H. Cho and S. Yoon. Divide and Conquer-Based 1D CNN Human Activity Recognition Using Test Data Sharpening. *Sensors (Basel, Switzerland)*, 18, 04 2018.

[4] L. Wang, H. Gjoreski, M. Ciliberto, P. Lago, K. Murao, T. Okita, and D. Roggen. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2019. pages 849–856, 09 2019.

[5] C. Ito, X. Cao, M. Shuzo, and E. Maeda. Application of CNN for Human Activity Recognition with FFT Spectrogram of Acceleration and Gyro Sensors. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp ' 18, page 1503–1510, New York, NY, USA, 2018. Association for Computing Machinery.

[6] T. Hur, J. Bang, T.Huynh-The, J.Lee, J. Kim, and S. Lee. Iss2Image: A Novel Signal-Encoding Technique for CNN-Based Human Activity Recognition. *Sensors*, 18:3910, 11 2018.

[7] W. Jiang and Z. Yin. Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM ' 15, page 1307–1310, New York, NY, USA, 2015. Association for Computing Machinery.

[8] D. Ravi, C. Wong, B. Lo, and G. Yang. Deep learning for human activity recognition: A resource efficient implementation on low-power devices. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 71–76, 2016.

[9] R. Tambi, P. Li, and J.Yang. An Efficient CNN Model for Transportation Mode Sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, SenSys ' 18, page 315–316, New York, NY, USA, 2018. Association for Computing Machinery.

[10] C. Ito, M. Shuzo, and E. Maeda. CNN for Human Activity Recognition on Small Datasets of Acceleration and Gyro Sensors Using Transfer Learning. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC ' 19 Adjunct, page

724–729, New York, NY, USA, 2019. Association for Computing Machinery.

[11] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen. Enabling Reproducible Research in Sensor-Based Transportation Mode Recognition With the Sussex-Huawei Dataset. *IEEE Access*, 7:10870–10891, 2019.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.