博士論文

多変量時系列データからの異常検知を目的とした深層ニューラルネットワークおよびその学習手法に関する研究

三木　大輔

目次

第 1 章

序論

1.1 緒言

多変量時系列データからの異常検知技術は、故障診断や状態監視を実現するために重要な技術であり、統計的手法や機械学習手法に基づく多くの手法が提案されている。原子力関連施設においても時系列データの解析技術は様々な分野で活用が期待されており、例えば動的機器に搭載されたセンサから取得されるデータを解析することで動的機器の故障を検知できる可能性がある。また、原子力関連施設に多く設置されている監視カメラは、核セキュリティのみならず作業者が安全に業務を遂行する上でも欠かせない設備であるが、日々収集される膨大な映像から異常を検知することができれば、安全対策をより強固とできる可能性がある。このようなセンサデータや映像データのような時系列データの解析のため、近年では深層ニューラルネットワーク（Deep Neural Network, DNN）を用いた解析手法がその表現能力の高さから有用な手法として注目されている。しかし、DNN モデルを実環境に適用するためには事前に膨大なデータを用いた学習によりそのパラメータを最適化することが必要である。そのためには取得した時系列データに対し、その異常を決定付ける上で重要な特徴を含む箇所にアノテーションを付与することが必要となるが、データに複雑な特徴が含まれている場合にはその作業は困難である。そこで本研究では時系列データに潜在する特徴を抽出し、異常を検知可能な DNN モデルおよびその学習手法を提案する。さらに人物動作解析のための新たな特徴抽出手法を確立し、先に確立した DNN モデルと組み合わせることで人物動作解析および異常検知を実現する。本章では、研究の背景として原子力関連施設に求められる状態監視技術と映像解析技術について述べた後、本研究の課題と目的についてまとめる。最後に本論文の構成について述べる。

1.2　本研究の背景

東日本大震災では、地震とそれに伴う津波により福島第一原子力発電所への電力供給ラインが損傷し、敷地内外の電力が喪失したことで運転中の原子炉や使用済燃料プールの冷却機能が失われた。その結果として放射性物質が環境中に大量に放出され、世界各国に衝撃を与えた [1]。この事故は多くの近隣住民に避難を強いると共に放射性物質による汚染を福島県を中心とした広域に引き起こし、わが国の国際的立場、経済、エネルギー政策、国民の意識に対して絶大な影響を与えた。我が国政府 [2]、国会 [3]、民間 [4] による事故調査および検証では、安全対策への消極性、想定を超える事象に対する危機感や事故対応の準備不足が指摘されている。こうした事故を二度と引き起こさないためにも、より強固な安全対策が求められており、地震や津波等の天災のみならず枢要機器の故障や不具合、テロリズム等のあらゆる脅威を想定し、それらに対して万全な対策を講じることが求められている。

原子力安全対策は原子力施設の事故・トラブルに対し、その発生確率と安全対策の関係から定められる多重防護を基本的な考え方としており、国際原子力機関（International Atomic Energy Agency, IAEA）は 5 層の安全対策からなる深層防護 [5] を提案している。深層防護における第 1 層および第 2 層は事故発生前、第 3 層以降は事故発生後の対応である。特に事故を未然に防ぐためにも第 1 層おける「異常運転や故障の防止」および第 2 層における「異常運転の制御及び故障の検知」においては、施設の高い信頼性によって安全を確保することが必要である。具体的には、第 1 層では設備に十分な安全性を持たせるほか、定期点検、運転操作についてもきめ細かな管理体制を整備する。また、第 2 層では異常を早期に検知する監視装置および制御装置が必要とされており、この防護が働かない場合には被害が甚大となる恐れがある。このような背景から原子力関連施設において動的機器の保全の重要性が高まっており、機器の劣化傾向を管理する技術が求められている。

一方、事故原因の究明等を通した数多くの報道から、枢要機器の破壊によって人為的に原発事故と同様の損害を与えられることが知れ渡り、悪意を持つ集団にとって原子力発電施設が攻撃対象として認知された可能性も指摘されている。原子力発電施設をテロの標的にするにあたって、原子炉を攻撃せずとも電源を喪失することでメルトダウンが誘発されることが明るみに出たことや、破壊された原子炉建屋内の構造がインターネット等を通じ広く知れ渡ったことは、関心を持つ者が多くの情報を得る機会となっており、我が国政府の報告書 [6] でもこの点ついて言及されている。このような背景から核セキュリティ対策も急務である。

こうした課題に対し、先に述べた状態監視技術や映像監視技術が開発されれば原子力安全に大きく貢献できる可能性がある。具体的には高度な多変量時系列データからの異常検知技術を確立することで、動的機器に搭載されたセンサから取得される時系列データを解析をすることができれば、機器の故障を事前に検知できる可能性がある。また、原子力関連施設に多く設置

されている監視カメラは、核セキュリティのみならず作業者が安全に業務を遂行する上でも欠かせない設備であるが、日々収集される膨大な映像から異常を検知することができれば、安全対策がより強固となることが期待される。

状態監視技術

動的機器の中でも最も頻繁に故障が発生する部位の一つが転がり軸受であり、IEEE Industry Application Society が実施した調査によれば、転がり軸受に関連する故障は回転機器の故障の約 40% を占める [7, 8, 9]。転がり軸受は内輪と外輪およびその間を荷重を受けながら接触する転動体から構成されており、その経年劣化や施工不良、給油不足等の保全上の問題によって摩耗や疲労剥離、焼付き等が生じることがある。このような機器の保全のため、機器の劣化傾向を管理し故障が生じる前の最適な時期に最善の保全を行う状態基準保全（Condition-Based Maintenance, CBM）技術が求められている。高度な CBM 技術を実現することで、予め定められた周期に従い定期的に保全を行う時間計画保全（Time-Based Maintenance, TBM）の代替とすることができれば、事故の発生を未然に防ぐのみならず、機器の稼働期間の延長やコストの削減に繋がる可能性がある。こうした軸受の故障診断法として、振動診断法、潤滑油分析法のほか、温度や電流値等の時間変化を解析する方法等があるが、最も多く現場で用いられる方法が振動診断法である。振動診断法として、軸受に固定された加速度センサ等から取得される振動データに対して包絡線検波処理等を行った後、高速フーリエ変換（Fast Fourier Transform, FFT）により得られるスペクトル情報から異常を検知する方法が一般的である。スペクトル情報から得られたピーク周波数は、軸受に生じた亀裂の箇所と転がり軸受に含まれる転動体の直径と個数、回転周波数、ピッチ円半径および接触角によって決定されることが知られており、これらを用いて故障部位を推定する方法も提案されている。振動診断法は回転機器の故障診断に最もよく用いられる方法であるが、故障の原因やその進行の程度の理解には熟練者や専門家の知識が必要であることが課題である。また、上記のように FFT で得られたスペクトル情報を特徴量として用いる場合には転動体と軸受軌道面との間に摺動が発生しないこと、すなわち転動体が軌道面上を滑ることなく転がることを前提としている。また、複数箇所で同時に亀裂が発生した際には亀裂の大きさによっては必ずしも意図した通りの波形が得られるとは限らない。動的機器の状態監視にはこのような多様な事象に対して頑健な異常検知手法が求められる。さらに、熟練者の不足と後継者の育成が課題になっており、機械学習等を用いた熟練者や専門家の知識に頼ることのない自動的、客観的、定量的な自動診断方法も求められている。上記のような特徴抽出作業は特徴量エンジニアリング等と呼ばれるが、従来の機械学習手法等ではこれらの特徴抽出器の設計は手作業で行われることが多く、結局のところ熟練者や専門家の知識を必要とすることが課題であった。一方で、機械学習手法の中でも特に DNN モデルを用いた手法は表現力の高さからデータの多様性に対して頑健であり、特徴抽出器を設計する作業が少なく実装が容易であり、時系列解析のみならず

様々な分野で活用が期待されている技術の一つである。Janssens らは、加速度センサによって取得された振動データに対し、畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）を用いた方法で、人手で設計された特徴抽出器により得られた特徴量に対してランダムフォレスト分類器を適用する場合と比較して良好な精度で故障の検知を可能としている [10]。これにより従来の機械学習手法では困難であったすべり軸受の劣化のような明示的な特徴周波数を持たない波形から異常を検知できることを報告している。また、Pan らは、DNN モデルに CNN 層と長期短期記憶（Long Short-Term Memory, LSTM）層を組み合わせた方法を採用することで、故障の検知のみならず、故障の識別精度を改善できることを示している [11]。以上のような DNN モデルは一般に誤差逆伝搬法等を用いた DNN モデルの学習により、DNN モデル構造における前半部分の層において適切な特徴抽出、後半部分において識別の機能を持つように学習が行われる。適切に DNN モデルを学習することができれば、煩雑なデータの前処理作業等を必要とせずにデータの解析が可能であるため、熟練者や専門家の知識を必要とする従来手法と比較して有用である。このような DNN モデルを用いてセンサデータのような多変量時系列データから異常を検知することができれば、回転機器のみならず様々な動的機器の状態監視への応用が期待できる。また、データの解析技術を映像解析等に応用することで先述の核セキュリティ等へ適用することも可能と考えられる。

映像監視技術

核セキュリティの重要性は IAEA からも勧告されており、2011 年には核物質及び原子力施設の物理的防護に関する核セキュリティ勧告（INFCIRC/255/Rev.5）[12]、放射性物質及び関連施設に関する核セキュリティ勧告 [13]、規制上の管理を外れた核物質及びその他の放射性物質に関する核セキュリティ勧告 [14] が発行されている。核セキュリティとは「核物質、その他の放射性物質、その関連施設及びその輸送を含む関連活動を対象にした犯罪行為又は故意の違反行為の防止、探知及び対応」のことであり、具体的には核兵器の盗取、放射性物質の盗取、放射性物質拡散のための装置の製造および原子力施設や放射性物質の輸送等に対する妨害破壊行為に対する措置である [15, 16]。これらに対する現行の対策として雇用者の信頼性調査や二人ルール、物理防護システム（Physical Protection System, PPS）等が挙げられる。信頼性調査は、氏名や住所、職歴や海外渡航歴等、本人が自己申告した情報について客観的に信頼性を評価するものである。また、枢要施設での作業時には二人ルールとして、必ず複数の人員で作業を行い互いを監視し合うことで危険な行動を阻止する。PPS は防御・検知・遅延・対応の４つのフェーズから成る防護策であり、防御のフェーズはフェンスや監視員の設置、検知のフェーズとして監視カメラやセンサ等による侵入者の位置情報等を取得し、取得した侵入者の位置に従って遅延のフェーズとしてターンスタイルゲートの制御等で侵入者の目的遂行の時間を遅延させた後、対応のフェーズとして外部武力で侵入者に対応する。ここで、妨害破壊行為を企てる人物は原子力施設関係者以外の外部脅威者と、原子力関連施設に通じる内部脅威者に

大別できる。原子力関連施設は一般的に防護区域・内部区域・枢要区域等の複数の領域に分けられており、それぞれの領域の境界にフェンスやセンサ、警備員、ID 認証等の防護システムによってアクセス権限の無い者の侵入を防いでいるが、原子力施設に通ずる内部脅威者は、施設内部への侵入と内部での作業等が許可されていることから、その妨害破壊行為は多岐に渡って想定される。特に物理的に侵入を阻むことが困難な内部脅威者による妨害破壊行為は脅威であり、我が国政府および内閣府原子力委員会は内部脅威者対策の強化の必要性について言及している [6, 15]。PPS の特性上、このような防御の難しい内部脅威者による妨害破壊行為の検知に失敗すると、遅延させ、対応するといった手続きをとることが困難になることからも妨害破壊行為の検知技術が必要である。このような内部脅威者による妨害破壊行為の検知には、監視カメラ等を利用した行動の監視が有効であるが、先述の通り妨害破壊行為の検知には早期の対応が必要であることや、監視すべき映像が膨大であることからも映像を解析する技術が有効と考えられる。また、内部脅威者による妨害行為は通常の作業になりすまされる可能性があるため、古典的な顔認識技術や服装の識別、単純な物体認識技術等は有効ではない。つまり、映像データから人物の動作に関する特徴量を適切に解析し、異常を検知する方法が求められる。

映像からの異常検知技術として特徴量エンジニアリングにより映像から得られた特徴量から異常を判別する方法がある。代表的な特徴量エンジニアリングに基づく方法として Otsu らは立体高次局所自己相関（Cubic Higher-order Local Auto-Correlation, CHLAC）特徴量を利用した異常行動検知手法 [17] を提案している。一方で、深層学習を利用した映像からの異常検知手法として Sultani らはマルチインスタンス学習およびランク学習に着想を得た DNN モデルの弱教師あり学習手法を提案し、映像からの異常検知に適用可能であることを報告している [18]。本手法は特徴量エンジニアリングの作業を必要としない点で上記のような手法と比較して有利である。しかし、映像中の人物の詳細な動作を識別するようなタスクにおいては良好な認識精度が得られなかった。これは、本先行研究で用いられた DNN モデルが映像データの全体から異常を判別する構造になっており、人物の動作のような映像中の局所的な特徴量を十分に抽出することが困難であったことに起因すると考えられる。映像データのような高次元データを解析するためには、人物の行動を理解する上で必要十分な特徴量の抽出を行うことが有効である。映像から人物の行動を認識するための適切な特徴抽出方法として、映像中の人物の姿勢を認識する技術がある [19, 20]。Shotton らは赤外線カメラやステレオカメラから得られる距離画像から人物の姿勢を認識する方法 [21] を提案している。また、Toshev らは DNN モデルを用いることで距離画像に代えてに監視カメラ映像として一般的に用いられている RGB 画像から直接人物の姿勢を認識することを可能としている [22]。さらに、最近ではその高度化およびオープンソースソフトウェア化が進んでおり、開発がより容易になりつつある [23]。これらの技術を応用することで、Chen らは人物の手指の動作の時系列データから人物の動作を解析し、核セキュリティへ適用する方法を提案している [24, 25]。本先行研究では、距離画像から人物の行動を決定づける上で重要な手指の姿勢情報の認識を可能とし、さらに手指の動作の

時系列データを解析することで、切る、叩く、まわす等の機器を破壊するような動作の識別やその発生時刻の特定を可能としている。このような技術を人物の手指の動作のみならず、人物の全身の動作解析へ応用することで、上記のような核セキュリティへの応用が期待される。

1.3　本研究の課題と目的

東日本大震災以降、原子力関連施設では地震や津波等の天災のみならず、枢要機器の故障や不具合、テロリズム等あらゆる脅威を想定し、それらに対して万全な対策を講じることが求められている。こうした課題に対し高度な状態監視技術や映像監視技術が開発されれば、原子力安全に貢献できる可能性がある。状態監視や映像監視を実現する技術として、統計的手法や機械学習手法を応用したセンサデータや映像データを解析する技術が注目されており、特に DNN モデルを用いた手法はその表現能力の高さから、多変量時系列データに含まれる異常の複雑な特徴を抽出し、検知する上で有効な手法として期待されている。時系列データからの異常検知に DNN モデルを用いる方法は、多変量時系列データを取り扱う上で有効であるほか、煩雑な特徴量エンジニアリングの作業が少なく済むため実応用の上でも有利である。DNN モデルを用いる課題として、事前に大規模なデータセットを用いた学習により、そのパラメータの最適化が必要なことが挙げられる。特に DNN モデルを時系列データの解析に応用するためには、一連のデータおよびそのデータを説明するアノテーションの対から成るデータセットを用意することが必要となる。現実的に扱われる異常を含む時系列データは、多変量時系列データであることが多いが、このような異常を含む多変量時系列データは一連のデータの中に異常が含まれていることを把握することが可能である場合においても、それらの異常が含まれる箇所やその程度を把握することは一般に困難である。

以上のような課題に対して本研究では多変量時系列データに潜在する特徴を抽出し、検知することが可能な DNN モデルおよびその学習手法を提案する。また、映像中の人物動作特徴量を抽出する手法を併せて開発することで映像解析への応用を行う。本研究の目的は、動的機器の状態監視や核セキュリティのための映像監視を目的とした多変量時系列データからの異常検知が可能な DNN モデルおよびその学習手法を確立することである。そのために本研究では、まず多変量時系列データから異常を検知するための DNN モデルとその学習手法を確立し、異常の検知のみならず異常の識別が可能となるように学習手法を改良する。さらに人物動作解析のための新たな特徴抽出手法を確立し、先に確立した DNN モデルと組み合わせることで人物動作解析および異常検知を実現する。

1.4　本論文の構成

本論文では、まず第 2 章で多変量時系列データの解析のための DNN モデルおよびその学習手法について述べ、時系列データからの異常検知への適用可能性について評価した後、実データに対する適用を通して本手法の有用性を確認する。第 3 章では DNN モデルを異常検知のみならず異常の識別へ応用するための学習手法を確立する。第 4 章では第 2 章および第 3 章で確立した多変量時系列データの解析技術を映像解析へ応用するために、映像から適切に特徴を抽出するための DNN モデルとその学習手法および評価について述べる。特に映像監視への応用に有利な広角撮像系により得られた映像を解析するための工夫について述べる。以上で確立した技術を組み合わせることで、第 5 章では人物動作からその行動を識別する手法および異常を検知する方法について述べる。最後に第 6 章で本論文をまとめる。

第 2 章

深層ニューラルネットワークおよびその弱教師あり学習手法

2.1 緒言

時系列データの解析技術は、機器の故障診断、状態監視、映像監視等を実現する上で重要な技術であり、統計的手法や機械学習手法等に基づく様々な手法が提案されている。特に DNN モデルを用いた手法はその表現能力の高さから、多変量時系列データに含まれる異常のもつ複雑な特徴を抽出し検知する上で有利な手法として注目されている。DNN モデルを時系列データ解析に適用するには、事前に時系列データおよびそれらに含まれる異常に関するアノテーションの対から成るデータセットを用いた学習により、そのパラメータを最適化することが必要である。さらに、時系列データにアノテーションを付与する際には、データにおける異常が含まれる箇所やその程度を定量的に把握することが必要であるが、一般にそれらは未知であることが多いため、アノテーションの付与作業は困難である。そこで、本研究ではそのようなアノテーションの付与が困難な時系列データから適切に異常を検知可能な DNN モデルおよびその学習手法を確立する。

本章ではまず、ニューラルネットワークに関する基本的な構成要素および学習手法について述べる。次に、ニューラルネットワークを時系列データ解析に適用する方法を述べた後、時系列データ中の異常が含まれる箇所に対してアノテーションが付与されていない弱教師ありデータを用いて、異常を検知可能な DNN モデルを学習する方法について述べる。最後に、提案手法が異常の検知のみならず定量に有効であることを確認した後、精度や定量特性、実データへの適用について述べる。

2.2 方法

2.2.1 ニューラルネットワークの構成要素

ニューラルネットワークは生物の神経回路を模倣して数理的にモデル化したものである。McCulloch と Pitts の研究 [26] では、図 2.1 のような形式ニューロンと呼ばれる素子を定義した。形式ニューロンは実際の生物のニューロンと同様に多数の入力信号 $\{x_1, x_2, \ldots, x_D\}$ を受け付ける。生物のニューロンにおいてはニューロン同士の結合の強さの度合いが異なることが知られており、形式ニューロンにおいても同様に結合の強さを表す重み $\{w_1, w_2, \ldots, w_D\}$ を導入することで層への総入力を

$$u = \sum_i w_i x_i \tag{2.1}$$

とし、この u を受け、次のニューロンに

$$h(u+b) = \begin{cases} 1 & (u \geq -b) \\ 0 & (u < -b) \end{cases} \tag{2.2}$$

を出力値として伝搬する。ここで h はヘヴィサイド関数であり、b は閾値を与えるパラメータである。以上まとめると

$$z = h(u+b) = h\left(\sum_i w_i x_i + b\right) \tag{2.3}$$

となる。ここで用いたヘヴィサイド関数のように総入力 u を出力 z に変換する関数を一般に活性化関数と呼ぶ。

Rosenblatt らは以上のような形式ニューロンを複数組み合わせ、それらの重み $\mathbf{w} = (w_1, w_2, \ldots, w_D)^\top$ および b を固定値ではなく、学習可能とした回路をパーセプトロンと呼んだ [27]。一般に複数の形式ニューロンの出力は次の形式ニューロンに入力され、各層の出力が次層へと入力される構造を繰り返す。初めの層は入力層と呼ばれ、ベクトル $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top$ の各成分 x_i を入力値として持つ。また、最後の層を出力層と呼び、ベクトル $\mathbf{y} = (y_1, y_2, \ldots y_K)^\top$ の各成分 y_j を出力値として持つ。さらに、それ以外の層を隠れ層と呼ぶ。第 l 層における j 番目のノードにおける総入力は

$$u_j^{(l)} = \sum_i w_{ji}^{(l)} z_i^{(l-1)} \tag{2.4}$$

で表される。さらに、ノードの出力は

$$z_j^{(l)} = g^{(l)}\left(u_j^{(l)} + b_j^{(l)}\right) = g\left(\sum_i w_{ji}^{(l)} x_i^{(l)} + b_j^{(l)}\right) \tag{2.5}$$

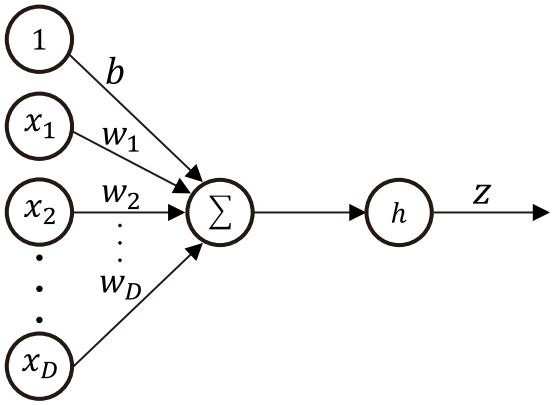


図 2.1 形式ニューロンの模式図

である。ここで、g は活性化関数であり、シグモイド関数

$$g(x) = \frac{1}{1 + e^{-\beta x}} \tag{2.6}$$

や、正規化線形関数（Rectified Linear Unit, ReLU）[28]

$$g(x) = \max(x, 0) \tag{2.7}$$

等が用いられる。

畳み込みニューラルネットワーク

畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）は畳み込み層およびプーリング層と呼ばれる層を持つ DNN モデルである。入力データが長さ T の D 変量の時系列データのような系列データであれば $T \times D$、高さ H、幅 W の D チャネルをもつ画像のような格子構造を持つデータであれば $H \times W \times D$ の大きさを持つデータと表すことができる。このようなデータに対する幅 L の 1 次元フィルタによる畳み込み演算における出力 u_i はフィルタ w_{id}（$i \in \{0, \ldots, L-1\}, d \in \{1, \ldots, D\}$）およびバイアス b_d を用いて

$$u_i = \sum_d \sum_{p \in \mathcal{P}_i} x_{pd} w_{p-i,d} + b_d \tag{2.8}$$

と算出される。また、大きさ $L \times L$ の 2 次元フィルタによる畳み込み演算における出力 u_{ij} は、同様にフィルタを w_{ijd}（$i \in \{0, \ldots, L-1\}, j \in \{0, \ldots, L-1\}, d \in \{1, \ldots, D\}$）およびバイアス b_d を用いて

$$u_{ij} = \sum_d \sum_{(p,q) \in \mathcal{P}_{ij}} x_{pqd} w_{p-i,q-j,d} + b_d \tag{2.9}$$

と算出される。ここで、$\mathcal{P}_i$ および $\mathcal{P}_{ij}$ は

$$\mathcal{P}_i = \{si + i' | i' = \{0, \ldots, L-1\}\} \tag{2.10}$$

$$\mathcal{P}_{ij} = \{(si + i', sj + j') | i' = \{0, \ldots, L-1\}, j' = \{0, \ldots, L-1\}\} \tag{2.11}$$

である。ここで s はストライドと呼ばれる係数であり、フィルタの移動幅を表す。

以上は層間のノードが特殊な形式で接続された疎な接続を持つネットワークとして説明できる。具体的には上位層の各ノードは下位層の一部のノードのみと接続されており、これを受容野が局所的であると呼ぶ。また、その接続の重みを表すフィルタは各ノード間で共有され、これを重み共有と呼ぶ。これらの仕組みにより DNN モデルへの入力変化に対する頑健性・メモリ容量の削減が実現される [29]。本研究では 1 次元的な畳み込みを行うニューラルネットワークを 1DCNN（One-Dimensional Convolutional Neural Network）、同様に 2 次元的な畳み込みを行うニューラルネットワークをを 2DCNN（Two-Dimensional Convolutional Neural Network）と標記する。

プーリング

プーリング層の目的はデータに対するフィルタの応答の強さに関する情報を一部切り捨て、データ内に含まれる特徴の微小な位置変化に対する応答の不変性を実現することである。プーリング層におけるノード (i, j) は畳み込み層と同様に、その入力側の層に受容野 $\mathcal{P}_{ij}$ を持つ。ノード (i, j) の出力は受容野 $\mathcal{P}_{ij}$ の内部のノード $(p, q) \in \mathcal{P}_{ij}$ の出力 z_{pq} を集約したものである。集約方法には、受容野 $\mathcal{P}_{ij}$ に属するノードからの入力の平均値

$$z_{ijd} = \frac{1}{|\mathcal{P}_{ij}|} \sum_{(p,q) \in \mathcal{P}_{i,j}} u_{p,q,d} \tag{2.12}$$

をノードの出力とする平均プーリングと呼ばれる方法と、受容野 $\mathcal{P}_{ij}$ に属するノードからの入力の最大値

$$z_{ijd} = \max_{(p,q) \in \mathcal{P}_{i,j}} u_{p,q,d} \tag{2.13}$$

をノードの出力とする最大プーリングと呼ばれる方法等がある。一般に入力が複数チャネルを有する場合にはチャネルごとに独立して以上の処理を行う。その他、これらの中間的な方法として L_p プーリング等があるが、本研究では平均プーリングおよび最大プーリングを扱う。

再帰構造をもつニューラルネットワーク

回帰結合ニューラルネットワーク（Recurrent Neural Network, RNN）はネットワーク構造に再帰的な構造を含み、出力が次の時刻における自身の入力になるニューラルネットワークの総称である。RNN はネットワーク内に過去の状態を記憶する構造が含まれるため、順伝搬型ネットワークと比較して時系列データを扱う上で有利である。しかし、多層の RNN においてはその学習時に誤差逆伝搬法 [30] による差分信号が意図した大きさで伝搬せず、隠れ層を経るごとに勾配が小さくなる勾配消失が生じ、学習が上手く行われないことが知られてい

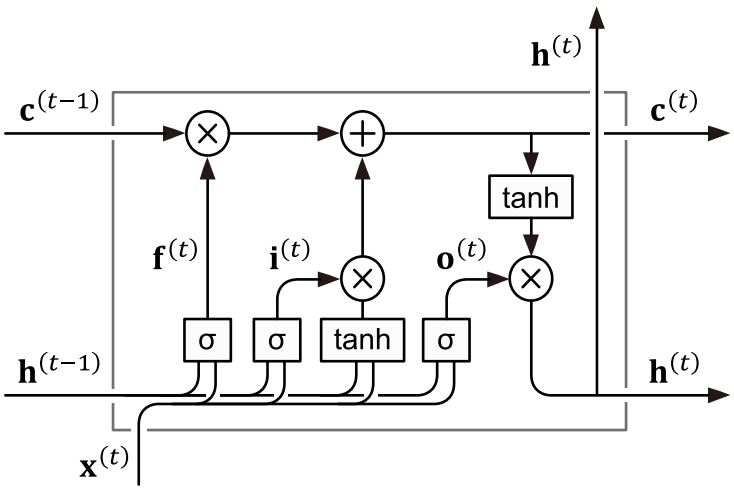


図 2.2 LSTM の模式図

る。長短期記憶ネットワーク（Long Short-Term Memory, LSTM）[31] はこれを解消するため、RNN の中間層をメモリセルおよびゲートと呼ばれる素子から成る層で置き換えたものである。図 2.2 に本研究で扱う LSTM の構造を示す。時刻 t における入力 $\mathbf{x}^{(t)}$ およびメモリセルの値を $\mathbf{c}^{(t)}$ としたとき、メモリセルの値は入力 $\mathbf{x}^{(t)}$、入力ゲート $\mathbf{i}^{(t)}$ の値および忘却ゲート $\mathbf{f}^{(t)}$ の値と、時刻 $t-1$ のメモリセルの値 $\mathbf{c}^{(t-1)}$ および出力 $\mathbf{h}^{(t-1)}$ を用い

$$\mathbf{c}^{(t)} = \mathbf{f}^{(t)} \odot \mathbf{c}^{(t-1)} + \mathbf{i}^{(t)} \odot \tanh(\mathbf{W}_{cx}\mathbf{x}^{(t)} + \mathbf{W}_{ch}\mathbf{h}^{(t-1)} + \mathbf{b}_c) \tag{2.14}$$

となる。ここで、$\odot$ はアダマール積を表し、重み $\mathbf{W}_c$ およびバイアス $\mathbf{b}_c$ が学習すべきパラメータである。メモリセルおよび出力ゲートの値を用い LSTM 層の出力は

$$\mathbf{h}^{(t)} = \mathbf{o}^{(t)} \odot \tanh(\mathbf{c}^{(t)}) \tag{2.15}$$

となる。時刻 t における入力ゲート $\mathbf{i}^{(t)}$ 忘却ゲート $\mathbf{f}^{(t)}$、出力ゲート $\mathbf{o}^{(t)}$ の各ゲートにおける値は入力データ $\mathbf{x}^{(t)}$ および時刻 $t-1$ における出力 $\mathbf{h}^{(t-1)}$ を用い

$$\mathbf{i}^{(t)} = \sigma\left(\mathbf{W}_{ix}\mathbf{x}^{(t)} + \mathbf{W}_{ih}\mathbf{h}^{(t-1)} + \mathbf{b}_i\right), \tag{2.16}$$

$$\mathbf{f}^{(t)} = \sigma\left(\mathbf{W}_{fx}\mathbf{x}^{(t)} + \mathbf{W}_{fh}\mathbf{h}^{(t-1)} + \mathbf{b}_f\right), \tag{2.17}$$

$$\mathbf{o}^{(t)} = \sigma\left(\mathbf{W}_{ox}\mathbf{x}^{(t)} + \mathbf{W}_{oh}\mathbf{h}^{(t-1)} + \mathbf{b}_o\right) \tag{2.18}$$

と表される。以上の仕組みを導入することで出力 $\mathbf{h}$ はメモリセルの状態と各ゲートの出力の加算演算およびアダマール積により算出されるため、RNN のような行列同士の乗算を繰り返さない。そのため、誤差逆伝搬時に勾配消失を回避でき学習を適切に進める上で有利である。

バッチ正規化

ニューラルネットワークの学習において学習用データをサンプリングした際の生成分布と、評価用データの分布に乖離が起き、共変量シフトが生じる。共変量シフトを防ぐためには、学習時に各層の出力値が従う分布が一定になるように調整が必要である。バッチ正規化 [32] は正則化手法の一つであり、隠れ層における出力を正規化し、出力を常に平均 0、分散 1 の分布に従うように調整する。あるミニバッチ $\mathcal{B}$ における入力 $\{x_1, x_2, \ldots, x_M\}$ を次の変換則に従い $\{z_1, z_2, \ldots, z_M\}$ に変換する。

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{M} \sum_{n \in \mathcal{B}} x_n, \tag{2.19}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{M} \sum_{n \in \mathcal{B}} \left(x_n - \mu_{\mathcal{B}}\right)^2, \tag{2.20}$$

$$\hat{x}_n \leftarrow \frac{x_n - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}, \tag{2.21}$$

$$z_n \leftarrow \gamma \hat{x}_n + \beta. \tag{2.22}$$

ここで、γ と β は学習すべきパラメータである。Ioffe らは論文内でバッチ正規化を用いることで、より以前から用いられていたドロップアウト層 [33] を用いずに DNN モデルの学習が行われることを示している [32]。また、Zhang らはバッチ正規化が DNN モデルの学習を安定させ、学習の収束を早める役割を担うと述べている [34]。

本研究では以上で述べたニューラルネットワークの構成要素を組み合わせることで DNN モデルを構築する。次項ではニューラルネットワークの学習手法について述べる。

2.2.2　ニューラルネットワークの学習手法

パラメータ $\mathbf{w}$ をもつニューラルネットワーク f の学習では、一般に入力データ $\mathbf{x}$ に対する推定値 $f(\mathbf{x}; \mathbf{w})$ と真値 y との誤差を最小化するようにパラメータ $\mathbf{w}^*$ を得ることを目標とする。つまり、データ集合 $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1,\ldots,N}$ において、ニューラルネットワークの学習は次のような二乗和誤差で与えられる損失関数の最小化として定式化できる

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}), \tag{2.23}$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y_n - f(\mathbf{x}_n; \mathbf{w})\right)^2. \tag{2.24}$$

現実的にはこのような最適化問題を解析的に解くことは困難であるため、十分な入力データと真値の対 $\mathcal{D}$ を用意し、損失 $E(\mathbf{w})$ を最小化するように反復的に学習を行うことが有効である。$\mathbf{w}^*$ の反復的求解方法として損失関数の 1 階微分を用いる勾配降下法がある。勾配降下法では、現在のパラメータ $\mathbf{w}^{(\tau)}$ における勾配

$$\nabla E(\mathbf{w}^{(\tau)}) = \left.\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^{(\tau)}} \tag{2.25}$$

を求め、

$$\mathbf{w}^{(\tau+1)} \leftarrow \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)}) \tag{2.26}$$

のようにパラメータの更新を行う。ここで、η は学習率を表す。DNN モデルの各層の重みパラメータに対してそれぞれの勾配を計算するには膨大な計算量が必要となるが、誤差逆伝搬法 [30] を用いることで効率的に計算を行うことができる。

この学習手法を時系列データ解析のための DNN モデルに適用するためには、損失 $E(\mathbf{w})$ の算出のため、一連の時系列データに対しその特性を決定付ける上で重要な箇所に対してアノテーションを付与し、学習に供することが有効である。つまり、時系列データにおける各時点の異常の大きさの程度が既知であれば、時点 t における DNN モデルの推定値 $f(\mathbf{x}^{(t)}; \mathbf{w})$ と真値 $y^{(t)}$ を用い、二乗和誤差

$$E(\mathbf{w}) = \sum_t \left(y^{(t)} - f(\mathbf{x}^{(t)}; \mathbf{w}) \right)^2 \tag{2.27}$$

を最小化するように DNN モデルのパラメータを最適化すればよい。しかし、深層学習で用いられるような膨大な時系列データにおける、それぞれの時点にアノテーションを付与する作業は困難であり、実際には t 時点における異常度の真値 $y^{(t)}$ は未知であることが多い。このようにデータに特定の事象が含まれていることが既知であるが、その事象が含まれる位置やその程度が未知であるデータを本研究では弱教師ありデータと呼ぶ。本研究では、このような弱教師ありデータを用いて DNN モデルを学習可能とし、異常検知へ応用する。

2.2.3 弱教師ありデータを用いたニューラルネットワークの学習手法

時系列データからの異常検知に DNN モデルを適用するためには、時系列データに対し異常を含む箇所のそれぞれにアノテーションを付与することが有効であるが、それらに含まれる異常箇所やその程度について定量的に把握しアノテーションを付与することは困難である。このようなデータの性質は弱教師あり学習手法の一つであるマルチインスタンス学習 [35, 36] で扱われるものと類似している。一般的な教師あり学習では、入力データ要素の集合と、それぞれの要素に対するアノテーションから成る学習用データセットが与えられるが、マルチインスタンス学習では個々の要素ではなく要素の集合ごとに教師信号が与えられる。データ集合に含ま

れるそれぞれの要素に対し教師信号を付与するには煩雑な作業を要するが、データ集合に教師信号を付与することは比較的容易である。Sultani らはマルチインスタンス学習に着想を得ることで、弱教師ありデータを用いた DNN モデルの学習手法を提案し、異常検知へ適用可能であることを報告している [18]。そこで、本研究ではこれらの先行研究を参考に時系列データを扱うための DNN モデルおよび学習手法の改良を行う。先行研究で提案された手法により学習された DNN モデルは、弱教師ありデータを用いて上手く学習を行えることから、映像データに含まれる異常のように定義が曖昧な事象の検知に有効である。しかし、先行研究では DNN モデルの構造が時間的な繋がりを考慮した構造となっていないことや、DNN モデルの出力が異常を含む、または含まないの二値識別に限られ、異常の識別にはそのまま適用できないことが課題であったため、それらの改良を行う。本章では時系列データからの異常検知のための基礎的検討を行い、次章では識別問題への応用を行う。

本研究で扱う時系列データからの異常検知のための DNN モデルは、長さ T の時系列データ $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$ を入力した際に $\{y^{(1)}, \ldots, y^{(T)}\}$ を出力する構造とし、入力データにおける異常の有無は既知であるがそれらが含まれる箇所は未知であるものとする。DNN モデルの学習では、まず時系列データに異常を含む（正）または、含まない（負）によって各データ集合に分割し、それぞれのデータを DNN モデルに入力した際に、正のデータ集合に対して高い値が、負のデータ集合に対して低い値が出力されるように DNN モデルのパラメータを最適化する。それぞれのデータ集合は、それぞれの要素の中に少なくとも一つ以上の要素の教師信号が正である正の集合、およびすべての要素の教師信号が負である負の集合に分割される。各データにおいてその特性を決定付ける上で重要な箇所は未知であるため、ここでは各データ集合から推定される最大値に着目し、

$$\max_t f\left(\mathbf{x}_{pos}^{(t)}\right) > \max_t f\left(\mathbf{x}_{neg}^{(t)}\right) \tag{2.28}$$

を満たすように DNN モデルを学習する。ここで、$f\left(\mathbf{x}_{pos}^{(t)}\right)$ および $f\left(\mathbf{x}_{neg}^{(t)}\right)$ はそれぞれ正および負のデータを DNN モデルに入力した際に出力される $t \in \{1, \ldots, T\}$ 時点における推定値である。このような条件を満たす DNN モデルの学習では、以下の損失関数

$$E = \max\left(0, 1 - \max_t f\left(\mathbf{x}_{pos}^{(t)}\right) + \max_t f\left(\mathbf{x}_{neg}^{(t)}\right)\right) + \lambda \tag{2.29}$$

を最小化するように DNN モデルのパラメータを最適化する。ここで λ は

$$\lambda = p_1 \sum_{t=1}^{T-1} \left(f\left(\mathbf{x}_{pos}^{(t)}\right) - f\left(\mathbf{x}_{pos}^{(t+1)}\right)\right)^2 + p_2 \sum_{t=1}^{T} f\left(\mathbf{x}_{pos}^{(t)}\right) \tag{2.30}$$

で表される正則化項であり、DNN モデルの学習用データへの過学習を防ぎ学習を安定させるために導入した。第 1 項は平滑化項であり、推定値における前後の時点で大きな変動がないように調整する項である。第 2 項はスパース化項であり、異常が長期に渡って生じることが考

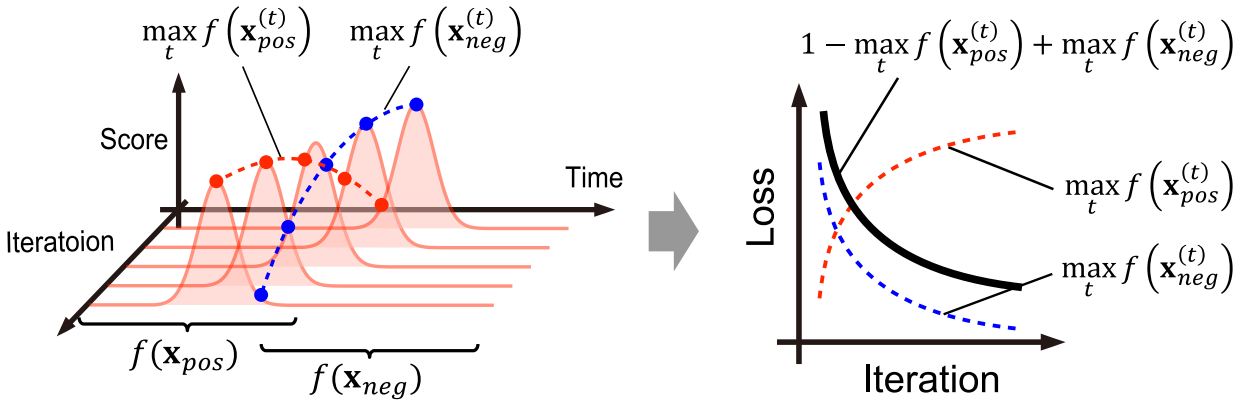


図 2.3　提案する損失関数の概要

えづらい場合や、その頻度が少ないことが想定される場合に誤検知を低減する効果がある。それぞれの項に、ハイパパラメータ p_1 または p_2 が設けられてあり、実際の問題に適用する際にはこれらを調整することで正則化の度合いが調整される。提案する DNN モデルの学習は式 (2.29) の損失関数の最小化問題となるが、一般的な DNN モデルの学習と同様にこれを解析的に解くことは困難であるため、反復的にパラメータを更新することで近似的な解を得る。

これらの学習が上手く行われることを図 2.3 を用いて示す。図 2.3（左）では横軸が時間、縦軸が DNN モデルにより推定される値の大きさ、奥から手前に向かって DNN モデルの学習の進行に伴う推定値の変化を示している。問題設定では、ある十分な長さをもつ区間における異常の有無は既知であるため、異常を含む区間に対する推定値と異常を含まない区間に対する推定値の最大値を描画すると図 2.3（右）のようになる。異常を含む区間に対する推定値と異常を含まない区間に対する推定値の最大値がこのように変化することで、損失関数（2.29）が順調に最小化されることがわかる。

図 2.4 もまた提案する損失関数を最小化するように DNN モデルのパラメータを最適化する際に DNN モデルの推定値がどのように変化していくかを模式的に示したものであるが、特に DNN モデルの推定値の初期値が極端に高い場合、または低い場合にも頑健となることを示したものである。DNN モデルのパラメータの初期値は適当な事前分布をもとに初期化されるため、それらの変動に頑健となることが好ましいが、大きく分けて全体的に高い値を推定する真陽性となりやすいが偽陽性にもなりやすい状態、または全体的に低い値を推定する、真陰性となりやすい一方で偽陰性にもなりやすい状態の 2 通りが考えられる。図 2.4（上）の場合は負のデータに対する一連の推定値のうち 1 時点でも高い値を推定してしまうと損失が大きくなり、それが改善されるに従って損失が小さくなることを示している。一方で図 2.4（下）の場合は正のデータに対する一連の推定値のうち最も高い値が適切に推定されていない際に損失が大きくなり、それが改善されるに従って損失が小さくなることを示している。以上から、DNN モデルの初期値の変動に頑健な学習が可能となるように損失関数が設計されていること

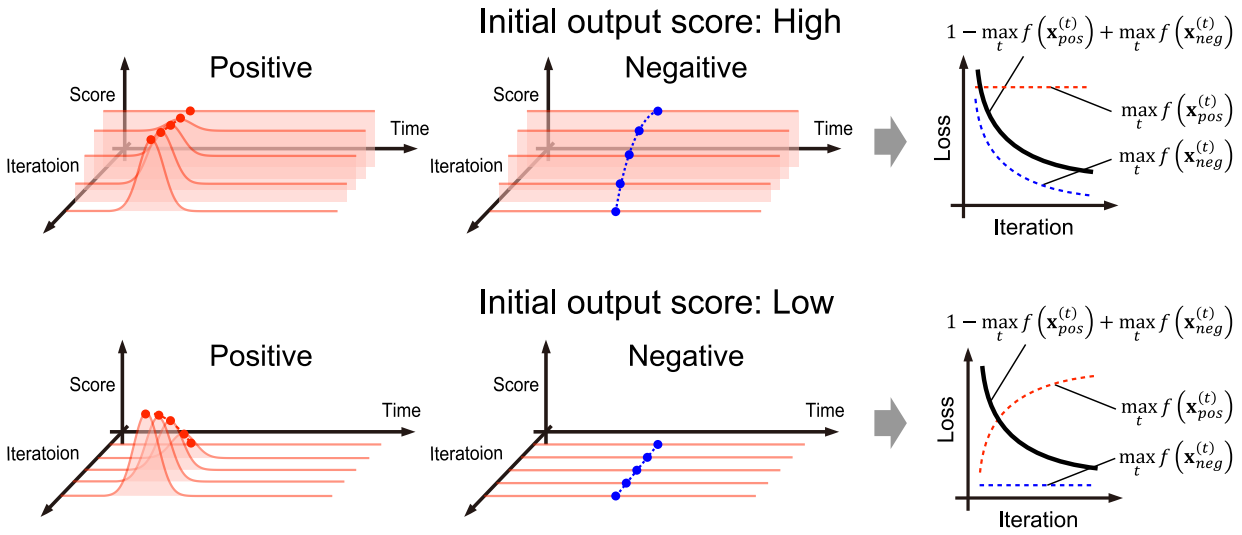


図 2.4　推定される値の初期値が（上）高い場合および（下）低い場合における推定値と損失の大きさの変化

がわかる。

2.3　実験

提案手法の有用性を評価するために、まず、人工的に生成された波形データからの外れ値検知を行うことで提案手法により学習された DNN モデルが異常検知に適用可能か評価する。さらに、実システムへの適用可能性を評価するために故障を含む軸受に装着された加速度センサから取得された振動データを用いた評価を行う。

2.3.1　異常の検知および定量に関する評価

人工的に生成された波形として 1 kHz で取得された 2 つの正弦波信号を模擬し、それぞれの合成波形に対し平均 $\mu_f = 0$、標準偏差 $\sigma_f = 0.05$ の正規分布に従うノイズを印加した波形を用意した。それぞれの正弦波の周波数は $0-4$ Hz とした。外れ値を含む波形として、上記と同様の条件で生成された波形に平均 $\mu_o = 0$、標準偏差 $\sigma_o = 0.20$ の正規分布に従う外れ値を各時点に対して 1% の確率で印加した。長さ $1,000$ 時点をもつ外れ値を含む波形および含まない波形をそれぞれ正のデータおよび負のデータとして $10,000$ 組ずつ生成し実験に供した。本実験に用いる DNN モデルは一連の時系列データを順に 10 時点ずつ入力として受け付け、それぞれの時点に対する異常の度合いを推定値として出力する。DNN モデルの構造は図 2.5 に示す 3 層の構造とし、隠れ層には 128 チャネル、カーネル幅 9 を有する畳み込み層、最終層には 128 チャネル、カーネル幅 1 を有する畳み込み層を用いた。1 層目および 2 層目の畳

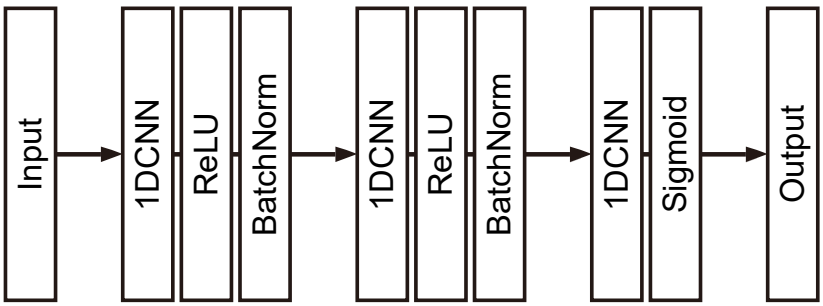


図 2.5 外れ値検知のための DNN モデルの構造

み込み層からの出力に対しては ReLU 活性化およびバッチ正規化を適用し、最終層の出力に対してシグモイド活性化を適用した。畳み込み層のそれぞれの重みパラメータは He の初期値 [37]、バッチ正規化における γ を平均 1 分散 0.01 の正規分布に従う値、β を 0 で初期化した。ハイパパラメータを $p_1 = 10^{-3}$、$p_2 = 10^{-3}$、学習係数 10^{-3} とし、式 (2.29) の損失関数を最小化するように $10,000$ 回のパラメータ更新を行った。上記を Python スクリプトとして実装し、Ubuntu18.04 環境における Python3.7.4 インタプリタ上で実行した。DNN モデルの実装には Tensorflow2.0.0 を用いた。学習および評価には GPU（NVIDIA TITAN V）を搭載したワークステーション（Intel Xeon E5-2698v4, 50M Cache, 2.20 GHz）を用いた。以降の実験についても同様の環境を用いて実施した。

2.3.2 軸受故障検知への応用

実環境により得られたデータからの異常検知が可能であるか確認するため、軸受故障診断の実現可能性を評価した。第 1 章で述べた通り、軸受は回転機器の重要な構成要素であり、その故障を早期に検知する技術の開発を目的とした様々なデータセットが公開されている [38, 39]。実験には、以下の 2 つのデータセットを用い、様々な故障事例に適用可能か評価した。

CWRU データセット

Bearing data Center of Case Western Reserve University（CWRU）[38] は軸受に生じる複数種類の故障を模擬したデータセットを公開している。本データセットは図 2.6 のように、電気モータ（左）、トルク変換器/エンコーダ（中央）、動力計（右）および制御回路で構成された系から取得された振動データが記録されている。用いられた軸受には放電加工により転動体、内輪および外輪のそれぞれに亀裂を恣意的に加えている。データは正常（Normal, N）な軸受から取得された波形および、内輪（Inner Race, IR）、転動体（Ball, B）、外輪の 3 時方向（Outer Race, OR@3）、6 時方向（OR@6）および 12 時方向（OR@12）に亀裂を生じた 5 種類の故障を模擬した軸受から取得された波形から成る（図 2.7）。振動データは、モータのドラ

イブ側およびファン側の両方の 12 時方向の位置に取り付けられた加速度計を用いて取得されている。本実験で用いたデータは 1,797 rpm の回転速度に対してサンプリング周波数 12 kHz で記録されており、1 回転あたり約 401 時点分のデータが取得されている。本実験では、それぞれの波形データを約半回転に相当する 200 時点のデータを含む 605 波形に分割し、さらにそれぞれを 9 : 1 の割合で学習および評価に用いた。それぞれの波形データには故障に特徴的な波形を含むが、上記の実験と同様にそれらの含まれる位置や程度に関する定量的な情報を用いずに DNN モデルの学習を行い、異常を適切に検知できるか確認した。

IMS データセット

上記のような軸受故障診断技術の開発を目的とした多くのデータは、事前に恣意的に故障させた軸受から取得されたもの、故障発生後に実機から回収された軸受から取得されたもの、シミュレータにより仮想的に生成されたもの等が用いられる。また、軸受の故障を模擬するために、軸受構成要素の表面に傷をつける方法や、潤滑油に異物を混入させたりする方法がある。このような方法は短時間で効率よくデータを収集する上で有利であるが、以上のような方法で取得されたデータを用いた実験方法では、故障初期の段階や自然な故障の発生を模擬することは困難である。NSF I/UCR Center for Intelligent Maintenance Systems（IMS）が公開するデータセット [39] では、実際に軸受に故障が生じる過程を記録するために設計された試験装置を用いてデータの取得が行われている。本試験装置（図 2.8）は、1 本のシャフトに対し 4 台の軸受が搭載されており、シャフトは AC モータで駆動され、回転速度は 2,000 rpm で一定に保たれている。すべての軸受に対し強制給油を行い、ばね機構によって軸および軸受に 6,000 lbs のラジアル荷重がかけられている。本データセットには様々な故障のデータが含まれるが、実験には軸受 3 および軸受 4 に故障が発生した際のデータの一部のみを用いた。それぞれの故障発生後の各部品の外観と故障した軸受から取得された各加速度データを図 2.9 および図 2.10 に示す。それぞれの故障は軸受の設計寿命を超えた後に発生しており、データは 10 分ごとに記録された 2,156 の計測データから成る。これらの計測データには、それぞれ 20,480 時点の振動データが含まれている。[39] で報告されたように、本データには 1,560 番目の計測以降に故障が発生していることが想定されるため、実験では 1,260 番目以前の計測により得られたデータを負のデータ、1,860 番目以降の計測により得られたデータを正のデータとして学習に供した。それぞれのデータにおける 20,480 時点についてデータの最初の 90%（18,432 時点）を学習用データ、残りの 10%（2,048 時点）を評価用データとした。上記の実験と同様に異常の含まれる箇所や程度などの定量的な情報を用いずに DNN モデルの学習を行い、異常を検知可能か確認した。

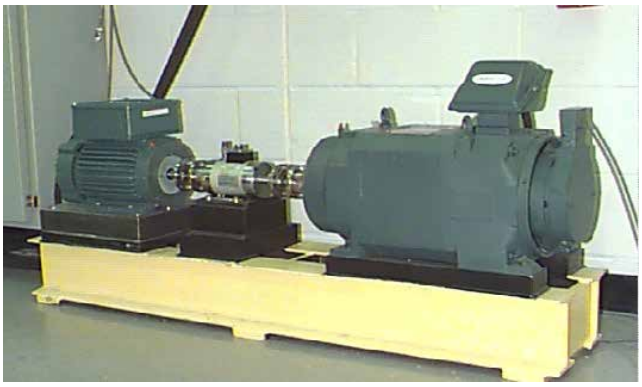


図 2.6　CWRU データセットにおける実験構成 [38]

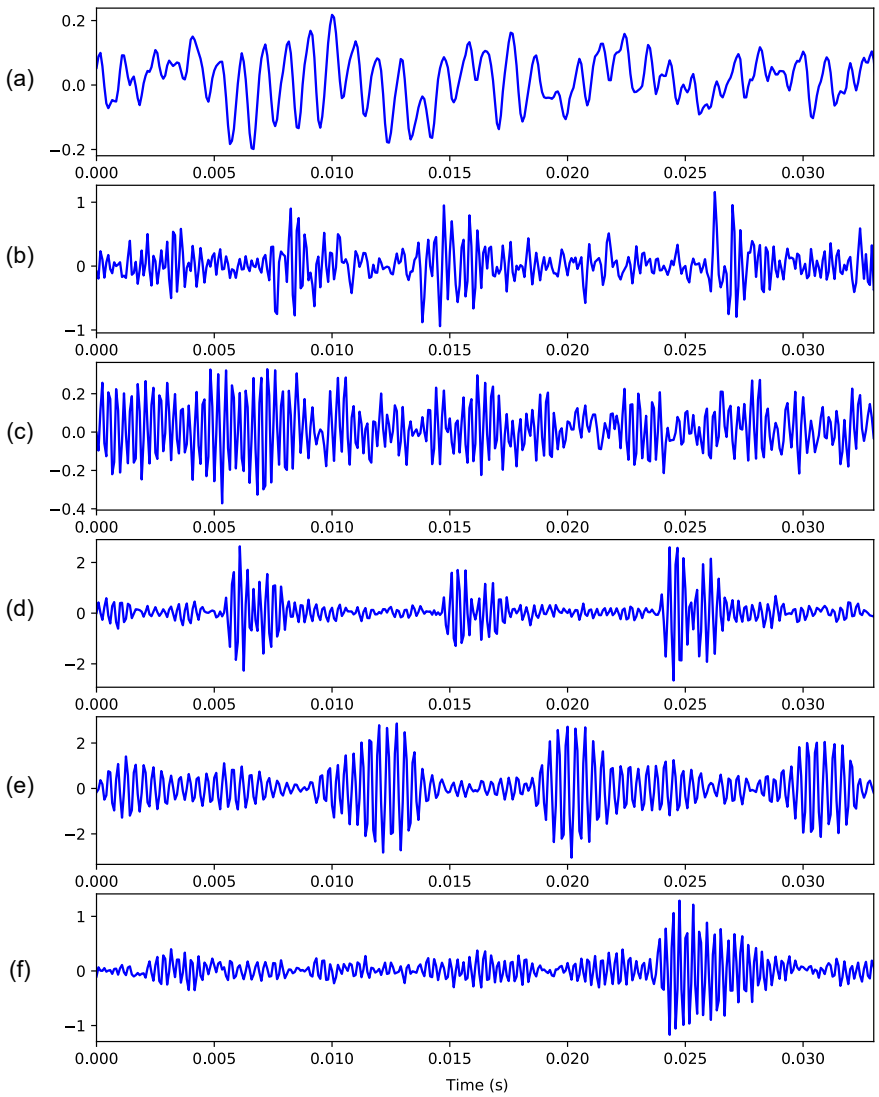


図 2.7　CWRU データセットに含まれる波形の例 [38]（a: N、b: IR、c: B、d: OR@6、e: OR@3、f: OR@12）

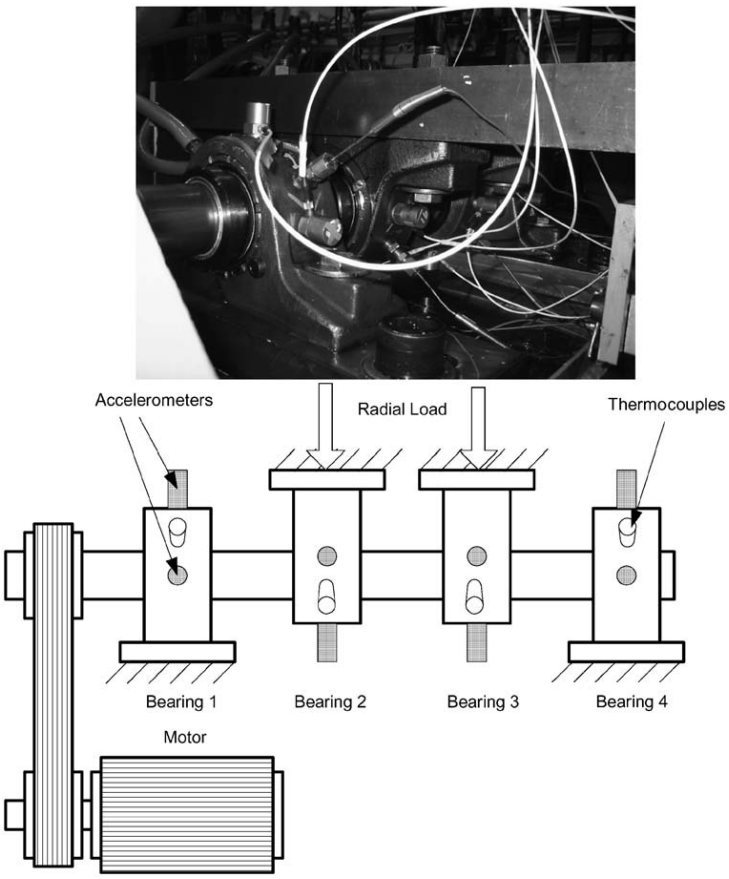


図 2.8　IMS データセットにおける実験構成 [39]

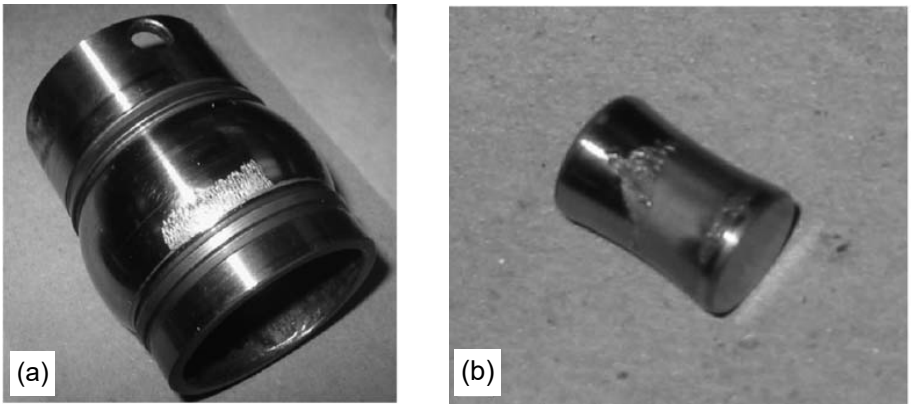


図 2.9　IMS データセットにおける軸受故障後の各部品の外観（a: 内輪、b: 転動体）[39]

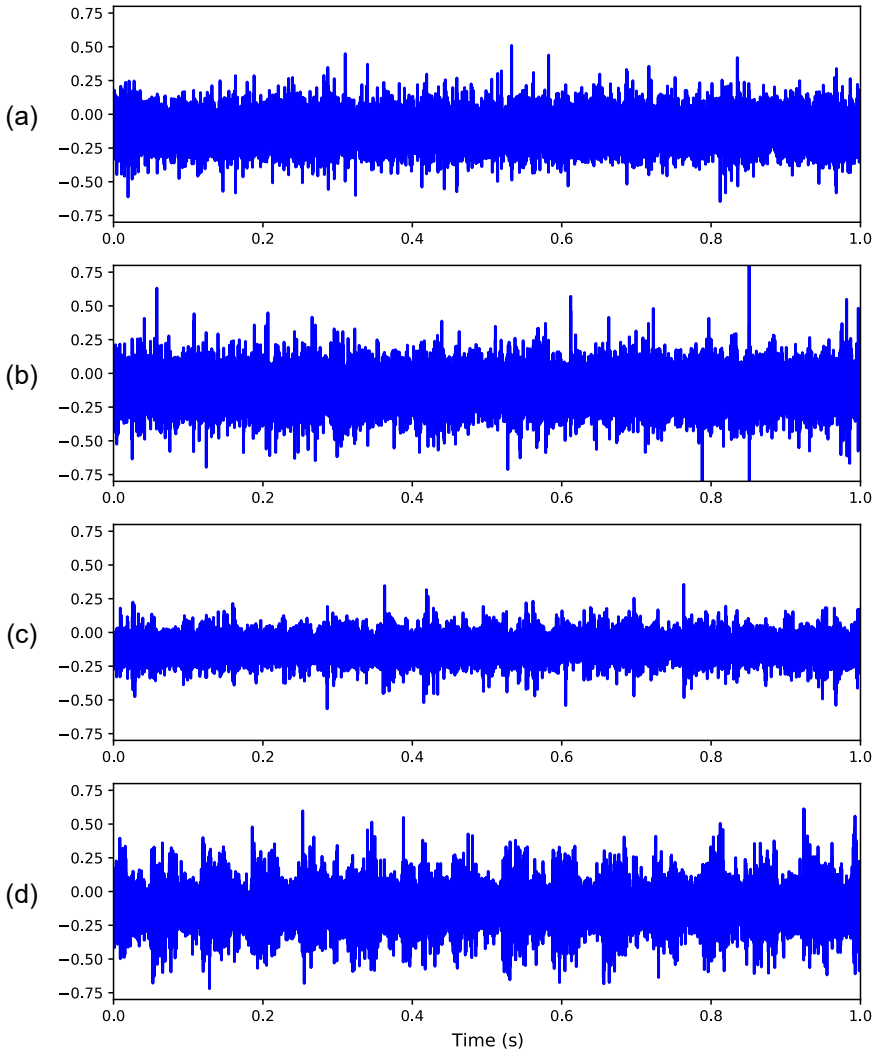


図 2.10 IMS データセットに含まれる (a, b) 正常時および (c, d) 故障時における波形

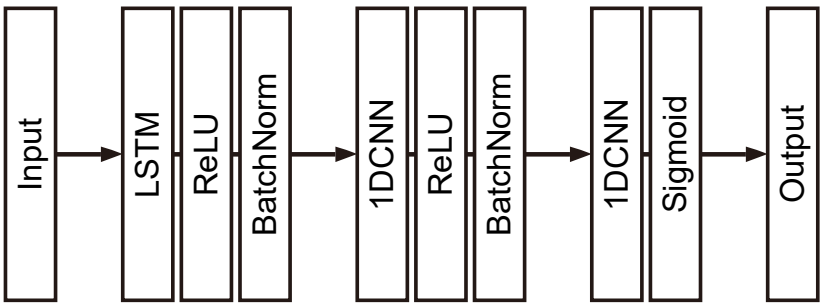


図 2.11　軸受振動データからの異常検知のための DNN モデルの構造

図 2.7 および図 2.10 に示すように軸受振動データにおける波形においてそれぞれの特徴的な箇所は数十時点にわたるため、先の実験で扱ったような畳み込み層のみで構成される DNN モデルでは異常の特徴を上手く捉えること困難である。そこで DNN モデルに図 2.11 のように、LSTM 層を含むネットワーク構造を採用した。LSTM 層は 128 チャネル、隠れ層には 128 チャネル、カーネル幅 9 を有する畳み込み層、最終層には 128 チャネル、カーネル幅 1 を有する畳み込み層を用いた。1 層目の LSTM 層および 2 層目の畳み込み層からの出力に対して ReLU 活性化およびバッチ正規化を適用し、最終層の出力に対してシグモイド活性化を適用した。ハイパパラメータはそれぞれ $p_1 = 10^{-1}$、$p_2 = 10^{-5}$ とし、式 (2.29) の損失関数を最小化するように学習率 10^{-3} で $5,000$ 回のパラメータ更新を行い、DNN モデルを学習した。

2.4 結果および考察

2.4.1 生成データからの異常検知

図 2.12 に実験に供した波形と提案手法により学習した DNN モデルの出力を示す。波形に外れ値が含まれる箇所に対してのみ高い値が推定され、その他の箇所については低い値が推定された。ここで、外れ値を含む箇所やその程度を示すアノテーションを用いずに DNN モデルが学習されており、本結果は提案する学習手法により、アノテーション作業を必要とせずに異常の検知が可能な DNN モデルの学習が可能であることを示している。また、表 2.1 はそれぞれの大きさの外れ値に対する検出精度である。正常波形に含まれるノイズの標準偏差 σ_f の 5 倍の大きさの外れ値に対して 98.0%、6 倍の大きさの外れ値に対して 99.5% と、良好な検出精度が確認された。さらに、波形に付与された外れ値の振幅と推定値の関係を図 2.13 に示す。外れ値の大きさに応じて推定値が大きくなる傾向が確認され、提案手法により学習された DNN モデルを用いることで異常の大きさの程度の定量可能性が示唆された。DNN モデルの学習に要した時間は約 80 分であり、評価に要した時間は 1 波形（1,000 時点）に対して 296 ミリ秒であった。

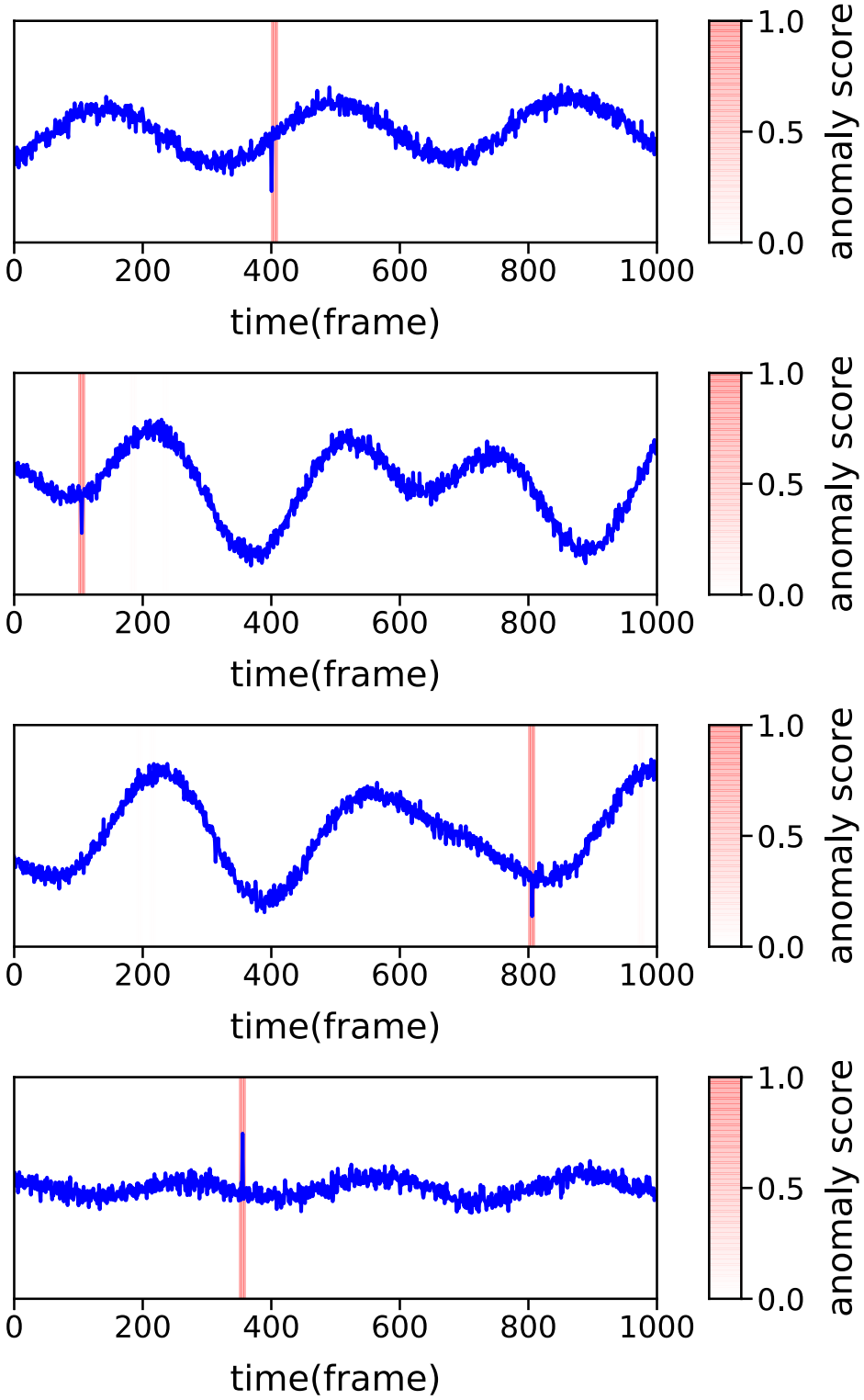


図 2.12　外れ値を含む波形データ（実線）および DNN モデルの推定値（網掛部に 10 時点近傍の最大値を描画）

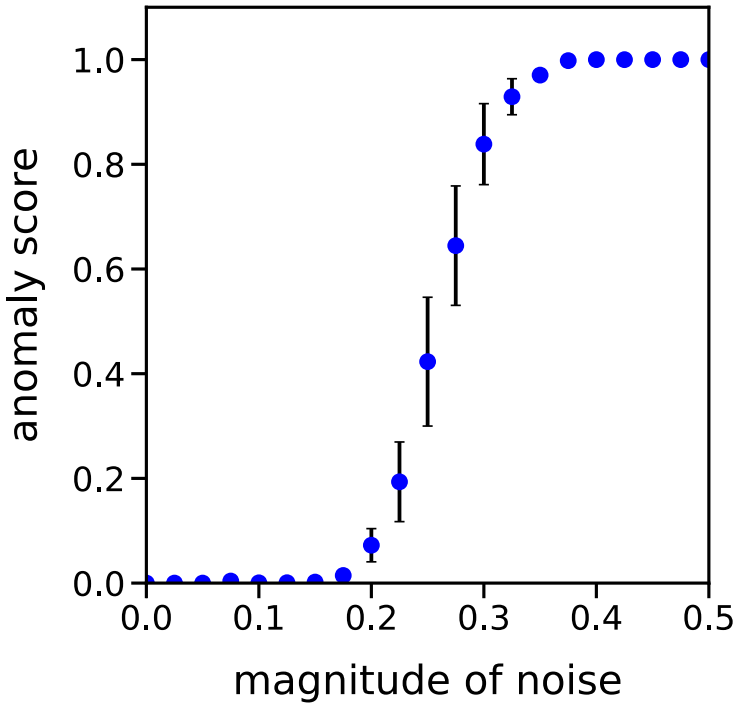


図 2.13　印加した外れ値の大きさと推定値の関係

表 2.1　外れ値の検出精度

Magnitude of noise	**Accuracy**
$4\sigma_f$	80.5%
$5\sigma_f$	98.0%
$6\sigma_f$	99.5%

2.4.2　振動データからの異常検知

図 2.14 から図 2.18 に、本手法を CWRU データセットに適用した結果を示す。それぞれの故障を含む軸受から得られた波形に含まれる特徴的な箇所に対して高い値が推定されることが確認された。一方でそれらを含まない箇所およびその他の波形に対しては、低い値を推定することが確認された。これらの結果は学習に供したデータのうち、正のデータ集合に含まれるそれぞれの波形に共通する特徴を自動的に抽出し、それらを検知可能となるように DNN モデルが学習されたことを示唆するものである。本データセットに含まれる波形には数十時点から成るそれぞれの故障に特徴的な波形を含んでいるが、DNN モデルに LSTM 層を導入したことで長期的な記憶が可能となり、それぞれの故障に独特な波形の特徴を適切に抽出され、それらの検知が行われていると考えられる。DNN モデルの学習に要した時間は約 40 分であり、評価に要した時間は 1 波形（200 時点）に対して 70 ミリ秒であった。

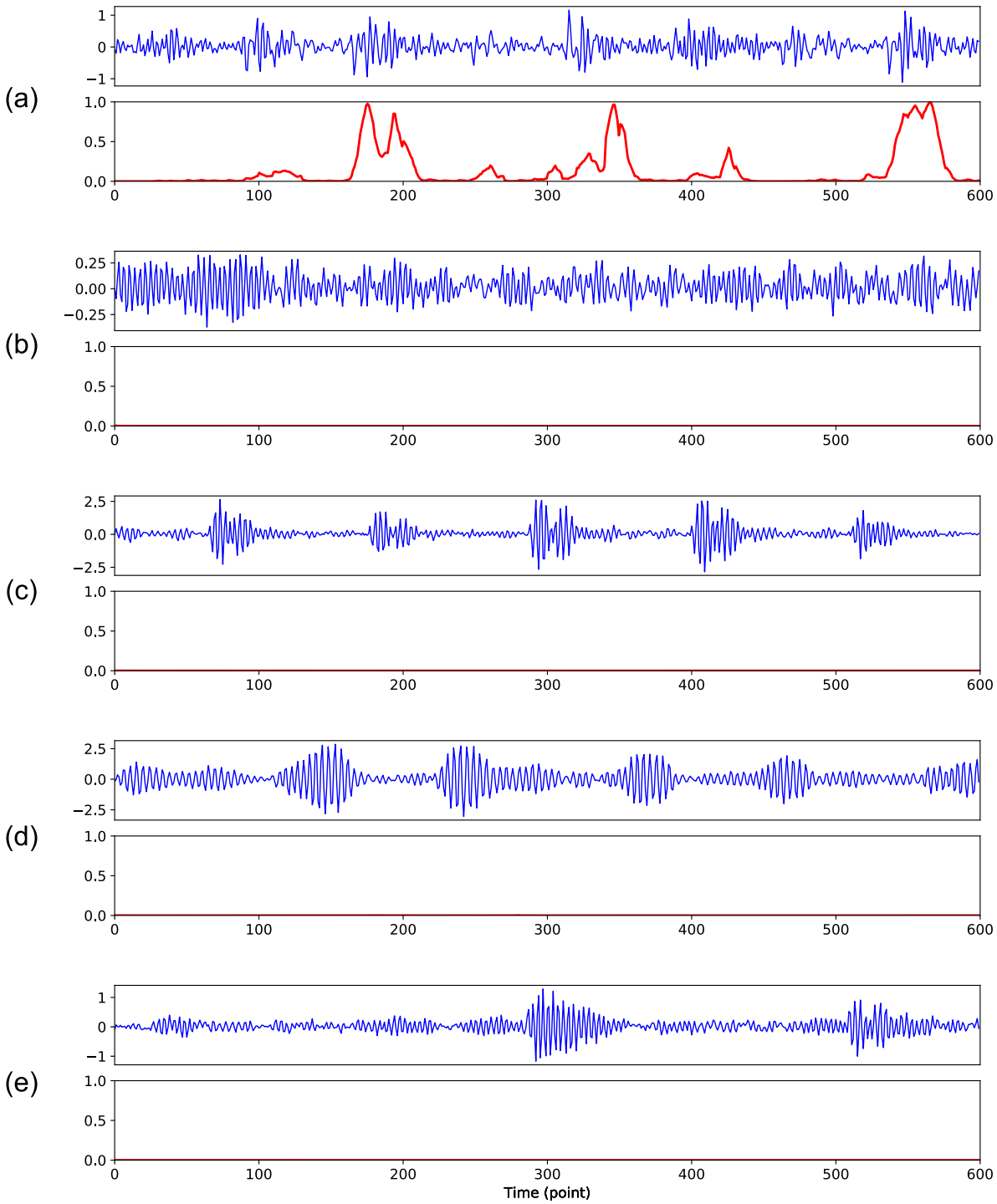


図 2.14 CWRU データセットに含まれる各振動データおよび IR を正のデータ、その他を負のデータとして学習された DNN モデルによる推定値（a: IR、b: B、c: OR@6、d: OR@3、e: OR@12）

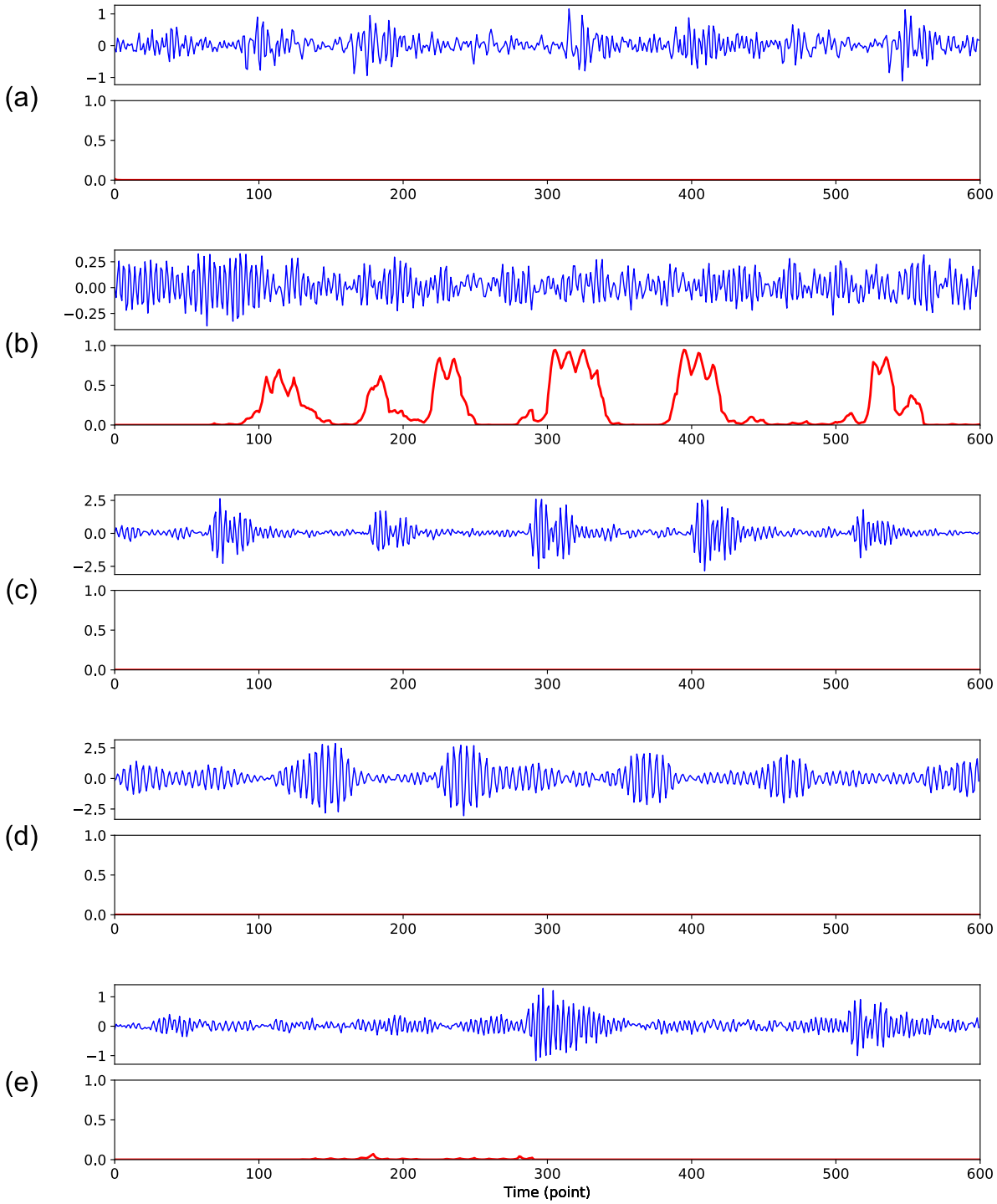


図 2.15　CWRU データセットに含まれる各振動データおよび B を正のデータ、その他を負のデータとして学習された DNN モデルによる推定値（a: IR、b: B、c: OR@6、d: OR@3、e: OR@12）

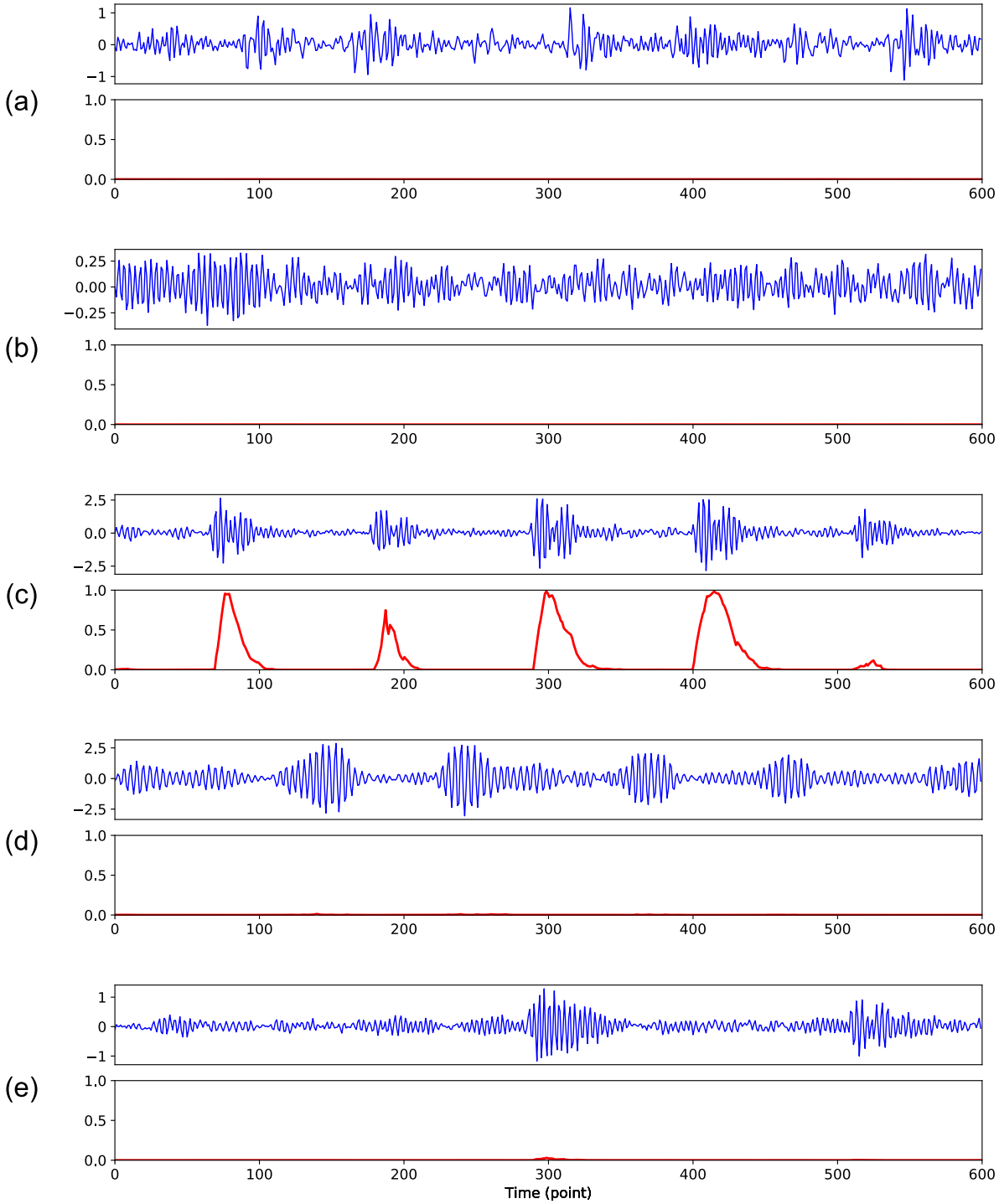


図 2.16 CWRU データセットに含まれる各振動データおよび OR@6 を正のデータ、その他を負のデータとして学習された DNN モデルによる推定値（a: IR、b: B、c: OR@6、d: OR@3、e: OR@12）

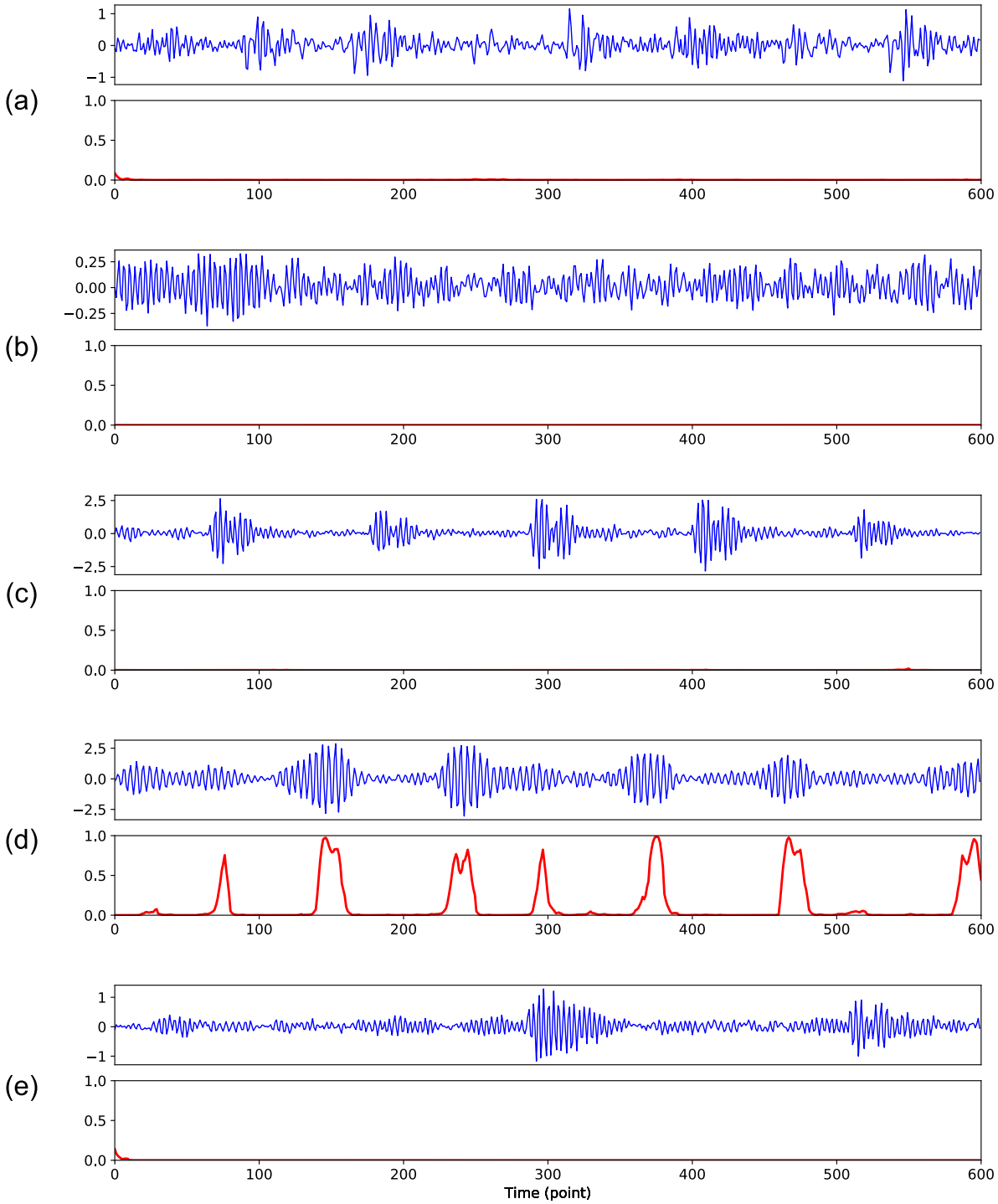


図 2.17　CWRU データセットに含まれる各振動データおよび OR@3 を正のデータ、その他を負のデータとして学習された DNN モデルによる推定値（a: IR、b: B、c: OR@6、d: OR@3、e: OR@12）

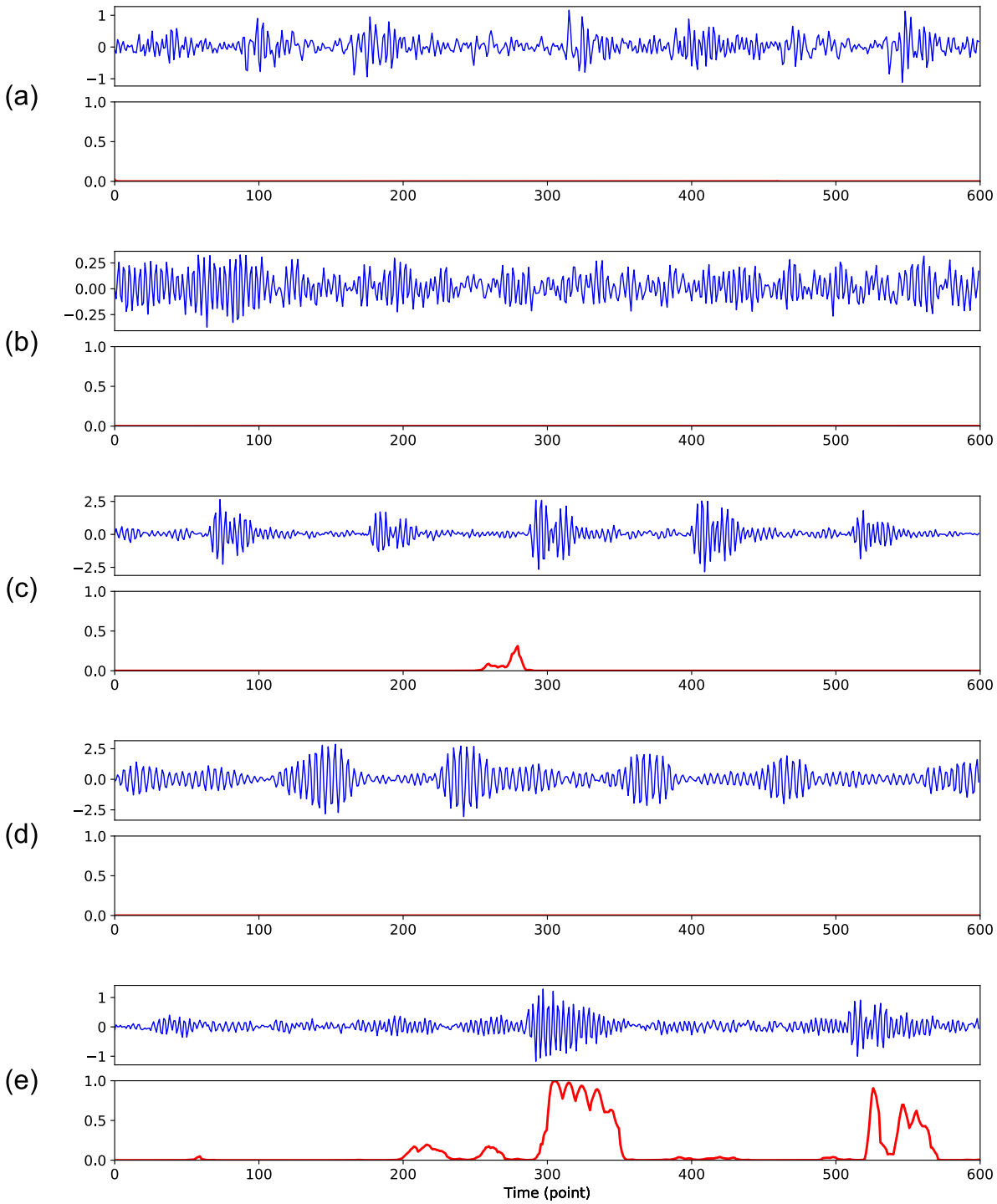


図 2.18 CWRU データセットに含まれる各振動データおよび OR@12 を正のデータ、その他を負のデータとして学習された DNN モデルによる推定値（a: IR、b: B、c: OR@6、d: OR@3、e: OR@12）

図 2.19 および図 2.20 に本手法を IMS データセットに適用した結果を示す。両データにおいて提案手法により学習された DNN モデルを用いることで、軸受故障発生後の波形から高い値が推定され、一方で故障の無い状態の軸受から得られた振動データに対しては低い値が推定された。この結果から、提案する学習手法により DNN モデルが異常を含む波形に特徴的な箇所の特徴量を適切に抽出できるように学習され、異常を含む箇所に対して限定的に高い値が推定されたと考えられる。DNN モデルの学習に要した時間は約 20 分であり、評価に要した時間は 1 波形（2,040 時点）に対して 630 ミリ秒であった。ここで DNN モデルの推定値から軸受の劣化具合を評価するための指標

$$H(t) = \begin{cases} H(t-1) & \text{estimated score} > 3\sigma, \\ H(t-1)+1 & \text{otherwise} \end{cases} \tag{2.31}$$

を導入した。σ は、故障の無い状態における軸受から得られた振動データから算出された標準偏差を示す。$H(t)$ が短時間で頻繁に増加した際に故障と判定し、本実験では連続する 3 時点で閾値を超えた値が検知された場合に異常と判定した。

軸受 3 に対する実験では、分散の値を比較する方法により評価した際には 1,988 時点で異常が検知された一方で、提案手法では 1,797 時点で検知されており、31 時間早期の異常検知が可能であった。同様に軸受 4 に対する実験では、分散の値を比較する方法により評価した際には 1,580 時点で異常が検知された一方で、提案手法では 1,438 時点で検知され、23 時間早期の異常検知が可能であった。本結果を、CWRU データセットを用いた実験結果を併せて考察すると、提案する学習手法により DNN モデルが故障した軸受の振動データに潜在する特徴を適切に抽出できるように学習され、故障を早い段階で検知できたと考えられる。

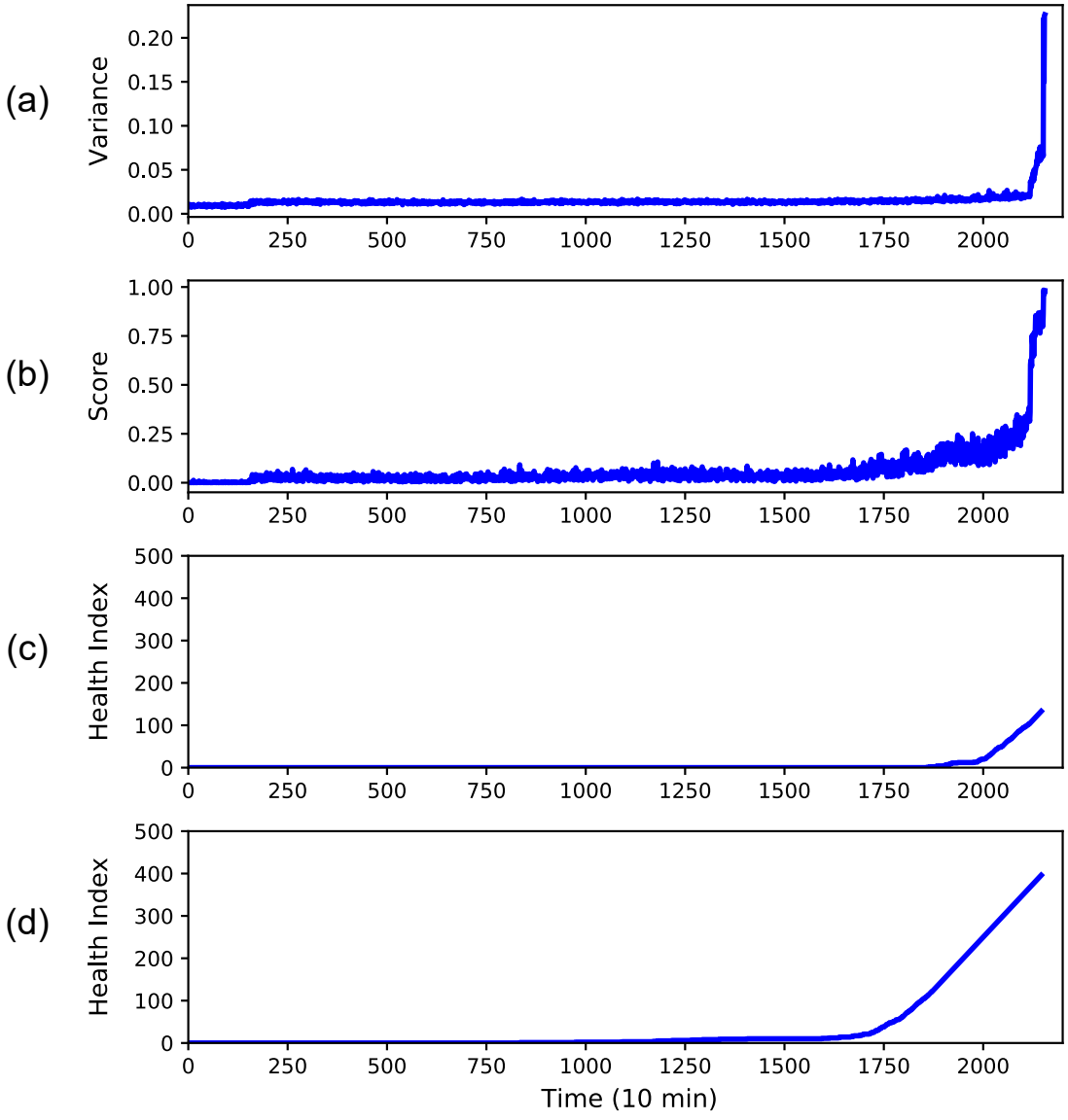


図 2.19 軸受 3 に対する異常検知結果（a: 分散の値を比較する方法により推定された異常度、b: 提案手法により推定された異常度、c: 分散の値を比較する方法により得られた $H(t)$、d: 提案手法により得られた $H(t)$）

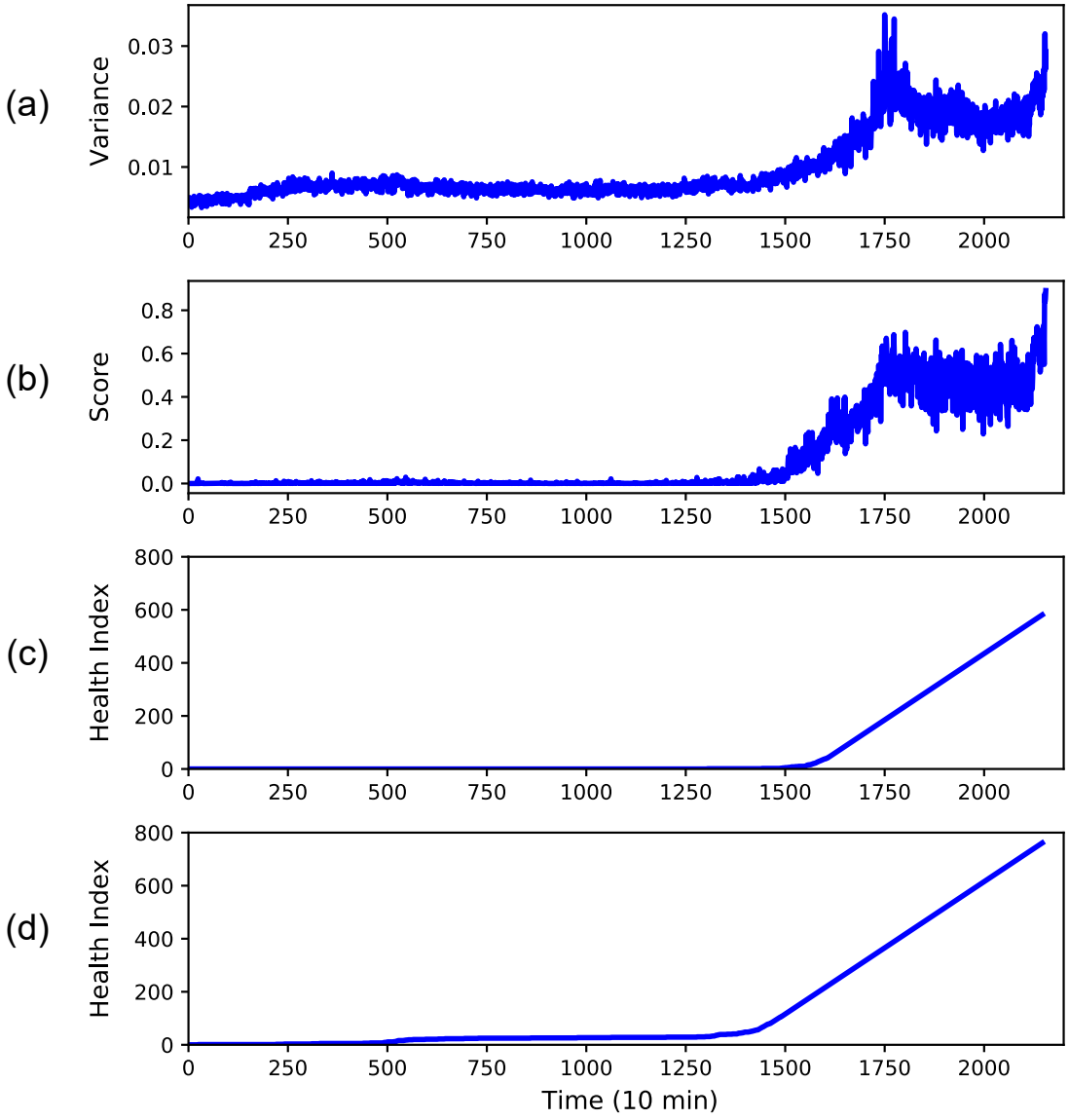


図 2.20　軸受 4 に対する異常検知結果（a: 分散の値を比較する方法により推定された異常度、b: 提案手法により推定された異常度、c: 分散の値を比較する方法により得られた $H(t)$、d: 提案手法により得られた $H(t)$）

2.5 結言

本章では時系列データから異常を検知するための DNN モデルおよびその学習手法に関する基礎的検討について述べた。特に時系列データに含まれる異常に関して、その有無が既知であるが具体的な箇所や程度が未知であるような弱教師ありデータを用いて DNN を学習する方法について述べた。学習には、マルチインスタンス学習に着想を得た方法により時系列データの各時点における異常度に関する真値を必要とせずに DNN モデルの学習を行った。実験では異常を含む弱教師あり時系列データに対して、データにおける異常を含む箇所のみに対して高い値を推定するように DNN モデルの学習が可能であることを確認した。また、外れ値を含む波形データからの異常検知に関する実験では、外れ値の大きさと DNN モデルの推定の推定との間に相関が確認され、異常の検知のみならず定量の可能性が示唆された。さらに、本手法の実システムへの応用可能性を確認するために、故障した軸受から取得された振動データの解析に適用したところ、それぞれの故障に特徴的な波形を含む箇所に対して高い推定値が確認され、本手法による故障診断のような実応用の可能性が見出された。次章では本手法を改良し、複数種類の異常を識別可能とするための DNN モデルおよびその学習手法と評価について述べる。

第 3 章

異常識別のための深層ニューラルネットワークおよびその学習手法

3.1 緒言

第 2 章では弱教師あり多変量時系列データからの異常検知を目的とした DNN モデルとその学習手法を確立した。特に DNN モデルの学習にマルチインスタンス学習に着想を得た方法を採用することで、異常を含むデータにおける異常を含む箇所のみに対して高い値を推定するように DNN モデルの学習を可能とし、異常検知への適用可能性が示唆された。本手法は従来手法と比べて異常を含む時系列データに対する詳細なアノテーションを付与することなく DNN モデルを学習可能である点で有利である。本手法を実システムに応用する上で、異常の有無のみならず異常の識別が可能となればより有用である。例えば故障した動的機器から取得される信号に対し、その故障に独特な信号の特徴を学習することができれば、故障原因の特定に役立つ可能性がある。そこで、本研究ではニューラルネットワークが回帰問題のみならず識別問題にも有効であることを活かし、先のニューラルネットワークを異常の識別に適用するための学習手法の改良を行う。本章では提案する DNN モデルを識別問題に適用するための方法について述べた後、提案手法により学習された DNN モデルの異常の識別能力に関する評価について述べる。

3.2 方法

3.2.1 識別のための DNN モデルおよびその学習手法

識別問題は多次元空間においてデータ $\mathbf{x}$ がどのクラスに属するかを判別する問題である。入力データ $\mathbf{x}$ が k 番目のクラスに属する確率を $P(y = y_k|\mathbf{x}; \mathbf{w})$ とするとき、多クラス識別の

ためのニューラルネットワークでは、出力値の総和が常に 1 になるように正規化を行うため、出力層の活性化関数にソフトマックス関数

$$z_k = \frac{\exp\left(u_k^{(L)}\right)}{\sum_{j=1}^{K} \exp\left(u_j^{(L)}\right)} \tag{3.1}$$

を用いることが多い。さらに、多クラス識別のための DNN モデルの学習では一般的に交差エントロピー損失

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk} \log f_k\left(\mathbf{x}_n; \mathbf{w}\right) \tag{3.2}$$

を最小化するように DNN モデルのパラメータを最適化する。ここで、N、K はそれぞれデータ数およびクラス数であり、$\mathbf{x}_n$ は n 番目のデータ、$y_{nk} \in \{0, 1\}$ は n 番目のデータが k 番目のクラスに属するかを示す教師信号、$f_k\left(\mathbf{x}_n; \mathbf{w}\right)$ は n 番目のデータが k 番目のクラスに属する期待値に関する DNN モデルの推定値を表す。現実的にはこのような最適化問題を解析的に解くことは困難であるため、第 2 章と同様に十分な入力データと真値の対 $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1,\ldots,N}$ を用意し、損失 $E(\mathbf{w})$ を最小化するように学習を行う。しかし、第 2 章で扱ったような弱教師あり時系列データには十分な長さを持つ区間において異常が含まれていることは既知であるが、異常の含まれる箇所やその程度については未知であることと同様に、各時点にどのような異常が含まれているかを示すアノテーションも付与されていない。第 2 章で確立した単一クラス認識のための DNN モデルを複数用意し、それぞれを並行して用いることで複数の異常識別が可能となるが、先に述べた DNN モデルを用い、その学習手法を改良することで複数種類の異常を識別を可能とすれば、異常検知の迅速化やリアルタイム化、メモリ消費の抑制、簡便化に貢献できる可能性がある。そこで、次節以降では学習手法の改良と提案手法により学習された DNN モデルを用いた異常識別について述べる。

3.2.2 識別のための DNN モデルの弱教師あり学習手法

先に述べた単一クラス認識のための DNN モデルの学習手法を識別問題へ応用するため、学習手法の改良を行う。具体的には、第 2 章で提案した式 (2.29) の損失関数を一般化し、

$$E(\mathbf{w}) = \max\left(0, \sum_{n}\sum_{k}\left(\phi_{nk} - \psi_{nk} z_{nk}\right)^2\right) + \lambda \tag{3.3}$$

とした。第 1 項はデータセットに含まれる $n \in \mathcal{N}$ 番目のデータについて $k \in \{1, \ldots, K\}$ 番目のクラスの異常を含んでいるかによりその損失の大きさを決定する。ここで、$\mathcal{N} \subset \{1, \ldots, N\}$ および $n(\mathcal{N}) = 2$ であり、ランダムに抽出された 1 組のデータ対のインデックスの集合を表す。z_{nk} は n 番目の時系列データに k 番目のクラスに属する異常が含まれる期待値について、

DNN モデルの推定値 $f_k\left(\mathbf{x}_n^{(t)};\mathbf{w}\right)$ における $t \in \{1,\ldots,T\}$ 時点の中で最も高い値を示し、

$$z_{nk} = \max_t f_k\left(\mathbf{x}_n^{(t)};\mathbf{w}\right) \tag{3.4}$$

で表される。また、ϕ_{nk} は

$$\phi_{nk} = \begin{cases} 1 & \text{if } k\text{th anomaly is included,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

と表され、n 番目のデータに k 番目のクラスに属する異常が含まれるかに応じて 1、0 のいずれかの値をとる。同様に ψ_{nk} もまた

$$\psi_{nk} = \begin{cases} 1 & \text{if } k\text{th anomaly is included,} \\ -1 & \text{otherwise} \end{cases} \tag{3.6}$$

と表され、n 番目のデータに k 番目のクラスに属する異常が含まれるかに応じて 1、-1 のいずれかの値をとる。さらに、式 (3.3) の第 2 項は正則化項であり、

$$\lambda = \lambda_1 + \lambda_2 + \lambda_3, \tag{3.7}$$

$$\lambda_1 = p_1 \sum_n \sum_k \sum_{t=1}^{T-1} \left(f_k\left(\mathbf{x}_n^{(t)};\mathbf{w}\right) - f_k\left(\mathbf{x}_n^{(t+1)};\mathbf{w}\right)\right)^2, \tag{3.8}$$

$$\lambda_2 = p_2 \sum_n \sum_k \sum_t f_k\left(\mathbf{x}_n^{(t)};\mathbf{w}\right), \tag{3.9}$$

$$\lambda_3 = -p_3 \sum_n \sum_k \phi_{nk} \log \frac{\exp(z_{nk})}{\sum_j \exp(z_{jk})} \tag{3.10}$$

と表される。ここで λ_1、λ_2 は第 2 章と同様にそれぞれ平滑化項、およびスパース化項であり、第 3 項は交差エントロピー損失である。また、p_1、p_2 および p_3 は各損失の大きさを制御するハイパパラメータである。

3.3 実験

提案手法により学習された DNN モデルによる故障検知の実現可能性を評価するため、公開データセットを用いた実験を行った。まず、提案手法により学習された DNN モデルによる軸受故障の識別能力を評価した。さらに、式 (3.3) の損失関数におけるハイパパラメータを調整しながら DNN モデルの学習を行い、故障の識別能力の変化について評価した。振動データには第 2 章で扱った CWRU データセットを用い、データの分割方法等についても第 2 章と同様の条件で実験に供した。振動データはそれぞれの故障に特徴的な波形を含むが、先の実験と同

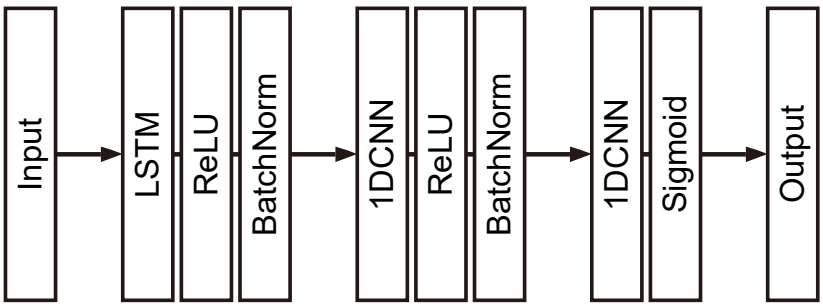


図 3.1 軸受振動データからの故障識別のための DNN モデルの構造

様にそれらの含まれる箇所や程度に関する定量的な情報を用いずに DNN モデルの学習を行った。実験に用いた DNN モデルの構造を図 3.1 に示す。第 2 章で用いた DNN モデルと概ね同様の構造であるが、出力層の形状を複数チャネル出力（本実験では 6 チャネル出力）とし、検知すべき故障の数だけ値を出力する構造とした。1 層目の LSTM 層および 2 層目の畳み込み層からの出力に対して ReLU 活性化およびバッチ正規化を適用し、最終層の出力に対してシグモイド活性化を適用した。式 (3.3) の損失関数を最小化するように学習率 10^{-3} で 5,000 回のパラメータ更新を行い、DNN モデルを学習した。さらに、認識精度を向上させるための最適なハイパパラメータ p_1、p_2、p_3 の検討として、p_1 を 10^{-1} から 10^{-3}、p_2 を 10^{-4} から 10^{-6}、p_3 を 10^{-1} から 10^{-3} の範囲で変動させ、識別精度にもたらす影響を確認した。

3.4 結果および考察

図 3.2 から図 3.4 は、実験に供した各振動データおよび提案手法により学習された DNN モデルの推定値である。それぞれの、ハイパパラメータ p_1、p_2、p_3 の値の組み合わせが認識精度の善し悪しに影響するが、適切にハイパパラメータが選ばれた場合には故障に独特な波形を含む箇所に対して適切に高い値が推定され、一方でそれらを含まない箇所に対しては低い値が推定されていることが確認できる。この結果は提案する DNN モデルおよびその学習手法によって、データに潜在する特徴を自動的に抽出し、異常の識別ができることを示している。また、平滑化項に対する重みパラメータ p_1 が増減することで、各時点間の値の変動の大きさ、スパース化項に対する重みパラメータ p_2 が増減することで、全体的な推定値の大きさが変動している様子が見受けられ、正則化項が適切に機能していることが確認できる。ここで、提案手法の故障識別能力を評価するために、t 時点目における、k 番目のクラスに属する異常を含む期待値 $f_k\left(\mathbf{x}^{(t)};\mathbf{w}\right)$ に対して

$$\text{detected anomaly} = \arg\max_k \sum_t f_k\left(\mathbf{x}^{(t)};\mathbf{w}\right) \tag{3.11}$$

とすることで、波形がどのクラスの異常を含むか判定し、その識別精度を先行研究と比較した。図 3.5 および表 3.1 は、式 (3.8)、式 (3.9)、式 (3.10) におけるそれぞれのハイパパラメータ p_1、p_2、p_3 の値を変えた場合の DNN モデルの識別精度を比較したものである。実験では、$p_1 = 10^{-1}$、$p_2 = 10^{-5}$、$p_3 = 10^{-1}$ のときに、最も良い認識精度が得られた。

また、表 3.2 は、DNN モデルを用いた他の手法との識別精度の比較である。提案手法により学習された DNN モデルは他の DNN モデルを用いた手法と比べて良好な精度で異常を識別可能であることが確認された。

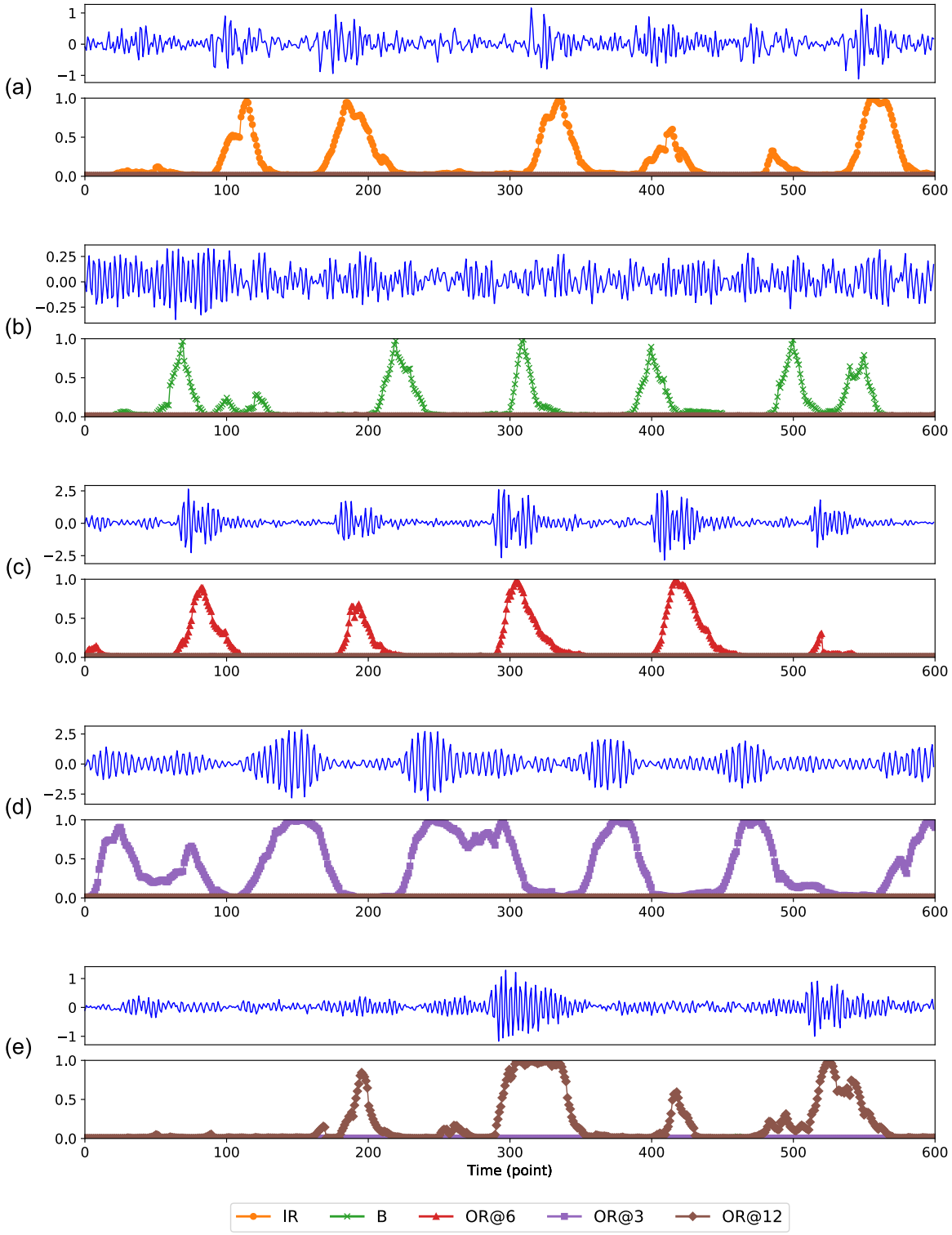


図 3.2　CWRU データセットに含まれる各振動データおよび DNN モデルの推定値（$p_1 = 10^{-1}$、$p_2 = 10^{-5}$、$p_3 = 10^{-1}$）

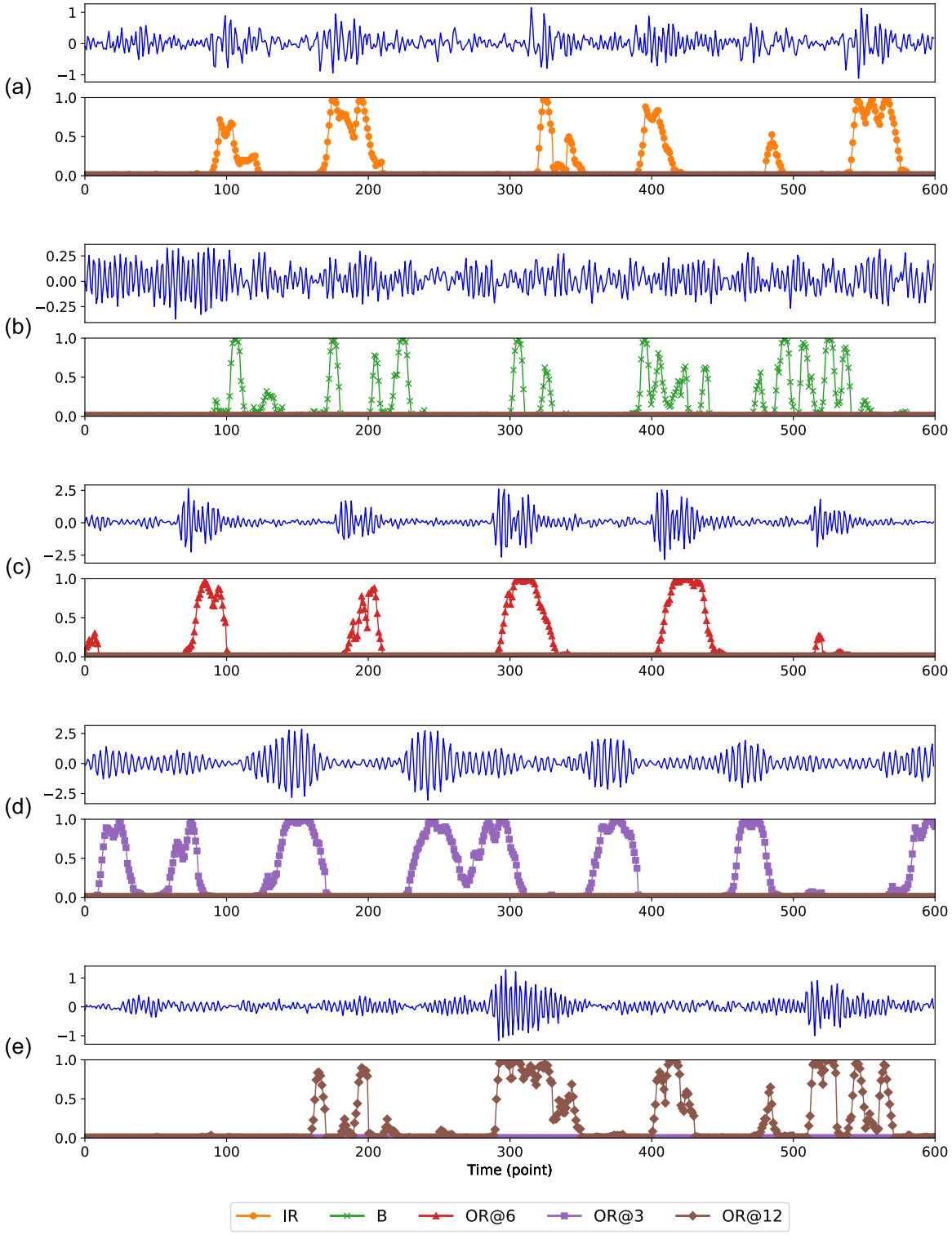


図 3.3 CWRU データセットに含まれる各振動データおよび DNN モデルの推定値（$p_1 = 10^{-2}$、$p_2 = 10^{-5}$、$p_3 = 10^{-2}$）

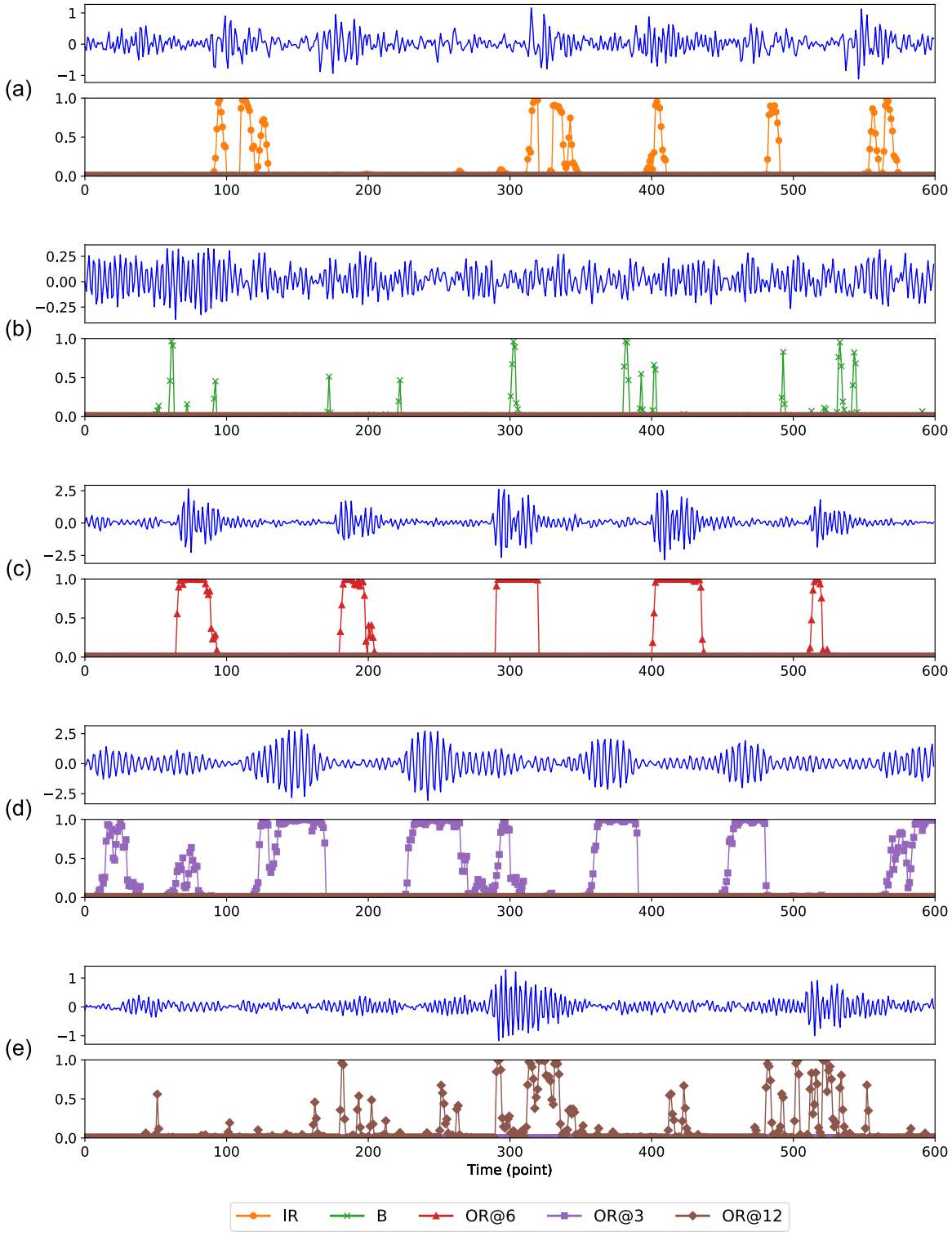


図 3.4　CWRU データセットに含まれる各振動データおよび DNN モデルの推定値（$p_1 = 10^{-3}$、$p_2 = 10^{-6}$、$p_3 = 10^{-1}$）

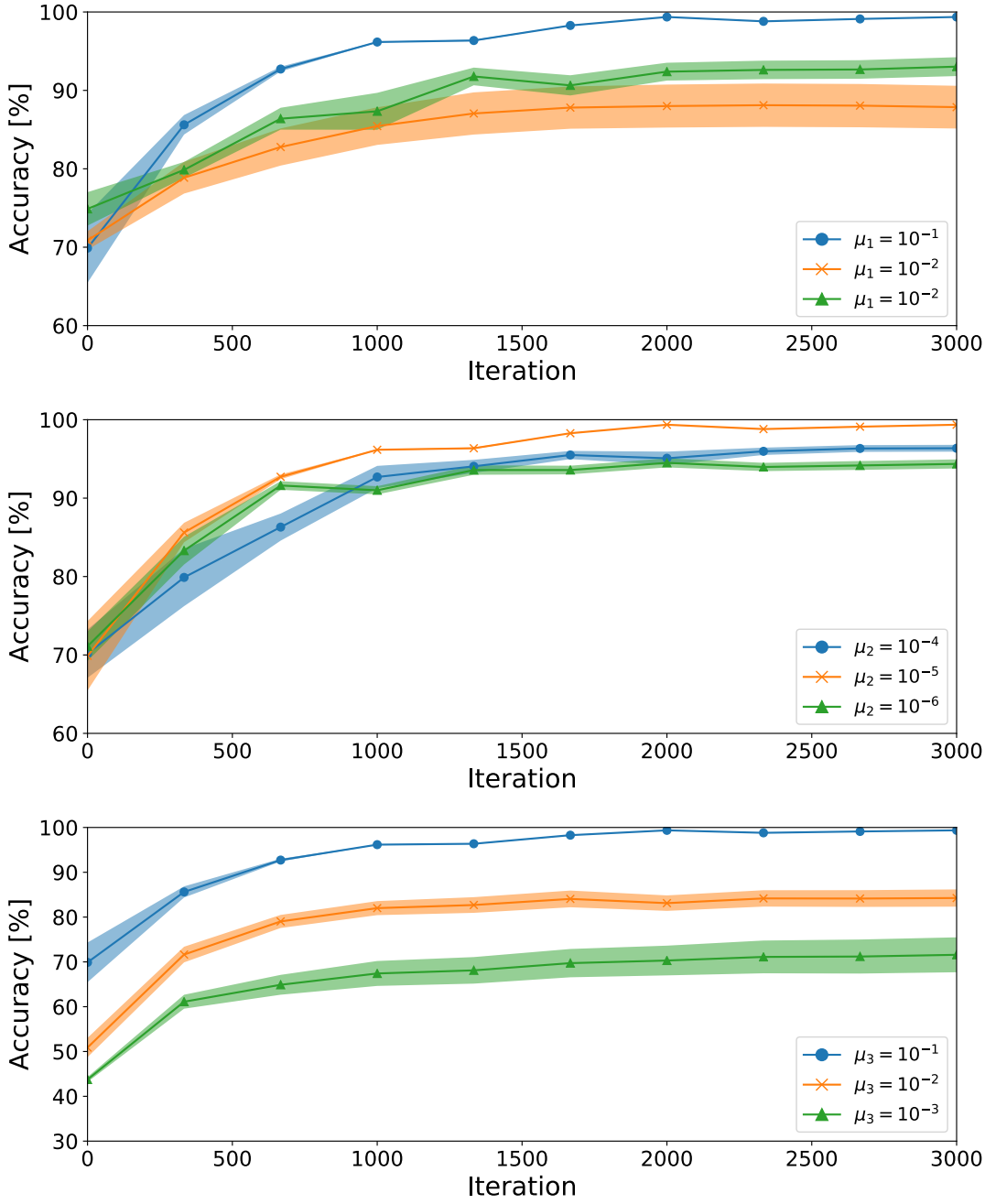


図 3.5　正則化項の各ハイパパラメータごとの DNN モデルのパラメータの更新回数と軸受故障の識別精度の関係

表 3.1 正則化項のハイパパラメータ p_1、p_2 および p_3 を変化させた際の軸受故障の識別精度の比較

Methods			Accuracy
$p_1 = 10^{-1}$	$p_2 = 10^{-5}$	$p_3 = 10^{-1}$	
10^{-1}	10^{-5}	10^{-1}	99.4%
10^{-2}	10^{-5}	10^{-1}	87.4%
10^{-3}	10^{-5}	10^{-1}	93.0%
10^{-1}	10^{-4}	10^{-1}	96.4%
10^{-1}	10^{-6}	10^{-1}	94.4%
10^{-1}	10^{-5}	10^{-2}	84.2%
10^{-1}	10^{-5}	10^{-3}	71.6%

表 3.2 識別精度に関する他手法との比較

Methods	Accuracy
Compact 1DCNN [40]	93.3%
DNN with temporal coherence (16.67 ms) [41]	97.4%
Ours	99.4%

3.5 結言

本章では前章で提案した DNN モデルおよびその学習手法について、異常の検知のみならず異常の識別を可能とするための改良について述べた。まず、第 2 章で確立した DNN モデルの構造を改良した後、学習のための損失関数の一般化を行った。実験では、軸受振動データのそれぞれの異常を含む箇所に特徴的な波形に対して高い推定値が確認されたほか、適切に異常が識別され、本手法により学習された DNN モデルの軸受故障診断のような実システムへの応用可能性が見出された。さらに、ハイパパラメータの探索では、適切に選択されたハイパパラメータを用いることでより優れた識別精度が得られることを確認し、他の手法と比較して良好な識別精度が得られることを確認した。次章では確立した時系列データ解析のための DNN モデルを人物動作解析へ適用するための映像データからの人物動作特徴量の抽出手法について述べる。

第 4 章

人物動作特徴抽出のための深層ニューラルネットワークおよびその学習手法

4.1　緒言

原子力関連施設において監視カメラは核セキュリティのみならず作業者が安全に業務を遂行する上でも欠かせない設備である。監視カメラから日々収集される膨大な映像から異常を検知できれば、原子力関連施設の安全対策がより強固となる可能性がある。映像データは画像データに関する時系列データであり、第 2 章および第 3 章で扱った時系列データと比較して遥かに高次元なデータである。そのため、適切に解析を行うためにはデータから画像中の人物の動作を表す特徴量を事前に抽出し、解析に供することが望ましい。そのような特徴量の抽出に有効な手法のひとつに、画像中の人物の関節位置座標に関する情報、つまり人物の姿勢情報を取得できるモーションキャプチャ技術が挙げられる。画像中の人物の姿勢情報を取得することで効率よくデータの次元削減が可能であり、第 2 章および第 3 章で確立した時系列データ解析手法が適用できる可能性がある。市場に広く普及しているモーションキャプチャ機器では、ステレオカメラや赤外線カメラ等によって撮像された距離画像を解析する方法が一般的に用いられているが、特殊な撮像機器を必要とすることや、画角が狭く限られることが課題であった。特に画角が狭く限られることは、広範囲に存在する人物や近距離に存在する人物の姿勢を認識することを困難とし、映像監視へ適用する上で不利であった。一方で、近年盛んに研究されている DNN モデルによって RGB 画像を解析し、人物姿勢を認識する技術は特殊な撮像機器を必要としない。これらを改良し、撮像系に広角カメラを適用することで、広範囲に存在する人物や近距離に存在する人物の動作を認識することができれば、映像監視への応用がより現実的になると考えられる。

そこで本研究では DNN モデルを用いた広角画像からの人物姿勢認識手法の開発に取り組む。本章では、まず画像中の人物の姿勢推定および画像の補正パラメータ推定のための DNN モデルおよびその学習手法について述べる。次に、DNN モデルにより得られた 2 次元的な人物姿勢情報を用いて 3 次元的な人物姿勢情報を再構築する方法について述べる。最後に、人物姿勢認識および人物動作特徴量の抽出に関する評価について述べる。

4.2　方法

映像中の人物の姿勢認識技術は映像監視のみならず、ヒューマンコンピュータインタラクション、医療等の様々な分野で広く活用が期待されている技術であり、特に近年の深層学習技術の発展に伴い様々な手法 [22, 42, 43, 44, 45, 46, 47] が報告されている。Toshev らは AlexNet[48] に着想を得た CNN モデルを用い、 2 次元画像中の人物の関節位置座標を直接推定する方法を提案した [22]。本手法は画像中に複数の人物が含まれる場合や、オクルージョン等によって関節位置の真値が未知である場合に学習が上手く行われないことが課題であった。Tompson らは画像中の関節位置が存在する期待値に関する複数の 2 次元的な確信度マップを CNN モデルの出力として推定する方法を提案した [42]。確信度マップを用いた人物姿勢の表現方法は、関節位置座標を直接推定する方法に比べて高い認識精度が得られるだけでなく、複数の人物の姿勢推定に応用する際や画像内の関節位置に関する複数の候補を推定する際にその不確実さを表現できる点で有利である。画像中に存在する複数の人物を効率よく認識する方法として、Cao らは RGB 画像から人物の関節位置を推定するため、 2 次元的な確信度マップおよび、各関節の結合の度合いを表現する Part Affinty Fields（PAFs）を同時に推定する方法を提案している [47, 23]。

さらに、これらの手法を応用することで画像から 3 次元的な人物姿勢を推定する方法が報告されている。これらは、画像から直接 3 次元的な人物姿勢を推定する方法 [49, 50, 51, 52] と、予め推定された 2 次元的な人物姿勢から 3 次元的な人物姿勢を復元する方法 [53, 54, 55, 56] に大別される。前者は、入力から出力まで一貫した CNN モデルを用いることで、画像中の人物の 3 次元的な関節位置座標を直接推定する。このような直接的方法では、検出器の学習の際に人物の含まれる画像データおよびその人物の 3 次元的な関節位置座標に関するデータの対から成るデータセットが必要となる。つまり、これらの方法を広角画像に適用する際には、改めて人物の含まれる広角画像および画像中の人物の関節位置情報の対から成るデータセットを用意することが必要であるが、データの収集およびアノテーション作業には膨大な手間を要する。Xu らは、 3 次元コンピュータグラフィックス（Three-Dimensional Computer Graphics, 3DCG）を用いて生成されたデータセットを作成し、DNN モデルの学習に供することで、魚眼カメラで撮像された一人称視点画像のような収集が困難な画像から直接 3 次元的な人物姿勢を推定する方法を提案している [57]。しかし、Xu らの方法では画像中の人物の位

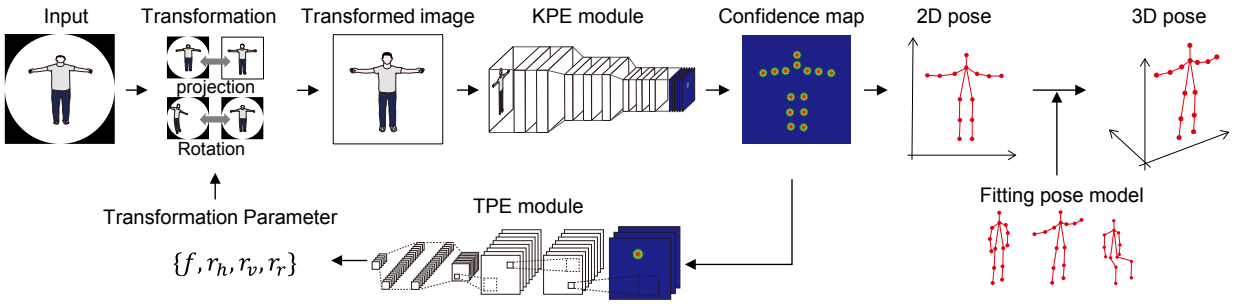


図 4.1　広角画像からの人物姿勢推定手法の概要

置が不変であり、人物の位置変化や姿勢変化に伴って生じる画像の歪みについては検討されていない。人物の 3 次元的な姿勢を間接的に推定する方法として、Iqbal らは 2 次元的な関節位置が既知である画像から成るデータセットと、それとは独立した 3 次元的なモーションキャプチャデータセットを用いる方法を提案した [58]。本先行研究では、DNN モデルにて推定された 2 次元的な姿勢と、3 次元的なモーションキャプチャデータを 2 次元平面に投影したものの誤差を最小化するように、3 次元的なモーションキャプチャデータを探索および変形させることで、画像中の人物の 3 次元的な姿勢を推定する。3 次元的な姿勢情報は間接的に推定されるため、新たなデータセットを構築する必要が無いことは、公開データセットの乏しい広角画像に適用するうえで有利である。

4.2.1　人物姿勢推定のための DNN モデルおよびその学習手法

本研究では、図 4.1 に示すような広角画像の歪みに対して頑健な姿勢認識手法を提案する。2 次元的な人物姿勢推定は、画像中の人物の関節位置を推定する注目点検出器（Keypoint Estimator, KPE）と画像補正パラメータを推定する画像補正量推定器（Transformation Parameter Estimator, TPE）を組み合わせた DNN モデルにより実現され、広角カメラから撮像された画像は最初のフレーム以降の入力画像が、適切な補正パラメータにより補正される。次に、注目点検出器によって生成された人物の関節位置を示す確信度マップが推定される。その後、画像補正量推定器は注目点検出器から取得した 2 次元的な関節位置を示す確信度マップを用いて、次フレームの姿勢推定精度を向上させるための適切な画像補正パラメータを推定する。続いて、推定された 2 次元的な人物姿勢情報に基づいて 3 次元的な人物姿勢が推定される。

2 次元的な人物姿勢認識のための注目点検出器として Cao らの研究 [47] を元に CNN モデルを構築した。本研究で採用する CNN モデルにおいても PAF の推定を行うが、画像中に複数の人物が存在する際に誤検出を避けるために用いた。提案手法では、まず CNN モデルの入力として高さ H、幅 W の画像を入力することで 2 次元的な人物姿勢を推定する。CNN モデ

ルでは畳み込み、プーリングおよび活性化を繰り返すことにより、画像中の人物の関節位置の存在する期待値を示す高さ h、幅 w の確信度マップ $\mathbf{h}(\mathbf{p}) \in \mathbb{R}^{h \times w}$ を推定する。具体的には、関節 $j \in \{1, 2, \ldots, J\}$ の位置に関する真値についての確信度マップ $\mathbf{h}_j^*(\mathbf{p})$ は

$$\mathbf{h}_j^*(\mathbf{p}) = \exp\left(-\frac{\left\|\mathbf{p} - \mathbf{y}_j^*\right\|^2}{\sigma^2}\right) \tag{4.1}$$

と表され、ここで $\mathbf{y}_j^* \in \mathbb{R}^2$ は画像内の関節位置座標の真値であり、$\mathbf{p} \in \mathbb{R}^2$ は確信度マップ内の注目座標である。この確信度マップを高さ H、幅 W に変形した確信度マップ $\mathbf{H}_j(\mathbf{P}) \in \mathbb{R}^{H \times W}$ における画像中の 2 次元的な人物の関節位置 $\mathbf{y}_j$ は

$$\mathbf{y}_j = \arg\max_{\mathbf{y}} \mathbf{H}_j(\mathbf{P}) \tag{4.2}$$

として推定される。注目点検出器の学習では上記の確信度マップの真値と推定値から算出される損失関数

$$E_{KPE} = \left\|\mathbf{h}_j(\mathbf{p}) - \mathbf{h}_j^*(\mathbf{p})\right\|^2 \tag{4.3}$$

を最小化するように CNN モデルのパラメータを最適化する。以上の方法を用いることで画像中の 2 次元的な人物姿勢を推定可能であるが、これらを可能とする DNN モデルは膨大な画像と真値から成るデータセットを用いた学習により実現される。このようなデータセットの構築作業の手間を省くため、人物が含まれる画像およびそれぞれの人物の関節位置座標の真値の対から成る公開データセットが提供されている [59, 60]。本研究ではこれらを活用し、 2 次元的な人物姿勢推定のための DNN モデルの学習を行う。一方で本研究で扱うような広角画像に対してこのようなデータセットは十分に提供されていないため、次節では広角画像の歪みに対して頑健とするための画像補正および、補正パラメータ推定方法について述べる。

4.2.2　画像補正量推定のための DNN モデルおよびその学習手法

本研究では、広角画像として 180° の画角を有する光学系によって撮像された画像を用いる。広角画像に対して解析を行う際には、一般に広角画像に特有の歪みに頑健とする工夫が必要となる。具体的には広角画像を補正せずに用いる方法と、画像を補正し歪みの無い画像に変換してから用いる方法がある。広角画像を補正せずに用いる方法では、認識対象人物の位置によりその映り方が異なるため、様々な歪みを想定した複数の検出器を用意することが必要である。しかし、DNN モデルを用いた方法においては検出器の学習に長時間を要するため、複数の検出器を用意することは開発の上で大きな負担となるほか、実装においてもメモリ消費や処理の高速化等の観点で不利となる。そこで、本研究では画像を補正し歪みの無い画像に変換してから用いる方法を採用する。

3 次元空間上における点 (X, Y, Z) の 2 次元平面上への投影 (u, v) は投影行列 $\mathbf{\Pi}$ を用いることで

$$(u, v) = \mathbf{\Pi}(X, Y, Z) \tag{4.4}$$

と表すことができる。さらに画像中の (u, v) は、画像の中心を原点とすることで動径 d および偏角 ϕ を用いた極座標形式でも表すことができる。ここで、広角画像の撮像系に採用されている等距離投影（Equidistant Projection, EP）および歪みを補正した透視投影（Perspective Projection, PP）における動径 d は、光学系に侵入する光線の入射角 θ を用いることでそれぞれ

$$d_{EP}(\theta) = f_{EP}\theta, \tag{4.5}$$

$$d_{PP}(\theta) = f_{PP}\tan\theta \tag{4.6}$$

で表される。式 (4.5) および式 (4.6) はそれぞれ、f_{EP}、f_{PP} を焦点距離とするとき光軸から角度 θ で入射する光が撮像面の中心から距離 d_{EP} および d_{PP} の位置に投影されることを意味する。つまり、各焦点距離 f_{EP}、f_{PP} が既知であれば画像の中心からの距離 d を用いて入射角 θ を導出することが可能である。ここで、式 (4.5) より、EP 形式では、半径 πf_{EP} だけの撮像面を用意すれば 180° の画角を得ることができる。一方で、式 (4.6) の PP 形式では、180° の画角を得るために無限の大きさの撮像面が必要となるため、EP 形式からの変換の際には図 4.2 に示すように、画角を絞ることや必要に応じて画像の回転が必要となる。そこで、本研究では画像の歪みを補正するために図 4.3 に示す EP-PP 変換器を提案した。本変換器は、広角画像に関する焦点距離 f_{PP} と 3 つの回転パラメータ r_h、r_v、r_r を制御することで、広角画像の歪みを補正する。

上記の EP-PP 変換器を適切に制御するための補正パラメータを推定するため、画像補正量推定器を導入する。画像補正量推定器は 3 層の畳み込み層と 2 層の全結合層からなる DNN モデルとし、注目点検出器から推定された確信度マップを入力として、広角画像の補正パラメータ $\{f_{PP}, r_h, r_v, r_r\}$ を推定する。図 4.4 に画像補正量推定器の学習手法を示す。画像補正量推定器の学習ではまず 3DCG を用いて EP 形式の広角画像を生成する。生成された広角画像は、予め指定した定義域内の補正パラメータ $\mathbf{T}^* = \{f_{PP}^*, r_h^*, r_v^*, r_r^*\}$ によって補正される。次に補正された画像は注目点検出器に入力され、人物姿勢を示す確信度マップが推定される。さらに確信度マップは画像補正量推定器に入力され補正パラメータ $\mathbf{T} = \{f_{PP}, r_h, r_v, r_r\}$ を推定する。最後に、推定された補正パラメータ $\mathbf{T}$ と真値 $\mathbf{T}^*$ に関する損失関数

$$E_{TPE} = \|\mathbf{T} - \mathbf{T}^*\|^2 \tag{4.7}$$

を最小化するように画像補正量推定器のパラメータを更新する。

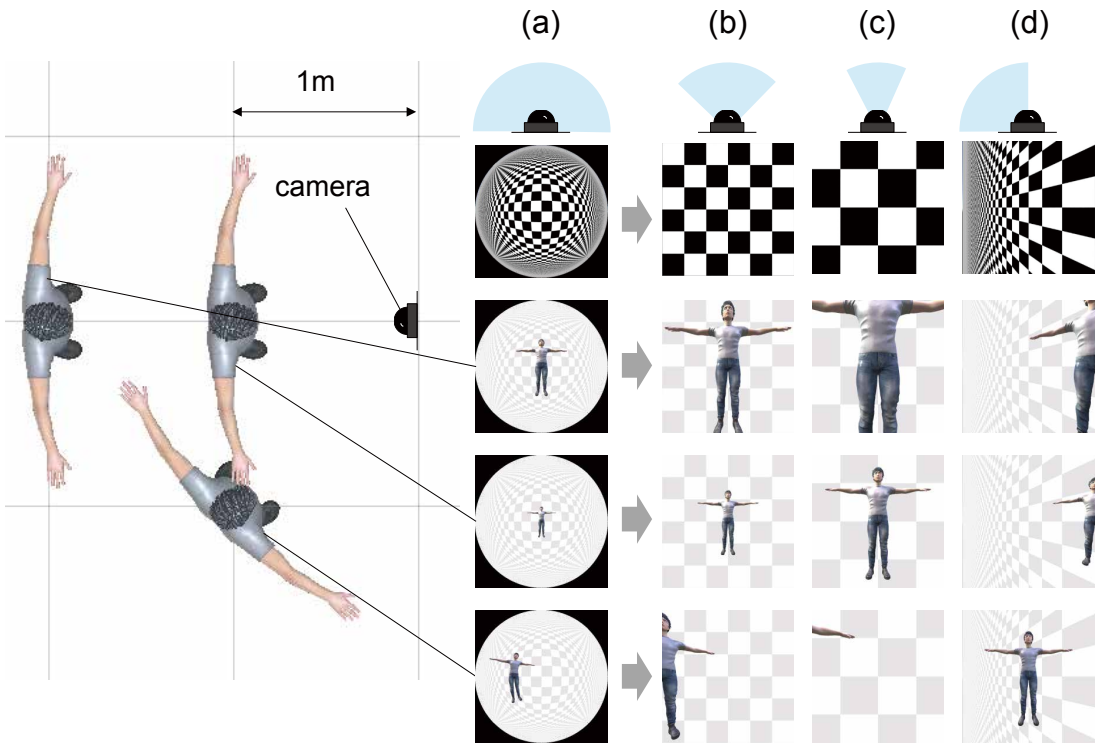


図 4.2　投影方法の比較 (a)EP 像および (b)$\theta_{FOV} = 45°$、(c)$\theta_{FOV} = 30°$、(d)$\theta_{FOV} = 45°$, $r_h = 45°$ における PP 像

4.2.3　3 次元人物姿勢の推定手法

先の注目点検出器および画像補正量推定器を併せて用いることで得られた 2 次元的な人物姿勢情報を用いることで、人物の 3 次元的な姿勢推定を行う。人物の 3 次元姿勢推定では、光学式モーションキャプチャ機器によって予め取得された学習用の 3 次元的な人物姿勢データセットに対して t 分布型確率的近傍埋め込み法（t-distributed Stochastic Neighbor Embedding, t-SNE）による次元削減を行った後、Expectation–Maximization（EM）法によるクラスタリングを行う。次に、各クラスタに含まれる要素について平均 3 次元姿勢 $\mathbf{Y}_n^*$ および正規直交行列 $\mathbf{e}_n$ を求める。さらに、平均姿勢 $\mathbf{Y}_n^*$ の 2 次元平面への射影 $\mathbf{y}_n^*$ および先の注目点検出器によって推定された人物の画像中の 2 次元姿勢 $\mathbf{y}$ を用い

$$\arg\min_{n,\mathbf{a},\mathbf{R}} \|\mathbf{y} - \mathbf{y}_n^*(\mathbf{a},\mathbf{R})\|^2 \tag{4.8}$$

を満たす平均姿勢のインデックス n、正規直交行列 $\mathbf{a}$ および回転行列 $\mathbf{R}$ を求め、対応する 3 次元的な人物姿勢

$$\mathbf{Y}_n(\mathbf{a},\mathbf{R}) = \mathbf{R}(\mathbf{Y}_n^* + \mathbf{a}\cdot\mathbf{e}_n) \tag{4.9}$$

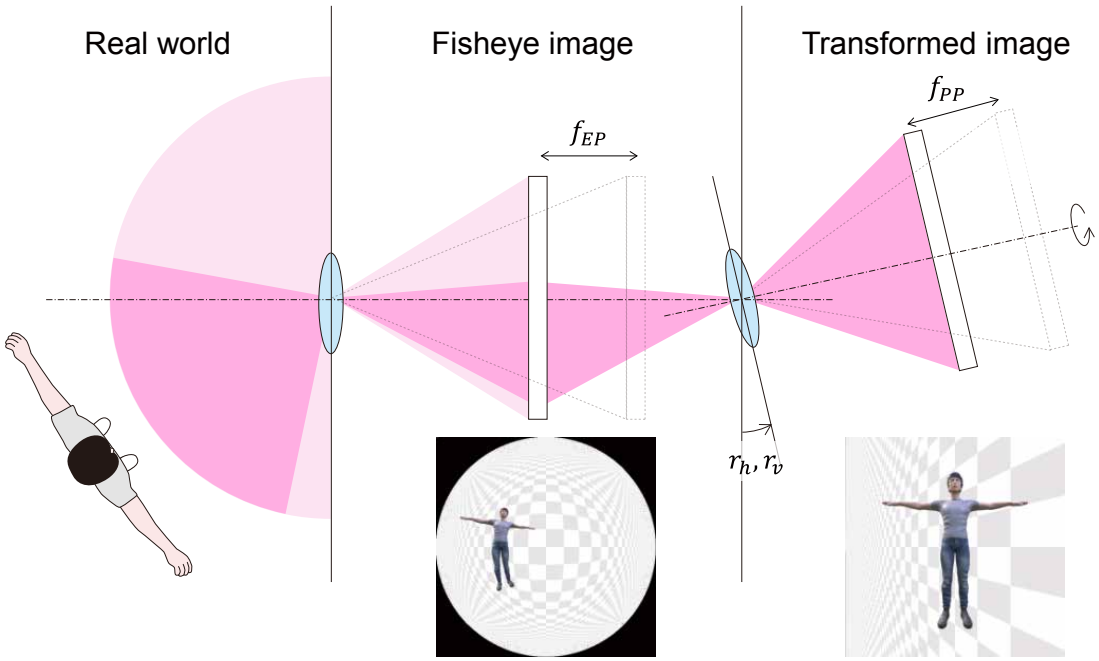


図 4.3　EP-PP 変換器

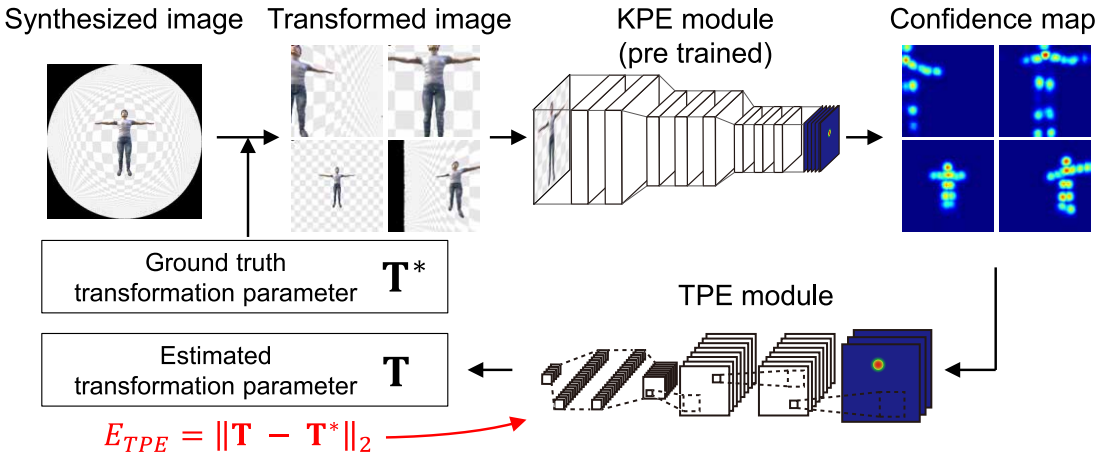


図 4.4　画像補正量推定器の学習手法

を得た。ここで、平均姿勢 $\mathbf{Y}_n^*$ の 2 次元平面への射影 $\mathbf{y}_n^*$ は投影行列 $\mathbf{\Pi}$ を用いて

$$\mathbf{y}_n^*(\mathbf{a}, \mathbf{R}) = \mathbf{R}\mathbf{\Pi}(\mathbf{Y}_n^* + \mathbf{a} \cdot \mathbf{e}_n) \tag{4.10}$$

で表される。

4.2.4　人物位置推定のための DNN モデルおよびその学習手法

図 4.5 に DNN モデルを用いた画像中の人物位置の推定方法を示す。画像中の人物位置の推定では、まず画像中における先の 2 次元的な関節位置座標をカメラ空間上での人物の位置を推

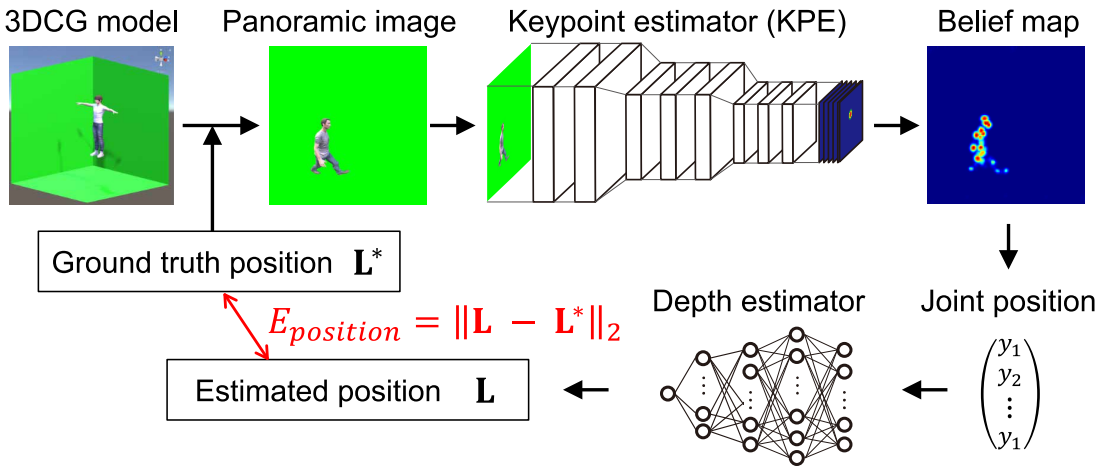


図 4.5　人物位置推定器の学習手法

定する全結合ニューラルネットワーク（Fully-Connected Neural Network, FCNN）へ入力する。FCNN モデルはカメラからの距離を推定し、距離情報と画像の射影方式を用いて実空間上の 3 次元位置を推定する。距離推定 FCNN モデルは 2 層の全結合層からなる DNN モデルとし、学習の際にはカメラ空間上の位置の真値 $\mathbf{L}^*$ に投影された 3DCG 人物モデル像を生成し、注目点推定のための CNN モデルおよび距離推定のための FCNN モデルを用いて 2 次元的な人物姿勢と人物位置を推定し、推定された人物位置 $\mathbf{L}$ および真値 $\mathbf{L}^*$ に関する以下の損失関数

$$E_{position} = \|\mathbf{L} - \mathbf{L}^*\|^2 \tag{4.11}$$

を最小化するように距離推定 FCNN モデルの学習を行った。

4.3　実験

4.3.1　２次元人物姿勢推定に関する定量的評価

注目点検出器の学習では、Max Planck Institute for Informatics の提供するデータセット [60] から抽出された人物を含む画像約 17,000 枚および、それぞれの 2 次元的な関節位置座標の真値を用いた。入力画像の大きさは縦、横それぞれ 368 ピクセルとした。確信度マップは画像中の関節位置を中心とした正規分布で示され、式 4.1 における標準偏差は $\sigma = 1.6$ とした。確信度マップの大きさは入力画像の縦、横それぞれを 8 分の 1(46 ピクセル) とした。

画像補正量推定器の学習では、3DCG ソフトウェア [61] により 50 人分の人物モデル（男性 25 人、女性 25 人）を用いることで 80,000 枚の画像を生成し学習に供した。3DCG モデルの関節位置座標データとして、Carnegie Mellon University Motion Capture Database（CMU MoCap）データセット [62] を用いた。本データセットは、光学式モーションキャプチャ機器によって取得された人物の関節位置に関する時系列データで構成される。各補正パラメータ $\mathbf{T}^*$

は、以下の一様分布 $U(\hat{f}_{PP}-\delta f, \hat{f}_{PP}+\delta f)$、$U(\hat{r}_h-\delta r_h, \hat{r}_h+\delta r_h)$、$U(\hat{r}_v-\delta r_v, \hat{r}_v+\delta r_v)$、$U(\hat{r}_r-\delta r_r, \hat{r}_r+\delta r_r)$ から抽出した。実験では、各上限および下限を $\hat{f}_{PP}=200$、$\hat{r}_h=0°$、$\hat{r}_v=0°$、$\hat{r}_r=0°$、$\delta f_{PP}=100$、$\delta r_h=40°$、$\delta r_v=40°$、$\delta r_r=20°$ により設定し、実験に供した。

本手法による 2 次元的な人物姿勢の認識精度に関する定量的な評価のため、Leeds Sports（LSP）データセット [59] に含まれる 1,000 枚の画像を用いて注目点検出器の認識精度を評価した。広角画像の歪み頑健であることを確認するため、LSP データセットに含まれる画像を式 (4.5) および式 (4.6) により歪ませ、特に広角での認識が可能であることを確認するため方位角を 0° から 90° の範囲で回転させ実験に供した。人物姿勢の認識精度の評価指標として、Percentage of Correct Keypoints（PCK）を用いた評価手法 [63] を採用した。本指標では、画像中における関節位置の真値と推定値との距離の閾値をを画像中の人物の大きさに基づいて決定し、認識精度を評価する。具体的には、LSP データセット [64] の評価プロトコルに従い PCK@0.2 の値を比較した。ここで、PCK@n は関節位置の推定値 y_j とその真値 y_j^* の間の距離を画像内における人物の外観に基づいて正規化した

$$d_j = \frac{|y_j - y_j^*|}{|y_{shoulder} - y_{hip}|} \tag{4.12}$$

について $d_j < n$ の際に正しく推定されたと判定し、得られる認識精度である。

4.3.2 3 次元人物姿勢推定に関する定性的評価

3 次元的な人物姿勢を推定するために、CMU MoCap データセット [62] に含まれる 3 次元的な人物姿勢データからランダムに抽出された 10,000 の姿勢データを学習用データセットとして用いた。ここでは、各クラスタから 8 の平均姿勢モデル $\mathbf{Y}_n^*$ を選択した。各 3 次元的な人物姿勢情報は、回転行列 $\mathbf{R}$ により、360° の範囲において 5° 間隔で床面に垂直な軸を中心に回転させた後、2 次元平面に投影される。投影された 2 次元的な関節位置座標について y 軸座標が $[-1, 1]$ の範囲内に収まるように正規化した。

提案手法による 3 次元的な人物姿勢認識能力について定性的に評価するため、赤外線カメラ式民生用モーションキャプチャ機器（Microsoft, Kinect v2）[65] との比較として、撮像機器から距離 0.8 から 2.0 m、方位角は 0° から 45° の範囲に存在する人物に対する姿勢の認識実験を行った。広角カメラとして魚眼レンズを有するカメラ（RICOH, THETA S）[66] を用い、撮像された画像の解像度を 368×368 ピクセルに調整し入力データとした。また、人物動作に関する特徴量を適切に抽出できるか確認するため、3DCG ソフトウェアを用いて人物が歩行する広角映像を生成し、人物の姿勢および位置に関する真値と推定値の比較を行った。映像データは様々な体型を模擬可能な人物の 3DCG ソフトウェアを用いて作成し、モーションキャプチャデータとして CMU Mocap データセット用いた。特に 3 次元的な動作の認識のた

め、CMU Mocap データから自然な歩行動作（Subject 91, trial 2）および不自然な歩行動作（Subject 91, trial 18）に対して人物姿勢の認識および得られた人物姿勢から体の向き、視線方向、移動速度の推定を行い真値と比較した。

4.4　結果および考察

4.4.1　２次元人物姿勢推定

２次元的な人物姿勢の認識精度の評価結果を図 4.6 および表 4.1、表 4.2 に示す。提案する画像補正量推定器を用いた場合においては認識精度が向上し、r_h が 0° から 70° の間では提案手法の認識精度が優れていることが確認された。水平方向の回転量 r_h が 0° から 70° の間において、PCK@0.2 は約 70% 程度であった。また、$r_h > 70°$ の場合、認識精度が低下することが確認されたが、これは画像内の人物の一部が画角に収まりきらないことが原因と考えられる。これらの結果は、提案手法が広角画像の歪みに頑健な人物姿勢認識を実現するうえで有利であることを示すものである。

次に補正パラメータの更新回数が認識精度に及ぼす影響を確認した。図 4.7 は LSP データセットに対する補正パラメータの更新回数と認識精度の関係である。5 回の補正パラメータ更新で姿勢が概ね適切に認識され、その後は同等の認識精度が得られた。したがって、以降の実験では 5 回の補正パラメータ更新を適用した。

赤外線式モーションキャプチャ機器と提案手法による人物姿勢認識結果の比較を図 4.8 に示す。カメラから距離 2.0 m の位置においては、両者において人物姿勢を適切に推定可能であった（図 4.8(a)）。提案手法では、画像の歪みにより人物の見かけの大きさが小さくなったにも関わらず、姿勢を推定することが可能であった。より近距離（0.8 m）では、赤外線式モーションキャプチャ機器を用いた場合は画角が制限されるため、人物の像の一部が欠損し、人物姿勢の認識に失敗している（図 4.8(b)）。一方で提案手法では、広角画像内に全身の画角に収まり、適切に人物姿勢の認識が可能であった。また、人物が赤外線式モーションキャプチャ機器の画角に収まらない位置に存在する場合においても、広角カメラにおいては人物の全身が画角に収まった画像を撮像可能であり、提案手法では人物姿勢の推定が可能であった（図 4.8(c)）。

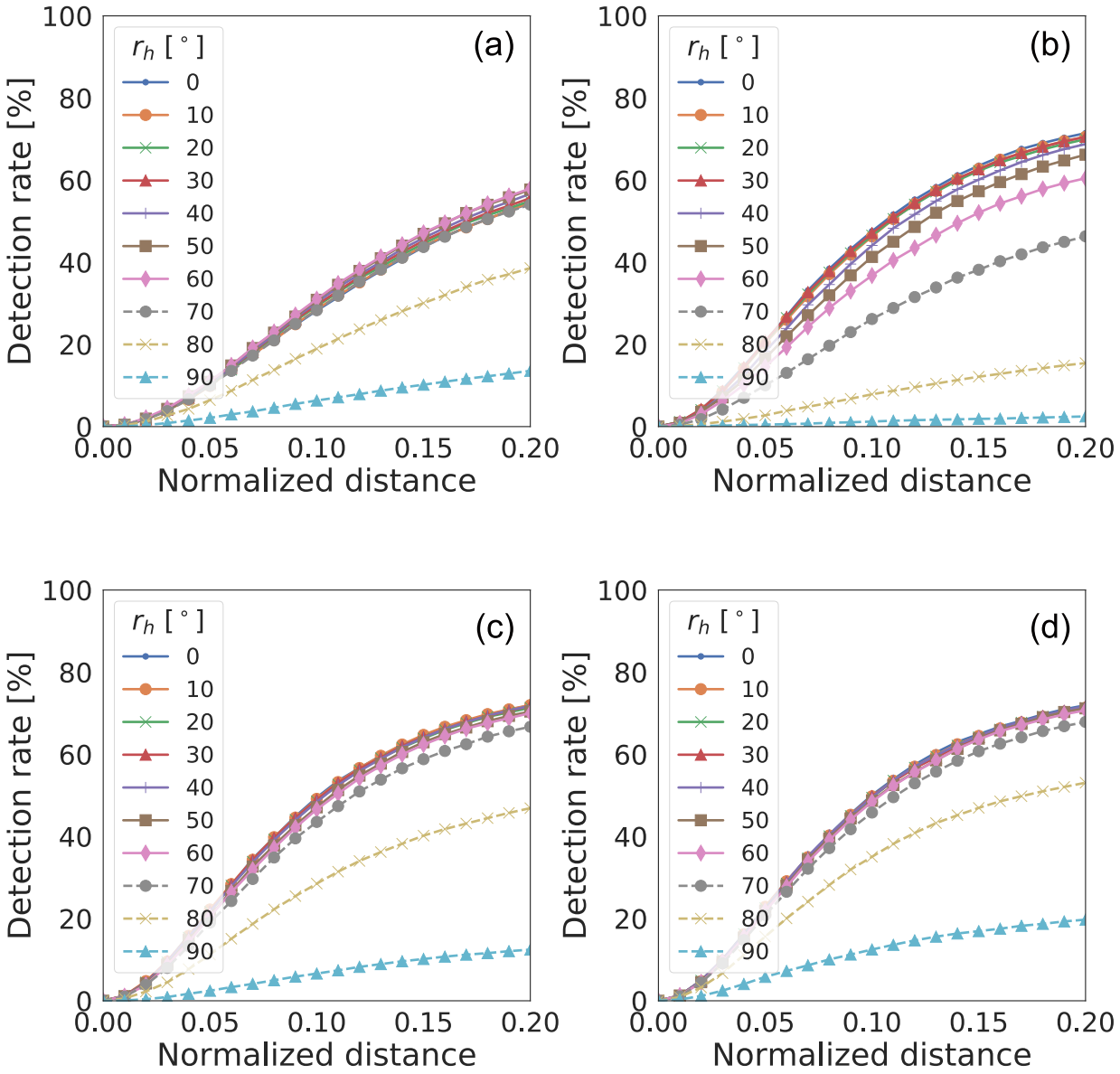


図 4.6　LSP データセットに対する (a)TPE なしおよび (b)1 回、(c)3 回、(d)5 回の補正を行った際の認識精度の比較

表 4.1　補正パラメータをそれぞれ $f^*_{PP} = 200$、$r^*_h = 0°$、$r^*_v = 0°$ および $r^*_r = 0°$ とした際の各部位の認識精度に関する比較

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
no TPE	55.1	66.6	52.0	42.5	62.4	54.8	48.5	54.4
TPE (x1)	83.5	78.4	65.3	58.5	76.2	73.3	65.0	71.4
TPE (x3)	84.8	78.6	65.9	58.4	77.6	73.6	64.9	72.0
TPE (x5)	84.4	79.7	66.0	59.4	77.3	72.8	64.1	71.9

表 4.2　補正パラメータをそれぞれ $f^*_{PP} = 200$、$r^*_h = 60°$、$r^*_v = 0°$ および $r^*_r = 0°$ とした際の各部位の認識精度に関する比較

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
no TPE	61.8	68.5	54.5	46.3	68.1	59.0	47.4	58.0
TPE (x1)	69.9	68.2	53.7	45.3	68.8	64.3	55.1	60.5
TPE (x3)	82.5	77.3	64.6	55.5	76.4	71.6	62.4	70.0
TPE (x5)	83.9	77.7	65.2	57.7	76.3	72.1	62.2	70.7

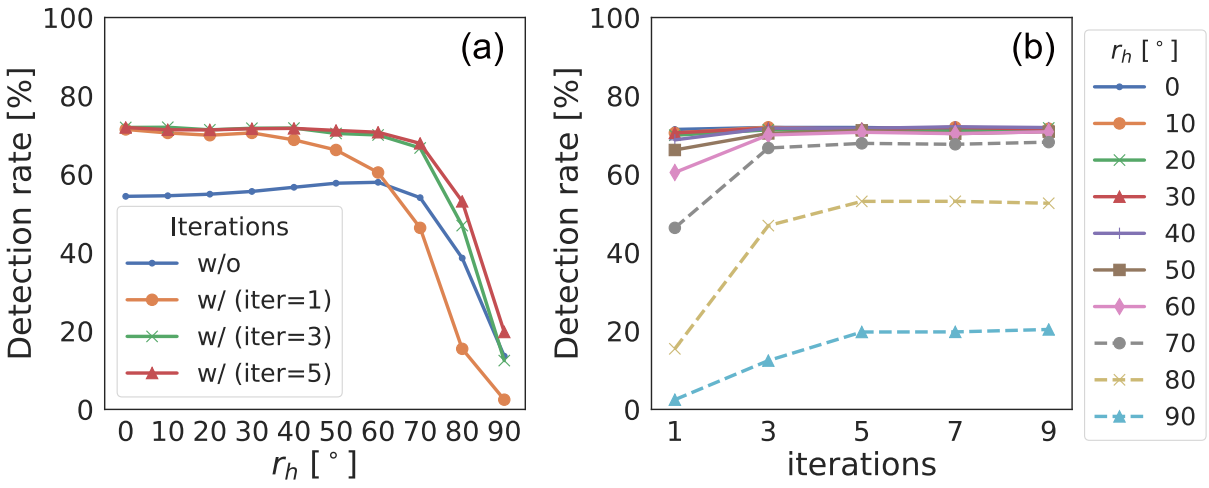


図 4.7　LSP データセットに対する (a) 方位角 (b) 補正パラメータの更新回数ごとの認識精度の比較

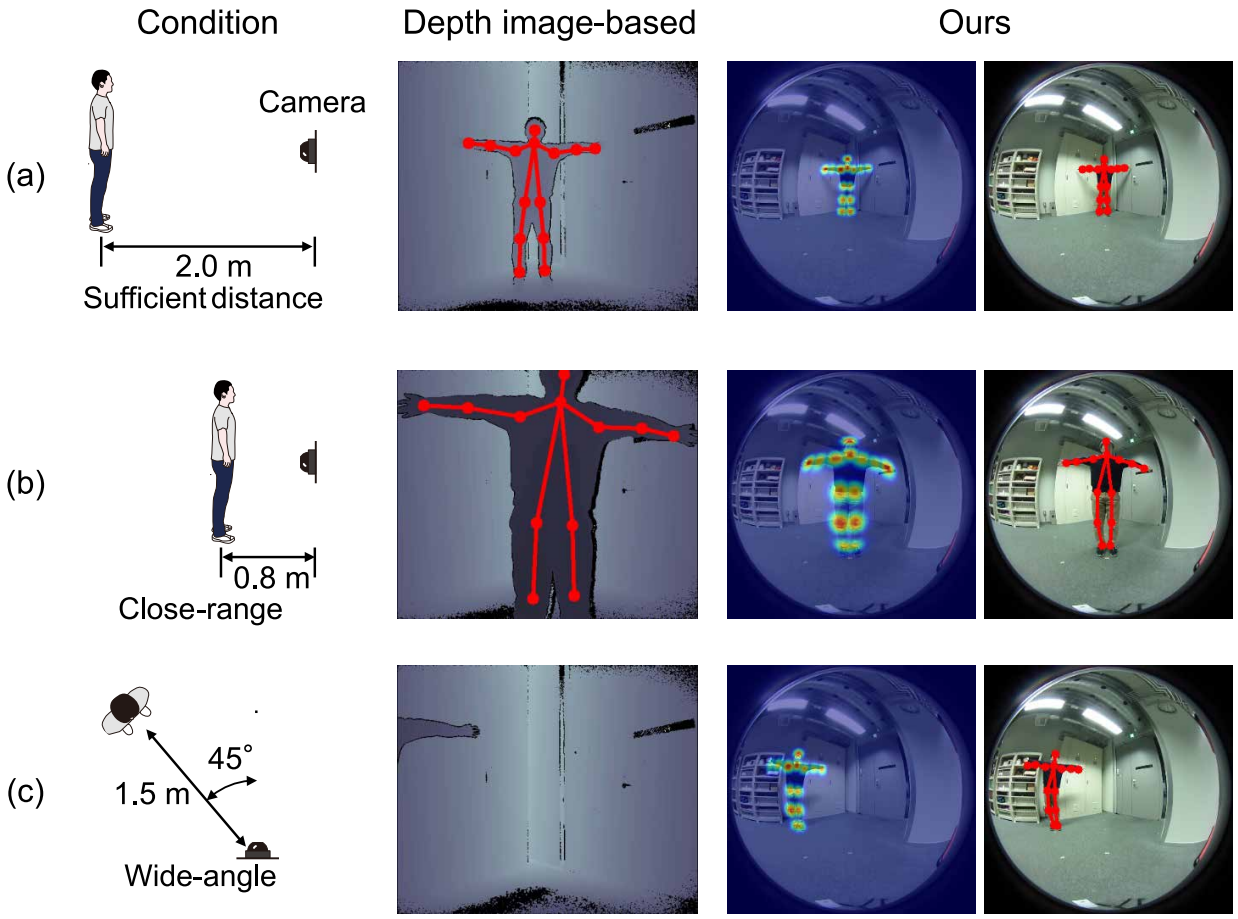


図 4.8　赤外線式モーションキャプチャ機器と提案手法による 2 次元姿勢認識結果

4.4.2　３次元人物姿勢推定

図 4.9 に、赤外線式モーションキャプチャ機器を用いた手法および提案手法による人物の 3 次元姿勢推定の結果を示す。図 4.9(a) のように人物とカメラの距離が適切である場合、両者において 3 次元的な人物姿勢を推定可能であった。また、図 4.9(b) および (c) のように人物が近距離に存在する場合、赤外線式モーションキャプチャ機器を用いた方法では画像内に人物の全身が収まらず、適切に姿勢を認識できていない。一方で提案手法では広角画像に人物の全身が収まり、3 次元的な姿勢を適切に認識可能であった。さらに、図 4.9(c) では、赤外線式モーションキャプチャ機器を用いた方法においては画像内の人物の一部が僅かに映る場合や、人物が周辺の物体と一体化している場合には姿勢の認識が行われなかったが、提案手法ではいずれの場合においても 3 次元的な姿勢を適切に認識可能であった。

次に、人物の自然な歩行動作および、不自然な歩行動作を含むモーションキャプチャデータから生成した 3DCG 映像に対し、提案手法による 3 次元姿勢の復元および人物位置の推定を行い、人物動作の特徴量を抽出した結果を図 4.10 に示す。それぞれの歩行動作に関する特徴量を、真値と比較して良好な認識精度で推定可能であることが確認された。特に不自然な歩行動作において体の向きや視線方向の変動、移動速度の変動に対して特徴的な傾向を把握する上で十分な精度で推定を行うことが可能であった。

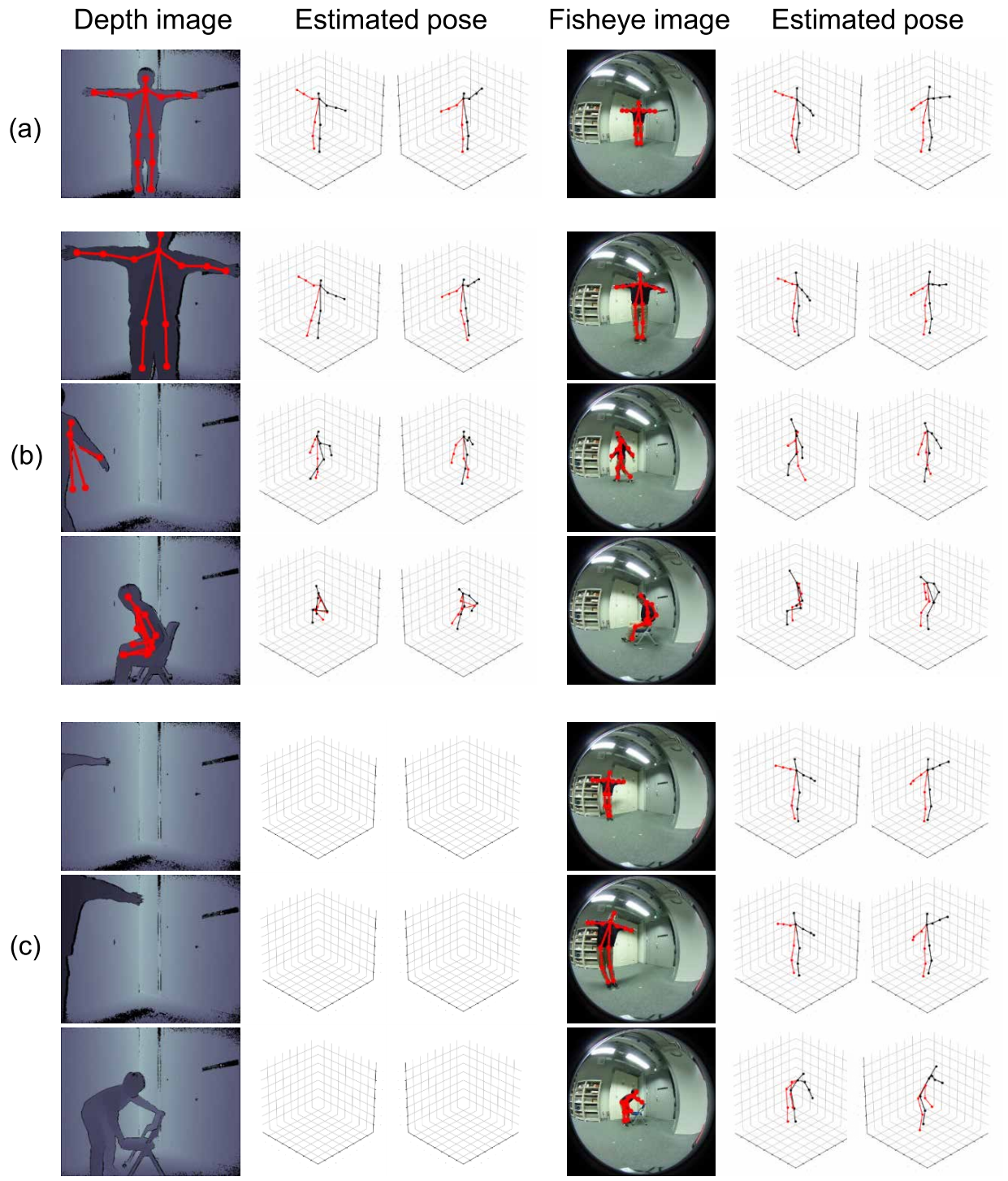


図 4.9 赤外線式モーションキャプチャ機器および提案手法による 3 次元姿勢認識結果

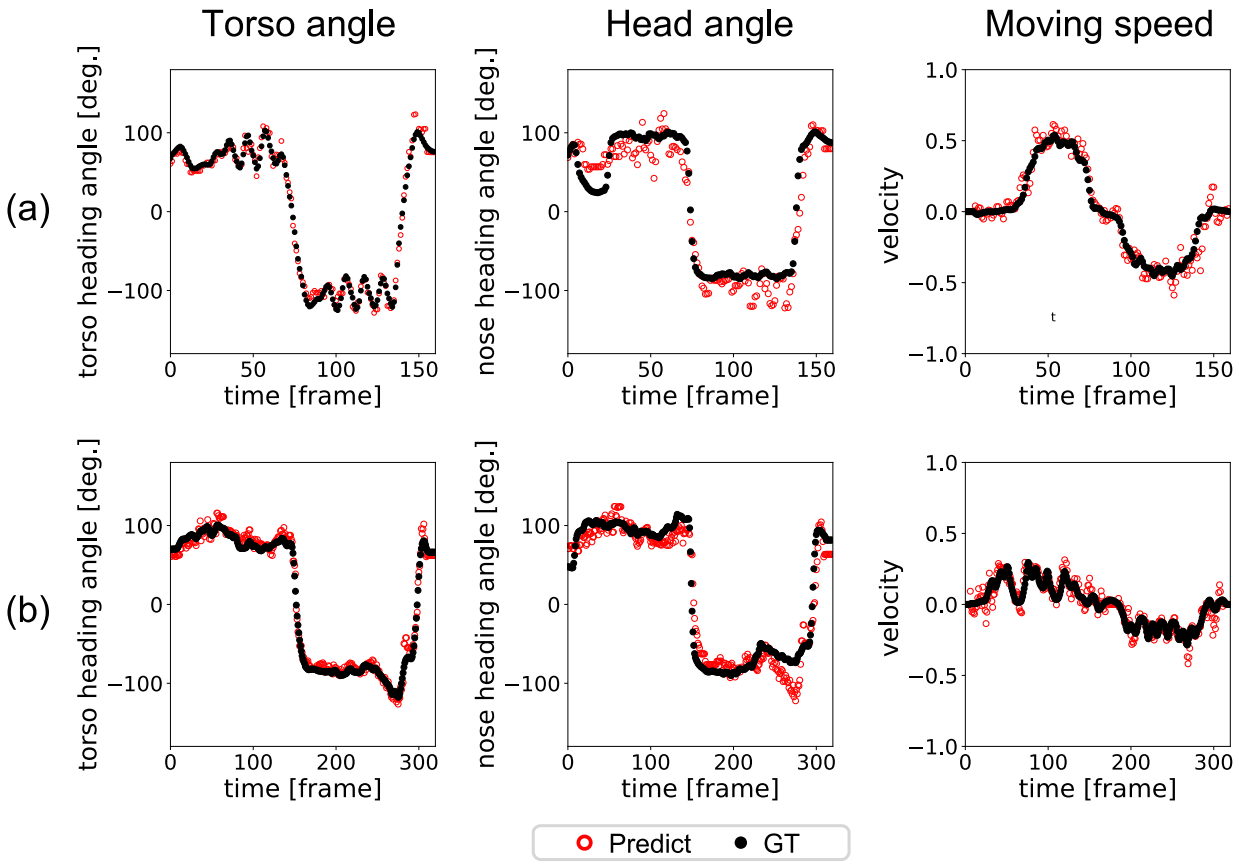


図 4.10　不自然な動作を (a) 含まないおよび (b) 含む人物動作データから抽出された特徴量

4.5 結言

本章では画像中の人物の動作を表す特徴量を抽出するための広角画像の歪みに頑健な人物姿勢の推定手法を提案した。広角画像の歪みに頑健な人物姿勢の認識のため、画像から人物の姿勢および適切な補正パラメータを推定可能な新たな DNN モデルおよびその学習手法を提案した。実験では水平方向 0° から 70° の範囲内に存在する人物を良好な精度で認識可能であり、広角での人物姿勢認識が可能であることを確認した。さらに提案手法は、赤外線式モーションキャプチャ機器を用いた方法と比較して、より近距離および広角での人物姿勢認識が可能であった。また、本手法を人物の自然な歩行動作および不自然な歩行動作を模擬した映像データの解析に適用することで、それぞれの歩行動作に関する特徴を真値と比較して良好な精度で抽出することが可能でった。特に不自然な歩行動作において体の向きや視線方向の変動、移動速度の変動に対して特徴的な傾向を把握することが可能であった。以上の結果から、監視カメラ映像から人物動作の特徴抽出の実現可能性が見出された。次章では人物動作特徴量に関する時系列データから人物の行動解析および異常検知を行うための方法について述べる。

第 5 章

時空間グラフ畳み込みネットワークを用いた人物動作解析

5.1　緒言

前章では提案した多変量時系列データの解析手法を人物動作解析に適用するための人物動作特徴抽出手法について述べた。得られた人物動作特徴量に関する時系列データを解析することで、人物の動作を識別したり通常動作からの逸脱を検知できれば、原子力関連施設の安全対策がより強固になる可能性がある。そこで本研究では、人物の関節位置に関する時空間的情報から適切に人物行動に含まれる特徴量を抽出可能な DNN モデルについてその構造を検討し、第 2 章および第 3 章で述べた弱教師あり学習手法によりそのパラメータを最適化することで人物行動の解析および異常検知を行う。第 2 章のように詳細なアノテーションを必要とせずに時系列データに含まれる異常が検知できれば、人物の不自然な動作のように定義の曖昧な動作の検知へ応用できる可能性がある。さらに第 3 章で述べた多クラス識別への応用ができれば人物動作の識別が可能となり、人物の危険な状態や行動の検知等に役立つ可能性がある。

本章では、人物姿勢情報を適切に扱うための DNN モデルの検討と人物行動認識への適用について述べる。次に、提案手法による人物行動の識別およびローカライゼーションに関する評価について述べる。最後に、人物動作からの異常検知への応用について述べる。

5.2　方法

映像中の人物動作を解析する技術は映像監視、ヒューマンコンピュータインタラクション、医療等への応用が期待されており、人物姿勢の認識技術と同様に活発な研究が行われている。先行研究として、人物動作に関する特徴量エンジニアリングに基づく方法 [67, 68, 69, 70, 71] が多く報告されているが、特に最近では人物姿勢認識技術と同様に、DNN モデルを用いた方法

が提案されており、CNN のような映像中の人物動作に関する空間的な情報を用いた方法 [72]、RNN や LSTM のような再帰型ニューラルネットワークを用いた時間的な情報を上手く取り扱う方法 [73, 74, 75]、グラフ畳み込みニューラルネットワーク（Graph Convolutional Network, GCN）を用いることで人物動作をグラフ構造として扱う方法 [76, 77, 78, 79, 80, 81] 等が提案されている。Du らは、人物姿勢の空間的な情報と時間的な情報を表現する階層的な RNN モデル [73] を提案した。Liu らは時空間的な情報を扱うための 2 次元的な接続をもつ LSTM 層をもつ DNN モデル [74] を提案した。また、距離画像センサによって推定される関節位置の測定誤差に頑健とするために trust gate mechanism[75] を提案した。Yan らおよび Li らは、時間的および空間的に接続された GCN（Spatial-Temporal Graph Convolutional Network, ST-GCN）[76, 77] を用いた人物動作解析手法を報告した。特に ST-GCN を用いた手法は人物の関節位置座標に関する一連の座標情報の時間変化をグラフ構造として扱うことで優れた識別精度が得られることを報告しており、本手法を改良した様々な手法 [78, 79, 80, 81] が提案されている。Li らは行動に関するリンク（Actional-link）と構造に関するリンク（Structural-links）を考慮した GCN である Actional-Structural Graph Convolutional Network（AS-GCN）[78] を提案した。また、Si らは ST-GCN に、Attention Enhanced Graph Convolutional LSTM（AGC-LSTM）層 [79] を導入することで認識精度の向上を実現した。

これらの ST-GCN およびその応用では、人物行動識別において優れた性能を報告しているが、いずれの方法も数十から数百の時点に渡る人物姿勢に関する時系列データを入力データとして用い、それらを全体的に評価した上で認識すべき行動種別ごとに単一の推定値を出力する仕組みとなっている。つまり一連の時系列データにおける行動を決定付ける動作を含む箇所の推定（行動ローカライゼーション）への応用には適さなかった。映像中の各時点における人物の姿勢情報は人物の行動を決定付ける上で必ずしも重要とは限らないため、解析により映像中に人物の行動を決定付ける動作が含まれている箇所を明らかにすることができれば、映像監視等の高度化に役立つ可能性がある。このような人物行動認識に関する研究のため、様々なデータセットが提案されており、これらはそれぞれ数十から数百時点の映像および人物姿勢情報と、その一連の動作が表す行動を説明する単一の教師信号から成る。しかし、人物の動作のような自由度の高い時系列データにおいて、その動作を決定付ける上で必要な動作が含まれる箇所や程度を定量的に把握し、それらに関する教師信号を人手で付与することは困難である。一方で、第 2 章および第 3 章で述べた時系列データ解析のための DNN モデルおよびその学習手法を応用することで、上記のようなデータセットを用いた人物行動ローカライゼーションの実現が期待される。そこで本研究では、第 2 章および第 3 章で確立した DNN モデルを改良することで、人物動作に関する時系列データを用いた人物行動識別および異常検知を行う。

5.2.1 時空間グラフ畳み込み演算

人物の姿勢情報は 2 次元、または 3 次元的な人物の関節位置座標に関する一連の時系列データから成り、それらの時系列データを用いることで人物の動作を表現することができる。人物の行動は、人物の関節位置に関する座標情報について、同一時点における異なる関節位置間の空間的な位置関係と、異なる時点における同一の関節位置間の時間的な位置変化を特徴量としてもつ。つまり、人物姿勢について関節位置をノード、その接続をエッジとしたグラフ構造として表現し、データの空間的な情報を扱う際には単一時点における人物の各関節同士を自然に接続したグラフ構造、データの時間的な情報を扱う際には前後時点における同一ノードを時間的に接続したグラフ構造を用いることで両者の特徴量を上手く扱うことができる。これらの空間的および時間的な特徴量は、姿勢情報を用いた人物行動認識においてどちらも重要であり、このような時空的な接続をもつ GCN を扱うため、Yan らは人物姿勢に関する時系列データに対して空間的なグラフ畳み込みを行う Spatial-GCN（S-GCN）と時間的なグラフ畳み込みを行う Temporal-GCN（T-GCN）の両方を持つ DNN モデルによって人物の行動を識別する ST-GCN モデル [76] を提案している。

長さ T 時点を持ち、単一時点に I ノードをもつノード集合 $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, I\}$ および、エッジ集合 E から成るグラフ $G = (V, E)$ における注目ノード v_{ti} の隣接ノード集合は

$$B(v_{ti}) = \{v_{tj} | d(v_{tj}, v_{ti}) \leq D\} \tag{5.1}$$

と表される。ここで、$d(v_{tj}, v_{ti})$ は、v_{tj} から v_{ti} までの距離を表し、本研究では $D = 1$ とした。また、注目ノード v_{ti} の隣接ノード集合 $B(v_{ti})$ に対し、マッピング関数 m_{ti} によって各ノードの部分集合への分割が行われる。具体的には、隣接ノード集合を注目ノード（self-connections）、重心に近いノード（centripetal-node connection）、および残りのノード（centrifugal-node connections）に分割するために、同一時点での姿勢情報におけるのすべての関節位置の平均座標をその重心とし、マッピング関数を

$$m_{ti}(v_{tj}) = \begin{cases} 0 & \text{if } r_j = r_i \\ 1 & \text{if } r_j < r_i \\ 2 & \text{if } r_j > r_i \end{cases} \tag{5.2}$$

とした（図 5.1）。ここで、r_i はグラフの重心から i 番目のノードまでの平均距離である。空間的なグラフの畳み込み演算は

$$h^{(l+1)}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} h^{(l)}(v_{tj}) \cdot \mathbf{w}(m_{ti}(v_{tj})) \tag{5.3}$$

で表される。ここで、$h^{(l)}(v_{tj})$ は l 層目の隠れ層の出力、$\mathbf{w}$ は重みパラメータである。重みパ

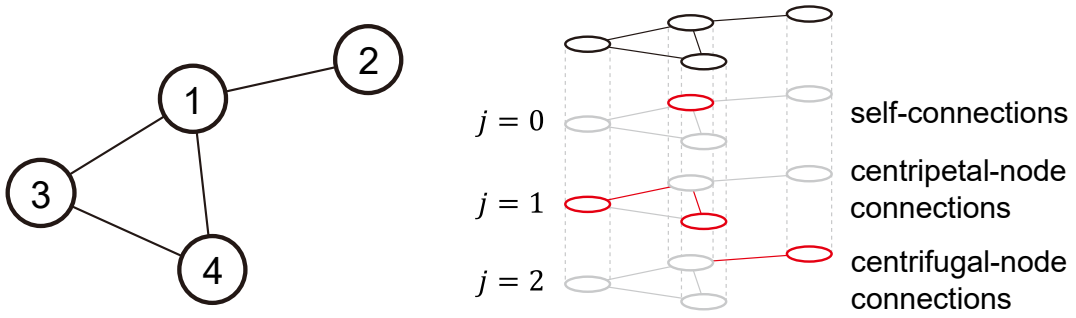


図 5.1　マッピング関数の概要

ラメータ $\mathbf{w}(v_{ti}, v_{tj})$ は先のマッピング関数を用いることで、

$$\mathbf{w}\left(v_{ti}, v_{tj}\right) = \mathbf{w}'\left(m_{ti}\left(v_{tj}\right)\right) \tag{5.4}$$

とも表すことができる。また、$Z_{ti}\left(v_{tj}\right) = \left|\{v_{tk}|m_{ti}\left(v_{tk}\right)\}\right|$ であり、出力の正規化に用いられる。

空間的なグラフの畳み込み演算を時間方向に等間隔で配列されたデータを扱えるようにするため、以上の内容を拡張する。グラフ構造を時間方向に拡張することで、時刻 t における i 番目のノード v_{ti} に対する隣接するノード集合は

$$B(v_{ti}) = \{v_{qj}|d\left(v_{tj}, v_{ti}\right) \leq K, |q-t| \leq \lfloor \Gamma/2 \rfloor\} \tag{5.5}$$

と表すことができる。ここで Γ は一度に畳み込むカーネル幅であり、本研究では $K = 1$、$\Gamma = 9$ とした。空間的グラフ畳み込みの際と同様にマッピング関数 m_{ST} を導入した。v_{ti} 周辺のノードを畳み込む際のマッピング関数 m_{ST} は単一時点におけるマッピング関数を $m_{ti}(v_{tj})$ とするとき

$$m_{ST}(v_{qj}) = m_{ti}(v_{tj}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K \tag{5.6}$$

である。

5.2.2　ST-GCN モデルの実装

本研究では、Yan らの実装 [76] を参考に、グラフ畳み込みを

$$\mathbf{H}^{(l+1)} = \sum_j \mathbf{\Lambda}_j^{-\frac{1}{2}} \mathbf{A}_j \mathbf{\Lambda}_j^{-\frac{1}{2}} \mathbf{H}^{(l)} W_j \tag{5.7}$$

で表現する。ここで、$\mathbf{H}^{(l)}$ および $\mathbf{H}^{(l+1)}$ は入力および出力特徴量、$\mathbf{W}_j$、$\mathbf{D}_j$ および $\mathbf{A}_j$ は、それぞれ重み行列、次数行列および隣接行列である。$\mathbf{A}_j$ は、$\mathbf{A}_0$、$\mathbf{A}_1$、$\mathbf{A}_2$ の 3 つの行列から成り、それぞれ注目ノード、重心に近いノード、および残りのノードに対応した隣接行列である。さらに、$\Lambda_j^{ii} = \sum_k A_j^{ik}$ である。

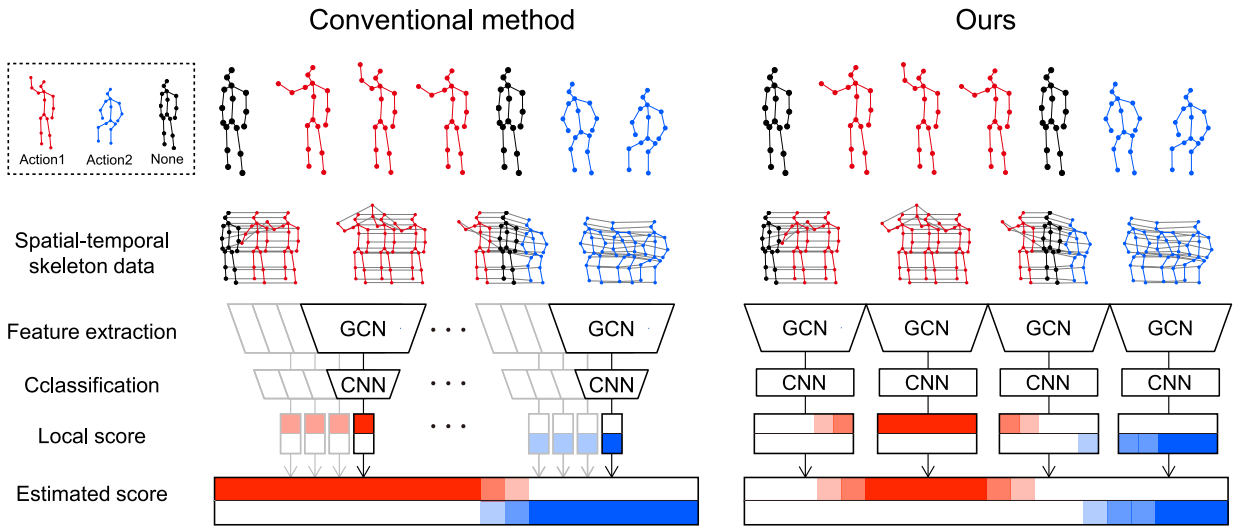


図 5.2　人物の行動認識手法の比較（左：ST-GCN を用いた従来法、右：提案手法）

Yan らの提案する ST-GCN [76] をはじめとして多くの研究で提案されている方法によって得られる特徴量は、時点数を T、認識すべき行動クラスの数を K とするとき $T \times K$ 次元の行列である。しかし、ST-GCN を用いた従来法では複数の時点におけるデータを入力し、それらを全体的に評価した単一の推定値を出力する。つまり、出力は K 次元のベクトルとなる。これは、最終層の出力ををソフトマックス関数により活性化した後、One-hot ベクトルの形式であらかじめ用意された教師信号との交差エントロピー損失を最小化するように DNN モデルを学習するためである。一方で提案手法では最終層を図 5.2（右）に示すように、$T \times K$ 次元の行列を出力するようにした。これにより複数時点におけるデータ入力に対して、それぞれに対応する複数の推定値が出力される。これまでに述べたようにそれぞれの時点において特定の行動の含まれる期待値の真値は未知であるが、第 2 章および第 3 章で述べた学習手法を以上の DNN モデルに適用することで、人物行動のローカライゼーションが可能な DNN モデルを学習する。

5.3　実験

5.3.1　人物行動認識に関する定性的および定量的評価

提案手法の人物行動の認識能力について定性的および定量的な評価を行うために、以下の公開データセットを用いた人物行動の識別およびローカライゼーションに関する評価を行った。

UTD-MHAD データセット

UTD-MHAD データセット [82] は民生用モーションキャプチャ機器（Microsoft, Kinect）により撮像されたデータセットであり、RGB 画像、距離画像、関節位置座標情報および加速度センサ情報を含む。本データセットに含まれる各人物姿勢情報は、20 点の関節位置に関する 3 次元座標から成り、8 名の人物それぞれが 27 種の動作を 4 回ずつ実施した様子が記録されている。評価では、データセット提供者の方法 [82] に従って、それぞれのデータを被験者番号 $\{1, 3, 5, 7\}$ および $\{2, 4, 6, 8\}$ に分割し、前者を学習用、後者を評価用とした。すべてのデータの長さを 128 時点に統一し、データ長が不足する場合にはゼロパディングを施した。

SYSU データセット

SYSU データセット [83] もまた民生用モーションキャプチャ機器（Microsoft, Kinect）で撮像されたデータセットであり、40 名の人物による 12 種の動作が記録されている。本研究では本データセットを、UTD-MHAD データセットに含まれないデータ、つまり負のデータとして用い、提案する行動認識手法が負のデータに対してどのような挙動を示すか確認するために用いた。

NTU RGB+D データセット

NTU RGB+D データセット [84] は上記データセットとは別の民生用モーションキャプチャ（Microsoft, Kinect v2）によって撮像され、$56,000$ 時点を超えるデータが含まれる。データセットには 40 名の人物による 60 種の動作データが含まれ、各人物の姿勢に関する 25 点の 3 次元的な関節位置の座標情報が記録されている。本データセットの提供者は評価方法として、異なる人物から得られたデータにより評価を行う、Cross Subject（CS）および同一の人物を視点が異なる 2 台のカメラで撮像されたデータにより評価を行う、Cross View（CV）の 2 つの評価方法を推奨しており、本研究では前者を $40,320$ の学習用データと $16,560$ の評価用データ、後者を $37,920$ の学習用データと $18,960$ の評価用データを含むように分割した。

人物行動識別のために提案するネットワーク構造として、GCN 層は 9 つのブロックに分割された構造とした。最初の 3 層は 64 チャネル、次の 3 層には 128 チャネル、最後の 3 層は 256 チャネルとし、時間方向について 4 番目と 7 番目の層の後に平均プーリング層、また、各 GCN 層の後に、ドロップアウト層を導入した。データ拡張として、人物姿勢データに対して床面に垂直な軸を中心に -30° から 30° の回転を施した。さらに、人物の体格や動作速度の違いに頑健にするために、90% から 110% の範囲でスケール変換および 0 から 10% の確率でデータをランダムに削除した。さらに、関節位置の推定誤差を模擬するために、正規分布に従うノイズをデータに付与した。DNN モデルの学習では、式 (3.3) の損失関数を最小化するように各ハイパパラメータを $p_1 = 10^{-5}$、$p_2 = 10^{-2}$、$p_3 = 10^{-2}$、学習率 10^{-4} として DNN モ

デルのパラメータを最適化した。

5.3.2 定義の曖昧な人物行動の認識

人物の不自然な動作が本手法で検知可能であるか確認するために、CMU Mocap データセット [62] を用いて実験を行った。実験では、人物の歩行動作から不自然な動作が検知可能か確認するために、データセットに含まれる「Walk」および「Weird Walks」のタグが付与されたデータを学習用データとして用いた（表 5.1）。本データに含まれる人物は歩行等によりその位置が変化するため、位置変化に頑健とするために DNN モデルの学習時に人物の重心位置に対する 14 点の関節位置の 3 次元的な相対位置を空間的特徴量（3 次元）として抽出した。同様に、それぞれの関節位置の 10 時点間の 3 次元的な移動量を時間的特徴量（3 次元）して抽出した。これらの空間的および時間的特徴量の両者を統合し、人物の 14 点の関節に対してそれぞれ 6 次元の特徴量をもつ学習用データおよび評価用データとした。入力および出力の次元数が異なることを除いて、ネットワーク構造は上記の実験と同様の構造とした。DNN モデルの学習では、式 (2.29) の損失関数を最小化するように各ハイパパラメータを $p_1 = 10^{-1}$ および $p_2 = 10^{-3}$、学習率 10^{-3} として DNN モデルのパラメータを最適化した。

表 5.1 不自然動作の検知に関する実験に用いたデータインデックス

	Training	**Testing**
Positive	132	91(trial 18)
Negative	7,8,16,35,36,37,38,39,69	91(trial 2)

5.4 結果および考察

5.4.1 人物行動識別および行動ローカライゼーション

図 5.3 に UTD-MHAD データセットに含まれる人物動作データに対して提案手法によって学習された DNN モデルにより推定されたそれぞれの行動が含まれる期待値を示す。提案手法では、それぞれの行動を決定づける上で重要な動作が含まれる時点に対して高い値が適切に推定され、その他の時点では低い値が推定された。さらに、図 5.4 に示すように、負のデータに対しては十分に低い値が推定された。図 5.5 は、学習時のパラメータの更新回数と提案手法により学習された DNN モデルの出力値の関係を示す。パラメータの更新回数が 1,000 回の場合には、第 2 章図 2.4（下）のように全体的に低い値を推定する、真陰性となりやすい一方で偽陰性にもなりやすい状態に陥っていると考えられる。学習が進み、パラメータの更新回数が

3,000 回の場合には、正のデータに含まれる行動に対する推定値が高くなるが、負のデータに含まれる行動を誤検知することがあり、第 2 章図 2.4（上）のように全体的に高い値を推定する真陽性となりやすいが偽陽性にもなりやすい状態に陥っていると考えられる。さらに学習が進むことで、正のデータに含まれる行動に対する推定値が高くなる一方で負のデータに含まれる行動に対する推定値が低くなり、適切に推定が行われることが確認された。以上の結果から、DNN モデルの学習が意図した通りに行われていることがわかる。

図 5.6(a) に UTD-MHAD データセットに対して得られた混同行列および表 5.2 に行動識別精度に関する従来手法との比較を示す。ここで、人物動作の認識精度を算出するため、推定値に関する各時点の合計値が最大であった動作クラス

$$\text{detected action} = \arg\max_{k} \sum_{t} y_k^{(t)} \tag{5.8}$$

を検知された動作とした。本手法では従来手法と比較して同等の識別精度が確認された。同様に、NTU RGB+D データセットに対する混同行列および識別精度を表 5.2 および図 5.6(b) に示す。提案手法は ST-GCN と同等の識別精度を有し、より大規模なデータセットに対しても人物行動を適切に識別できることを示している。実装した ST-GCN での識別精度は文献 [76] より低い値が推定されたものの、概ね同等の認識精度が得られ、評価の妥当性が確認された。

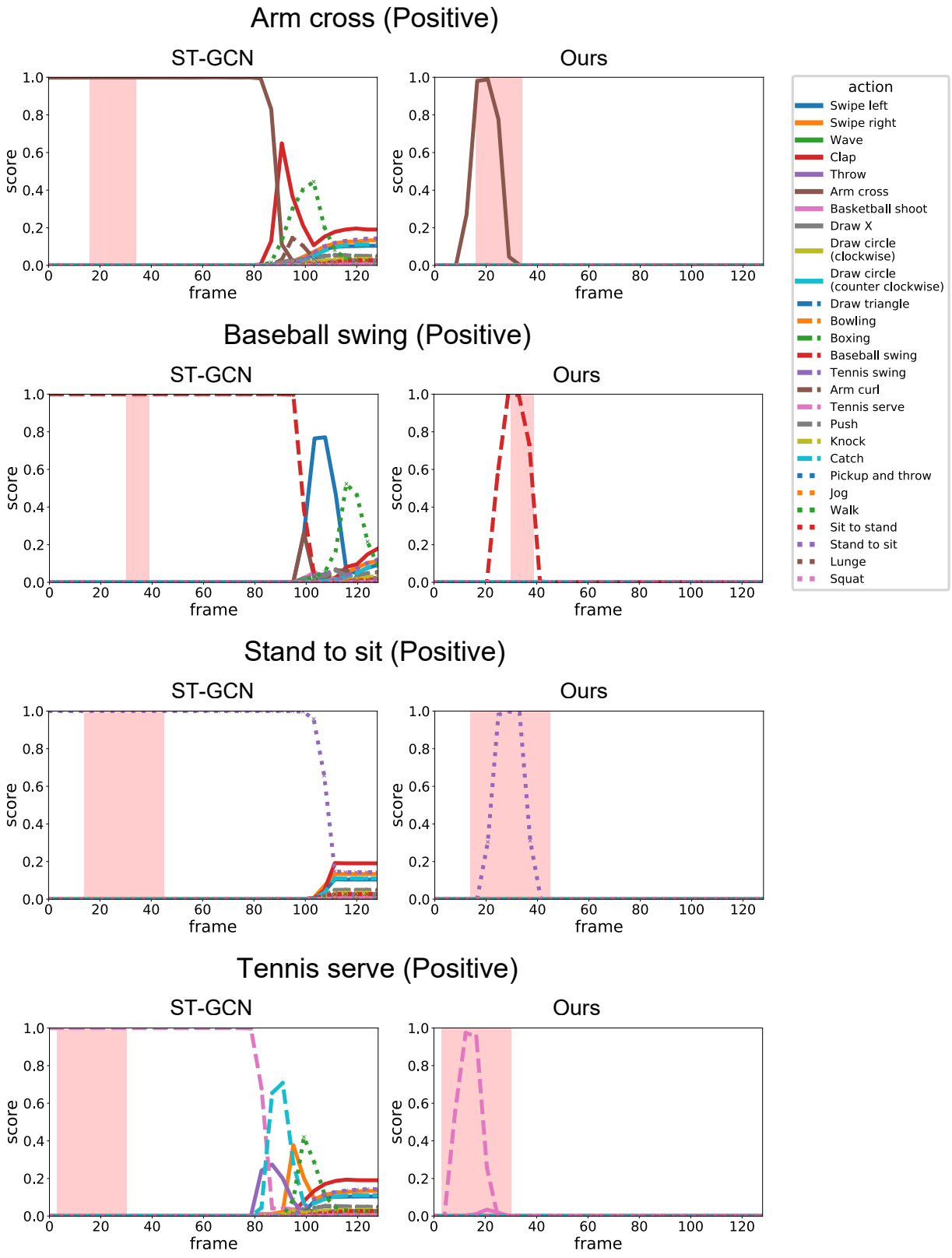


図 5.3 UTD-MHAD データセットに対する提案手法により学習された DNN モデルの推定値

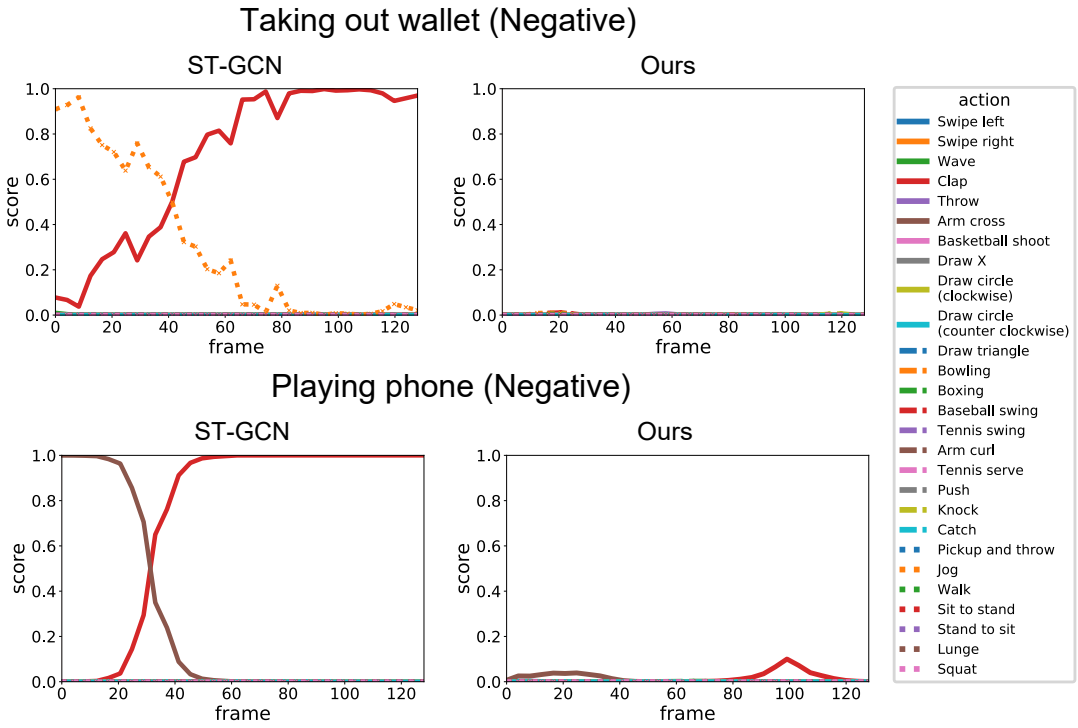


図 5.4　データセットに含まれない行動に対する DNN モデルの推定値の比較

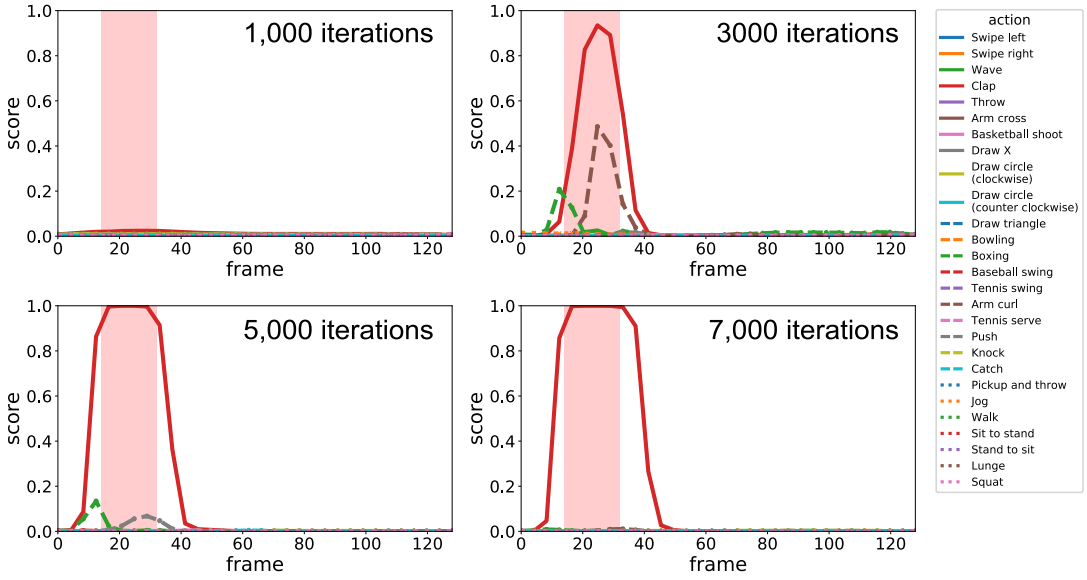
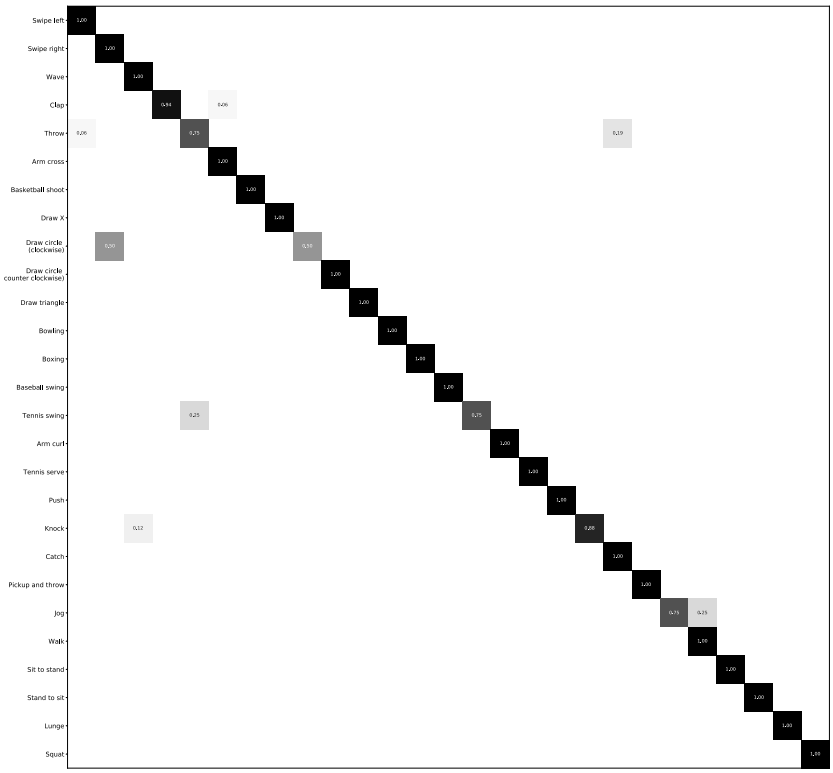


図 5.5　学習の進行に伴う DNN モデルの推定値の変化

(a)



(b)

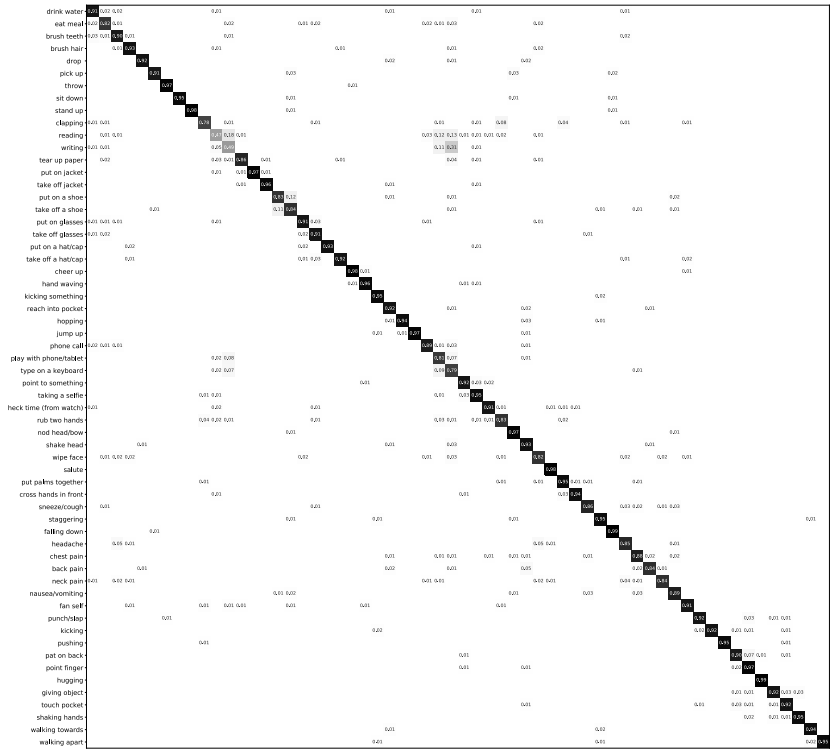


図 5.6 (a)UTD-MHAD および (b)NTU RGB+D データセットに対する混同行列

表 5.2　UTD-MHAD および NTU RGB+D データセットにおける識別精度の比較

Methods	UTD-MHAD	NTU RGB+D	
		CS	CV
ST-GCN [76]	–	81.5	88.3
ST-GCN (Our implementation)	94.2	79.5	87.3
Ours	94.6	79.9	89.8

人物行動ローカライゼーションに関する定量的な評価として、UTD-MHAD データセットに対して mean Average Precision（mAP）の値を Intersection over Union（IoU）の閾値 $0.1 - 0.5$ の範囲で評価した。図 5.7、図 5.8 および表 5.3 は、認識精度の定量的評価結果と UTD-MHAD データセットに対する行動ローカライゼーション結果およびアノテーションの一部を示したものである。これらの図からわかるように、従来手法では人物の行動認識が可能であることが確認できるが、行動ローカライゼーションは困難であることがわかる。これは、ST-GCN の出力においてソフトマックス関数による活性化が行われるため、一連の動作を全体的に評価した上で行動の決定に必要の無い動作に対しても、高い値が推定されることが原因と考えられる。一方、提案手法では行動ローカライゼーションが適切に行われていることが確認でき、両手法は行動識別において同程度の精度が確認されたが、提案手法は行動ローカライゼーションを行う上でより有利であった。この結果はアノテーションの付与が困難なデータに対して共通する特徴量、例えば人物の危険な動作や不自然な挙動のように、定義が曖昧でアノテーションの付与が困難な動作を認識したい際に有利と考えられる。

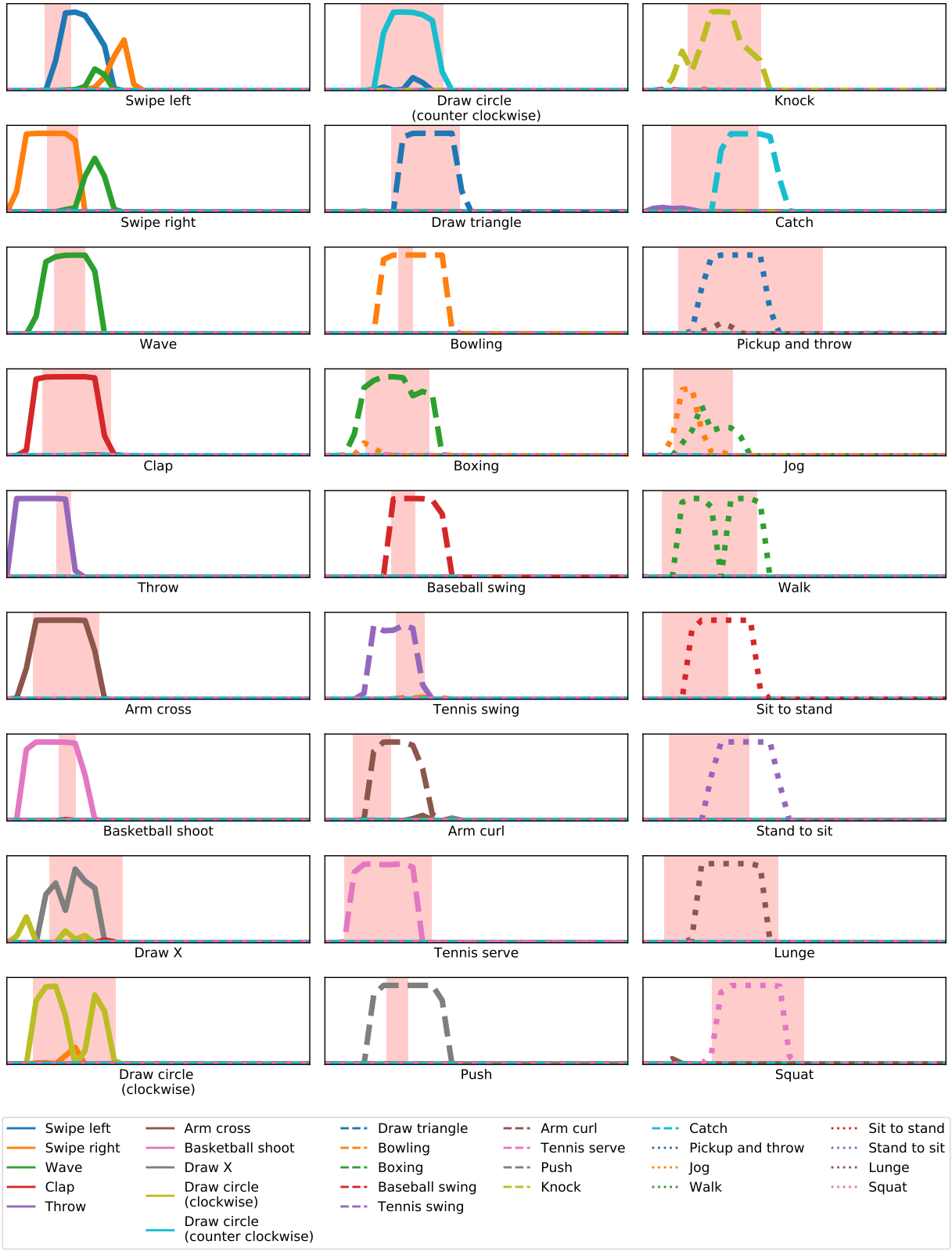


図 5.7　提案手法による UTD データセットに対する人物行動ローカライゼーション結果

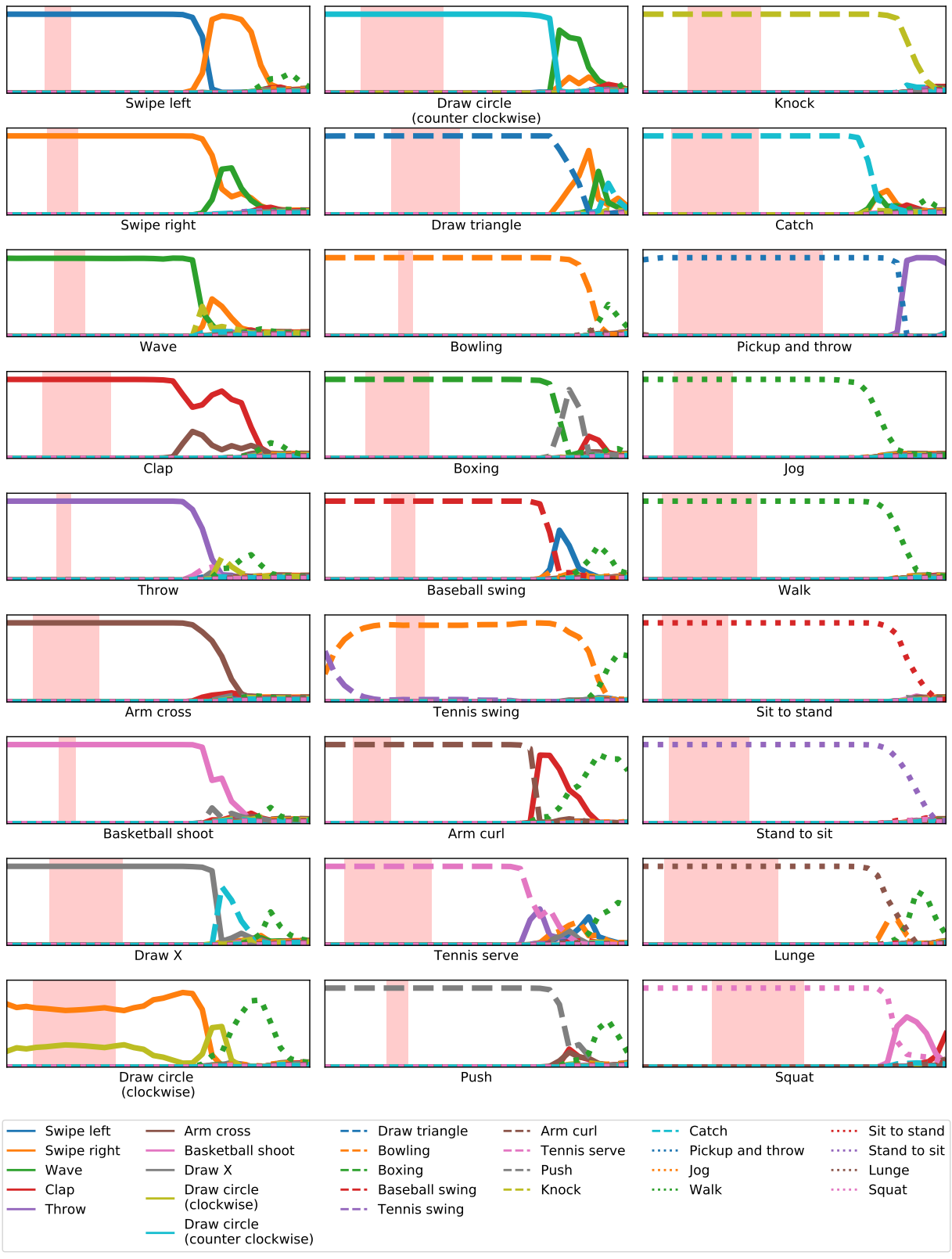


図 5.8 従来手法 [76] による UTD データセットに対する人物行動ローカライゼーション結果

表 5.3　UTD-MHAD データの行動ローカライゼーション結果の比較

	mAP@IoU				
IoU threshold	0.1	0.2	0.3	0.4	0.5
ST-GCN (Our implementation)	24.2	20.5	12.4	4.0	0.7
Ours	72.3	69.5	59.7	44.1	25.0

5.4.2 人物動作からの異常検知

図 5.9 に不自然な歩行動作および自然な歩行動作を含む一連の人物動作に関する時系列データに対し提案手法により学習された DNN モデルから推定された値を示す。不自然な歩行動作を含むデータに対して推定された値は、多くの時点において自然な歩行動作データから推定された値よりも高いことが確認できる。また、不自然な歩行動作を含むデータに対して推定された値の中でも、幾つかの時点に対しては高い値を示す一方で、他の時点に対しては低い値を示した。行動ローカライゼーション結果と併せて考察すると人物の不自然な歩行動作に共通する特徴を含む箇所に対して高い値を推定するように DNN モデルが学習されたと考えられる。以上の結果から、本研究の目的である DNN モデルを用いた多変量時系列データからの異常検知、特に DNN モデルの学習のため異常度の真値を把握することが困難な弱教師あり多変量時系列データからの異常検知の実現可能性が見出された。特に、このような定義が曖昧な行動の検知技術は、原子力関連施設において作業員が危険な行動、または危険な状態に陥っていないか客観的に把握するシステム等に応用できる可能性がある。

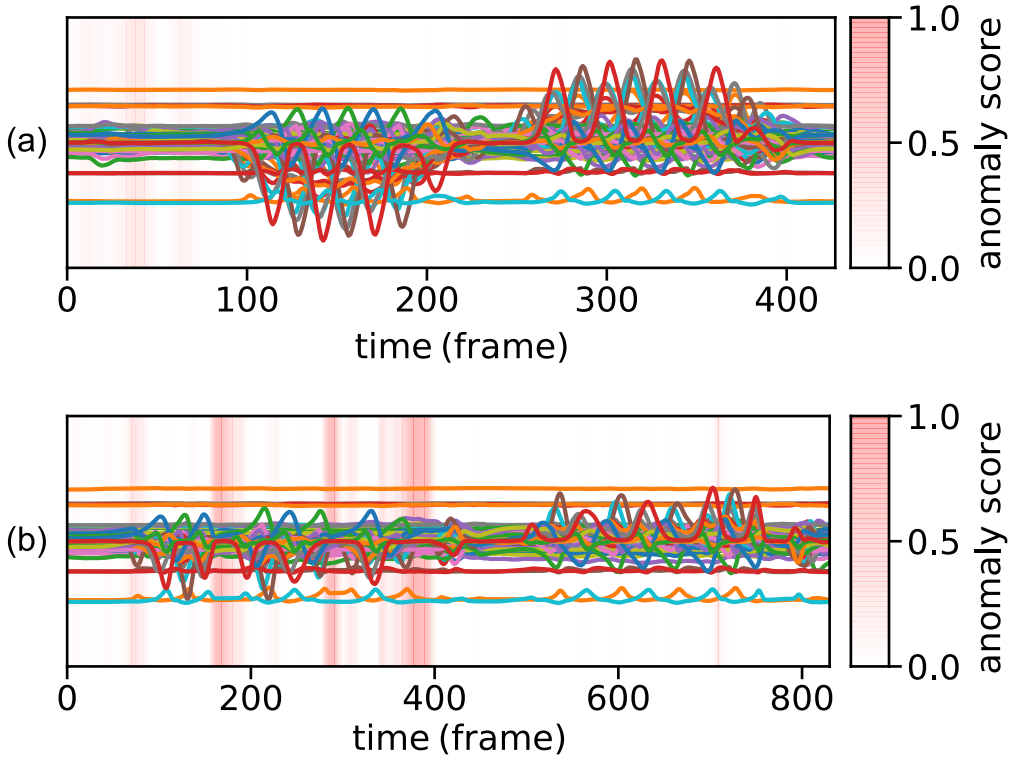


図 5.9　不自然自然な動作を (a) 含まないおよび (b) 含む人物動作データおよびそれらに対する DNN モデルの推定値

5.5 結言

本章では第 2 章および第 3 章で提案した時系列データ解析手法と第 4 章で提案した人物動作特徴量の抽出手法を併せて用いた DNN モデルの人物動作解析への応用とその評価について述べた。特に人物動作解析のため、DNN モデルには人物の関節位置情報に関する時空間的なデータ構造を扱うための ST-GCN モデルを適用した。DNN モデルの構造は従来手法である ST-GCN モデルの構造 [76] に着想を得ているが、DNN モデルの出力形状を時間および認識すべきクラス数についてそれぞれを行および列とする行列形式とすることで、人物行動の識別のみならず、時系列データに含まれるそれぞれの行動を決定づける上で重要な特徴を含む箇所を推定可能とした。局所的なアノテーションの付与されていない多変量時系列データを用いた人物行動ローカライゼーションを実現するために、DNN モデルを第 2 章および第 3 章で提案した手法により学習することで、人物行動の識別と同時にローカライゼーションが可能であった。公開データセットを用いた評価では、提案手法で学習された DNN モデルにより人物の行動識別が可能であり、従来手法と同等の識別精度を保ちながら同時にローカライゼーションが可能であることを確認した。また、不自然動作の検知のため、不自然な挙動が含まれるおよび含まれない人物動作データを用いた実験から、提案手法により学習された DNN モデルを用いて不自然な人物動作を検知可能であることを確認した。以上の結果は、局所的なアノテーションの付与が困難な膨大なデータセットを扱う際や、人物の危険な動作や不自然な挙動といった定義が曖昧な人物動作に関する情報を含むデータセットから特定の動作の識別およびローカライゼーションを行う上で有利であることを示唆するものである。このような定義が曖昧な行動の検知技術は、原子力関連施設において作業員が危険な行動、または危険な状態に陥っていないか客観的に把握するシステム等に応用できる可能性がある。

第 6 章

結論

時系列データからの異常検知技術は、故障診断、状態監視、映像監視等を実現するために重要な技術である。特に原子力関連施設では地震や津波等の天災のみならず、枢要機器の故障や不具合、テロリズム等あらゆる脅威を想定しそれらに対して万全な対策を講じることが求められている。状態監視や映像監視を実現する技術として統計的手法や機械学習手法を応用したセンサデータや映像データを解析する技術が求められており、特に DNN モデルを用いた手法はその表現能力の高さから、多変量時系列データに含まれる異常の複雑な特徴を抽出し検知する上で有利な手法として期待されている。動的機器に搭載されたセンサデータに代表される多変量時系列データから異常を検知可能とすれば、機器の故障検知に応用できる可能性がある。また、監視カメラから日々収集される膨大な映像のような、より高次元な時系列データから異常を検知することができれば、原子力関連施設の安全対策がより強固となることが期待される。

DNN モデルを実環境に適用するためには事前に膨大なデータを用いた学習によりそのパラメータを最適化することが必要である。そのためには取得した時系列データに対し、その異常を決定付ける上で重要な特徴を含む箇所にアノテーションを付与することが必要となるが、データに複雑な特徴が含まれている場合にはその作業は困難であった。そこで本論文では弱教師あり多変量時系列データに潜在する特徴を自動的に抽出し、異常を検知可能な DNN モデルおよびその学習手法を提案した。また、人物動作解析のための新たな特徴抽出手法を確立し、先に確立した時系列データ解析のための DNN モデルと組み合わせることで人物動作解析および異常検知を実現した。

第 2 章では時系列データから異常を検知するための DNN モデルおよびその学習手法について述べた。特に異常を含む時系列データに対して、それらの有無が既知であるが、含まれる箇所や程度が未知であるような弱教師あり時系列データを上手く取り扱うため手法について述べた。学習には、マルチインスタンス学習に着想を得た手法を提案し、特に損失関数を工夫することで時系列データの各時点における異常度の真値を必要とせずに DNN モデルを学習した。実験では、異常を含むまたは含まない時系列データに対して本手法を適用することで、異常を

含むデータにおける異常を含む箇所のみに対して高い値を推定するように DNN モデルの学習が可能であることを確認した。さらに外れ値を含む波形データを用いた実験では、外れ値の大きさと DNN モデルの推定値との間に相関が確認され、異常の検知のみならず定量の可能性が示唆された。また、実システムへの適用可能性を評価するため、軸受の振動データ解析に適用したところ、異常を含む振動データに対して高い値が推定され、本手法の実システムへの応用可能性が見出された。本手法は、特徴量エンジニアリングの作業が少なく、実応用の上で有利と考えられる。

第 3 章では前章で提案した弱教師あり時系列データからの異常の識別を目的とした DNN モデルの学習手法の改良について述べた。異常の識別のため、DNN モデルの構造を検討し、第 2 章で確立した DNN モデルの学習手法の改良として損失関数を一般化した。本手法の実システムへの応用可能性を確認するために軸受の振動データ解析に適用したところ、異常を含むデータに対してそれぞれの波形に特徴的な箇所に対して高い推定値が確認され、適切に異常が識別された。この結果は本手法により学習された DNN モデルが軸受の故障診断のような実システムへ適用可能であることを示唆するものである。

第 4 章では以上の時系列データ解析手法を映像解析に適用するための映像からの人物動作特徴量の抽出手法ついて述べた。特に、広角画像の歪みに頑健な人物姿勢の認識のため、画像から人物の姿勢および適切な補正パラメータを推定可能な新たな DNN モデルおよびその学習手法を提案した。また、推定された 2 次元的な人物姿勢から 3 次元的な人物姿勢を復元する手法と FCNN モデルを用いた人物位置推定手法を提案し、人物動作特徴量を抽出可能とした。実験では方位角 0° から 70° の範囲内で良好な認識精度で人物姿勢の認識が可能であることを確認した。提案手法は、赤外線式モーションキャプチャ機器を用いた手法と比べて、より近距離および広角での人物姿勢認識が可能であり、映像監視への適用可能性が示唆された。また、本手法を人物の自然な歩行動作および不自然な歩行動作を模擬した映像データに適用することで、それぞれの動作に関する特徴量を真値と比較して良好な認識精度で推定可能であった。特に不自然な歩行動作において体の向きや視線方向の変動、移動速度の変動に対して特徴的な傾向を把握することが可能であった。

第 5 章では確立した時系列データの解析手法と人物動作特徴量の抽出手法を組み合わせた人物動作解析への応用について述べた。特に人物動作解析のため、DNN モデルには人物の関節位置に関する時空間的な情報を扱う GCN モデルを採用した。本 DNN モデルの構造は従来手法である ST-GCN モデルに着想を得ているが、出力形状を行列形式とし、行および列をそれぞれ時間および認識すべき行動クラス数とすることで、行動の識別のみならず行動ローカライゼーションを可能とした。公開データセットを用いた評価では、人物行動の識別が可能であることを確認し、識別精度を従来手法と同等に保ちながら行動ローカライゼーションが可能であることを確認した。これらの結果は、局所的なアノテーションの付与が困難な膨大なデータセットを扱う際や、人物の危険な動作や不自然な挙動といった曖昧な動作を含むデータセット

から特定の動作の識別およびローカライゼーションが可能であることを示唆するものである。また、定義が曖昧な行動の検知技術は原子力関連施設において作業員が危険な状態に陥っていないか客観的に把握するシステム等に応用できる可能性がある。

以上の結果から、本研究の目的である DNN モデルを用いた時系列データからの異常検知、特に異常度の真値を把握することが困難な弱教師あり多変量時系列データを用いて DNN モデルを学習し、異常の検知が可能であることが示された。以上で確立した技術は振動データや映像データを用いた実験から、動的機器の状態監視や映像監視等、様々な分野への応用が可能と考えられる。

謝辞

本研究は筆者が東京大学大学院および東京都立産業技術研究センターにおいて行ったものです。本論文の執筆にあたり、多大なご指導、ご鞭撻を賜りました指導教員である出町和之准教授に心より感謝申し上げます。また、同研究室の笠原直人教授をはじめ皆様に厚く御礼申し上げます。

研究活動を遂行するにあたり多大なご協力を頂いた東京都立産業技術研究センター情報技術グループの皆様に心から感謝いたします。

最後に家族に感謝いたします。

本研究の一部は日本学術振興会科学研究費助成事業 JP19K05324, JP19K20310 の助成を受けて行われました。また、研究成果の一部には東京都立産業技術研究センターにおける基盤研究事業の成果が含まれます。

令和 2 年 12 月 1 日

三木 大輔

参考文献

[1] Yukiya Amano. *The Fukushima Daiichi Accident Report by the Director General*. 2015.

[2] 東京電力福島第一原子力発電所事故調査・検証委員会，政府事故調 最終報告書.

[3] 東京電力福島原子力発電所事故調査委員会，国会事故調 報告書.

[4] 福島原発事故独立検証委員会，調査・検証報告書.

[5] IAEA. *Defence in Depth in Nuclear Safety INSAG-10*. 1996.

[6] 福島第一原子力発電所事故を踏まえた核セキュリティ上の課題への対応. http://www.aec.go.jp/jicst/NC/senmon/bougo/siryo/bougo25/siryo1.pdf.

[7] Pat O'Donnell. Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part I. *IEEE Transactions on Industry Applications*, Vol. IA-21, No. 4, pp. 853–864, 1985.

[8] Pat O'Donnell. Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part II. *IEEE Transactions on Industry Applications*, Vol. IA-21, No. 4, pp. 865–872, 1985.

[9] Pat O'Donnell. Report of Large Motor Reliability Survey of Industrial and Commercial Installations: Part 3. *IEEE Transactions on Industry Applications*, Vol. IA-23, No. 1, pp. 153–158, 1987.

[10] Olivier Janssens, Viktor Slavkovikj, Bram Vervisch, Kurt Stockman, Mia Loccufier, Steven Verstockt, Rik Van de Walle, and Sofie Van Hoecke. Convolutional Neural Network Based Fault Detection for Rotating Machinery. *Journal of Sound and Vibration*, Vol. 377, pp. 331–345, 2016.

[11] Honghu Pan, Xingxi He, Sai Tang, and Fanming Meng. An improved bearing fault diagnosis method using one-dimensional CNN and LSTM. *Journal of Mechanical Engineering*, Vol. 64, No. 7-8, pp. 443–452, 2018.

[12] IAEA. *Nuclear Security Recommendations on Physical Protection of Nuclear Material and Nuclear Facilities*. 2011.

[13] IAEA. *Nuclear Security Recommendations on Radioactive Material and Associated Facilities*. 2011.

[14] IAEA. *Nuclear security recommendations on nuclear and other radioactive material out of regulatory control*. 2011.

[15] 内閣府原子力委員会，核セキュリティの確保に対する基本的考え方について. http://www.aec.go.jp/jicst/NC/pressrelease/files/20110809/honbun.pdf.

[16] 外務省ホームページ，原子力の平和的利用　核セキュリティ. https://www.mofa.go.jp/mofaj/dns/n_s_ne/page22_000968.html.

[17] Nobuyuki Otsu. Towards flexible and intelligent vision systems - From thresholding to CHLAC. *Proceedings of the 9th IAPR Conference on Machine Vision Applications, MVA 2005*, pp. 430–439, 2005.

[18] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

[19] Lei Wang, Du Q. Huynh, and Piotr Koniusz. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. *IEEE Transactions on Image Processing*, Vol. 29, No. 1, pp. 15–28, 2020.

[20] Chhavi Dhiman and Dinesh Kumar Vishwakarma. A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*, Vol. 77, No. June 2018, pp. 21–45, 2019.

[21] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. Efficient Human Pose Estimation from Single Depth Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, pp. 2821–2840, 2013.

[22] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

[23] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp. 172–186, 2021.

[24] Shi Chen, Kazuyuki Demachi, Tomoyuki Fujita, and Yutaro Nakashima. Insider Malicious Behaviors Detection and Prediction Technology for Nuclear Security. *E-Journal of Advanced Maintenance*, Vol. 9, pp. 66–71, 2017.

[25] Shi Chen and Kazuyuki Demachi. Proposal of an insider sabotage detection method for nuclear security using deep learning. *Journal of Nuclear Science and Technology*, Vol. 56, No. 7, pp. 599–607, 2019.

[26] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Vol. 5, No. 4, pp. 115–133, 1943.

[27] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, Vol. 65, No. 6, pp. 386–408, 1958.

[28] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2146–2153, 2009.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. 2016.

[30] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, Vol. 323, No. 6088, pp. 533–536, 1986.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, pp. 448–456, 2015.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, Vol. 15, No. 1, pp. 1929–1958, 2014.

[34] Chiyuan Zhang, Benjamin Recht, Samy Bengio, Moritz Hardt, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[35] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, Vol. 89, No. 1-2, pp. 31–71, 1997.

[36] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support Vector Machines for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems 15*, pp. 577–584, 2003.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015*

IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034, 2015.

[38] Case Western Reserve University Bearing Data Center. [Online]. http://csegroup.case.edu/bearingdatacenter/home.

[39] Hai Qiu, Jay Lee, Jing Lin, and Gang Yu. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of Sound and Vibration*, Vol. 289, No. 4-5, pp. 1066–1090, 2006.

[40] Levent Eren, Turker Ince, and Serkan Kiranyaz. A Generic Intelligent Bearing Fault Diagnosis System Using Compact Adaptive 1D CNN Classifier. *Journal of Signal Processing Systems*, Vol. 91, No. 2, pp. 179–189, 2019.

[41] Ran Zhang, Zhen Peng, Lifeng Wu, Beibei Yao, and Yong Guan. Fault Diagnosis from Raw Sensor Data Using Deep Neural Networks Considering Temporal Coherence. *Sensors*, Vol. 17, No. 3, p. 549, 2017.

[42] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *Advances in neural information processing systems*, pp. 1799–1807, 2014.

[43] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing ConvNets for Human Pose Estimation in Videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1913–1921, 2015.

[44] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 1, pp. 44–58, 2006.

[45] Shih-en Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732, 2016.

[46] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 483–499. 2016.

[47] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302–1310, 2017.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, Vol. 60, No. 6, pp. 84–90, 2017.

[49] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In *2016 IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), pp. 991–1000, 2016.

[50] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *2017 International Conference on 3D Vision (3DV)*, pp. 506–516, 2017.

[51] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-Shot Multi-person 3D Pose Estimation from Monocular RGB. In *2018 International Conference on 3D Vision (3DV)*, No. 335545, pp. 120–130, 2018.

[52] Grégory Rogez and Cordelia Schmid. Image-Based Synthesis for Deep 3D Human Pose Estimation. *International Journal of Computer Vision*, Vol. 126, No. 9, pp. 993–1008, 2018.

[53] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5689–5698, 2017.

[54] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2659–2668, 2017.

[55] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex C. Kot. Feature Boosting Network For 3D Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 2, pp. 494–501, 2020.

[56] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7745–7754, 2019.

[57] Weipeng Xu, Avishek Chatterjee, Michael Zollhofer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 Cap 2 : Real-time Mobile 3D Mo tion Capture with a Cap -mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 25, No. 5, pp. 2093–2101, 2019.

[58] Umar Iqbal, Andreas Doering, Hashim Yasin, Björn Krüger, Andreas Weber, and Juergen Gall. A dual-source approach for 3D human pose estimation from single images. *Computer Vision and Image Understanding*, Vol. 172, pp. 37–49, 2018.

[59] Sam Johnson and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *Procedings of the British Machine Vision*

Conference 2010, pp. 12.1–12.11, 2010.

[60] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

[61] Unity software. [Online]. https://unity.com.

[62] CMU Graphics Lab Motion Capture Database. [Online]. http://mocap.cs.cmu.edu/.

[63] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pp. 1385–1392, 2011.

[64] Evaluation tool for the LSP dataset. [Online]. http://human-pose.mpi-inf.mpg.de/results/lsp/evalLSP.zip.

[65] Microsoft, Kinect for Windows v2. https://developer.microsoft.com/en-us/windows/kinect/.

[66] RICOH, THETA S. hhttps://theta360.com/en/about/theta/s.html.

[67] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, 2012.

[68] Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2466–2472, 2013.

[69] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, Vol. 25, No. 1, pp. 24–38, 2014.

[70] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 588–595, 2014.

[71] Junwu Weng, Chaoqun Weng, and Junsong Yuan. Spatio-Temporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for Skeleton-Based Action Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–454, 2017.

[72] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, Vol. 68, pp. 346–362, 2017.

[73] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, 2015.

[74] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C. Kot. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing*, Vol. 27, No. 4, pp. 1586–1599, 2018.

[75] Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 12, pp. 3007–3021, 2018.

[76] Sijie Yan and Dahua Xiong, Yuanjun and Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 7444–7452, 2018.

[77] Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, and Jian Yang. Spatio-temporal graph convolution for skeleton based action recognition. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 3482–3489, 2018.

[78] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3590–3598, 2019.

[79] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1227–1236, 2019.

[80] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12018–12027, 2019.

[81] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7904–7913, 2019.

[82] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015.

[83] Jian-fang Hu, Wei-shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 11, pp. 2186–2200, 2017.

[84] Amir Shahroudy, Jun Liu, Tian-tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, 2016.