

修士論文

深層学習による
表情・体動の時系列情報を考慮した
新生児睡眠・覚醒状態自動推定

令和5年度修了

三重大学大学院 工学研究科 情報工学専攻
ヒューマンコンピュータインタラクション研究室

伊藤 由樹

はじめに

在胎週数が 37 週未満で生まれた早産児は、身体機能が未熟であることが多く、NICU (Neonatal Intensive Care Unit) と呼ばれる新生児特定集中治療室に入院する。NICU 内では早産児インキュベータと呼ばれる保育器内で医療監視下におかれ、昼夜様々な医療行為を受ける。NICU には看護師や様々な医療機器があり、それらが光や騒音を発する。このような環境は、早産児にとって胎内とは全く異なる環境であり、NICU 内の環境が新生児の神経学的発達に及ぼす影響についての調査が必要とされている。特に、早産児にとっての NICU 内の環境は、睡眠リズムであるサーカディアンリズムを形成する上でも重要であり、新生児の睡眠状態と NICU 内の環境について様々な調査が行われている。

現在、NICU における新生児の睡眠状態の判定は、看護師による目視での判定、もしくは新生児の足に装着するアクチグラフと呼ばれる装置を用いて行われている。しかし、これらの方法は新生児の状態を目視で判定する看護師や装置を装着する新生児にとって負担となる。これらの負担を無くすため、本研究では動画を用いて新生児の睡眠・覚醒状態を自動で推定する手法を提案する。新生児の睡眠・覚醒状態は、日本の NICU で広く用いられている Brazelton の分類に則り、6 段階で評価する。

本研究では動画中の新生児の睡眠・覚醒状態を自動推定する手法を提案する。はじめに、NICU で撮影された動画から、1 分ごとの身体全体・顔領域（手作業で指定または物体検出モデルによる自動検出）のクリップに分割する。そして、クリップ単位で 3DCNN や TimeSformer などの深層学習モデルを用いて睡眠・覚醒状態を推定する。最後に学習した身体全体・顔領域の動きから睡眠・覚醒状態を推定する 2 つの深層学習モデルの推定結果とその時系列変化を考慮した推定結果統合手法を構築する。

九州大学病院で撮影された新生児 8 名、昼と夜にそれぞれ約 2 時間撮影された動画を 16 本使用した実験の結果、クリップごとの推定においては、3DCNN が TimeSformer よりも身体全体の動画では Kappa スコアが 0.127、顔領域の動画では Kappa スコアが 0.175 高く、3DCNN の方が有効的であることがわかった。また、身体全体・顔領域の動画から推定される各クラスの確率分布とその時系列変化を考慮した平滑化を行うことで、身体全体の動画の時は Kappa スコア 0.082、顔領域の動画では Kappa スコア 0.074 向上するこ

とがわかった。上記の結果から、3DCNN と各クラスの確率を考慮した平滑化を用いる手法について、手作業により抽出した顔領域と、YOLO により自動抽出した顔領域を用いた場合での推定精度の比較を行った。その結果、手作業で抽出した場合での Kappa スコア 0.727 に対し、YOLO では Kappa スコア 0.684 がえられた。今後は身体と顔の動画から出力されたそれぞれの結果をよりよく組み合わせる手法を検討する必要がある。

目次

はじめに	i
第 1 章 緒言	1
1.1 研究背景	1
1.2 関連研究	2
1.3 先行研究	3
1.4 研究目的	4
第 2 章 準備	5
2.1 NBAS(Neonatal Behavioral Assessment Scale)	5
2.2 データセット	7
第 3 章 提案手法	9
3.1 睡眠・覚醒状態のクラス分類	9
3.2 顔領域の検出	12
3.3 時系列平滑化	15
3.4 身体全体・顔領域の推定結果の統合	16
第 4 章 実験	18
4.1 データセット	18
4.2 評価方法	19
4.3 実験 1: モデル選定	21
4.4 実験 2: 時系列平滑化	25
4.5 実験 3: 顔領域の自動抽出	28
第 5 章 結言	33
5.1 本研究のまとめ	33
5.2 今後の課題	33

目次	iv
付録 A 付録	34
A.1 プログラムの詳細	34
謝辞	35

第 1 章

緒言

1.1 研究背景

早産などにより管理や治療が必要な新生児は、NICU (Neonatal Intensive Care Unit) に入院し、特別な医療を受ける。NICU では、新生児は医学的な監視のもと、新生児用保育器の中で昼夜を問わず様々な医療ケアを受ける。しかし、NICU は看護師の会話や医療機器の作動によって騒音や光が発せられる環境となっており、これが満期産児に比べて新生児にとっては過剰な刺激となり、その結果として新生児の睡眠リズムであるサーカディアンリズムの形成を妨げる可能性がある [1,2]。そのため、新生児の神経学的発達にとってよりよい環境を整備することを目的として、NICU 内の環境調査が行われている [3]。NICU 内の環境調査をする上で、新生児の睡眠・覚醒状態の判定は不可欠である。日本の NICU では、新生児の睡眠状態を表す方法として、睡眠・覚醒状態を 6 つの状態に分類した Brazelton の評価尺度 [4] が広く用いられている。従来、新生児の睡眠・覚醒状態は、看護師が目視やバイタルデータによって判定するか、新生児の足に装着するアクチグラフと呼ばれる装置などを用いて判定されてきたが、これらの方法では判定を行う看護師や装置を装着する新生児への負担が大きい [3]。そこで、本研究では新生児の睡眠・覚醒状態を非接触で自動判定する手法を提案する。

1.2 関連研究

EEG (Electroencephalography: 脳波) を用いて新生児の睡眠覚醒状態を分類した研究がある [5–7]。EEG とは、頭皮に電極を付けて脳の活動を測定し、脳波として記録する技術である。この装置を用いることで、新生児の脳活動を観察し、睡眠覚醒状態を高い精度で分類することができる。しかし、新生児は皮膚が弱いため、装着による圧迫感や不快感があるため脳波計を装着するのは負担が大きい。また、装置の取り外しやメンテナンスのために頻繁な操作が必要となり、新生児へのストレスが増大する可能性がある。したがって、新生児に負担をかける EEG を用いた手法は、本研究では望ましくないと考えている。

Cabon らの研究では、NICU における新生児の睡眠状態を音声・体動・眼球運動に基づいて推定する手法を提案した [8]。この研究では、動画から抽出した特徴に基づいて、新生児の睡眠状態を Prechtl のルール [9] に基づいて表された 5 つの状態に半自動で分類している。Prechtl のルール [9] では新生児の睡眠状態を、Quiet Sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA), Active Alert (AA) の 5 段階で表す。推定の結果では、QA と AA においてそれぞれ 93.5 %, 99.0 % と高い正答率を得ている。また、彼らは新生児の睡眠状態の推定には、音声と体動が重要な要素であることを示唆した。しかし、新生児の目の状態を手動で判定する必要があり自動判定とはなっていない。また、日本の NICU においては新生児はインキュベータと呼ばれる保育器に入っているため、インキュベータの外から撮影された動画を用いざるを得ず、新生児の音声を取得することは困難である。

Awais らの研究では、赤外線カメラと RGB カメラで撮影した動画に基づいて、新生児を睡眠状態と覚醒状態の 2 つのクラスに分類した。彼らは顔領域の動画を使用し、DCNN (Deep Convolutional Neural Network) [10] と SVM (Support Vector Machine) [11] のハイブリッド・モデルを提案した。この研究では、RGB ビデオを使用した場合、より高い正解率 (93.8) と F1 スコア (0.93) が報告された。また、睡眠覚醒状態の分類には表情情報が必要であることが示された。

日本の NICU では睡眠覚醒状態を Brazelton のルール [4] を用いて 6 つのクラスに分類している。そのため、本研究では動画から自動で新生児の睡眠覚醒状態を推定する手法を提案する。

1.3 先行研究

先行研究として服部らの研究 [12] と盛田らの研究 [13] がある。服部らの研究 [12] では Optical Flow により動画中の新生児の体動を検出し、得られた動きベクトルの値から動きの大きさと方向を算出し、動きベクトルの L2-ノルムを用いるヒストグラム、方向ごとの L2-ノルムを用いるヒストグラム、L2 ノルムと方向を用いるヒストグラムの 3 手法でヒストグラムを生成し、手法ごとに、セル単位で生成したヒストグラムを結合した特徴ベクトルを SVM (Support Vector Machine) [11] を用いて分類し、精度を比較した。実験の結果、L2 ノルムと方向を用いるヒストグラムを生成する手法で、最大 0.765 の macro-F1 スコアが得られたが、実用に十分な精度を得られることができなかった。

盛田らの研究 [13] では手動で切り出した顔領域の画像から HOG 特徴量 [14] を抽出し、重み付きサポートベクターマシン (w-SVM) を用いて新生児の睡眠覚醒状態を分類する手法を提案した。この手法は、最大 0.732 の micro-F1 スコアを達成し、睡眠覚醒状態分類における表情情報の重要性を示唆した。しかし、この研究では、顔の領域は手動で抽出されている。

また、どちらの研究も学習とテストには同じ新生児の異なる時刻のビデオが使用された。そこで本研究では、顔領域の自動抽出を行い、異なる新生児のデータを学習とテストに用いる。

1.4 研究目的

本研究の目的は、インキュベーターの外から撮影された新生児の動画から Brazelton の睡眠覚醒状態分類 [4] の高精度推定を可能とすることである。これにより、NICU 内の環境調査を目的とした新生児の睡眠・覚醒状態判定に伴う看護師や新生児の負担をなくすることができる。

本研究では、動画中の新生児の睡眠・覚醒状態を自動推定する手法を提案する。はじめに、NICU で撮影された動画から、1 分ごとの身体全体・顔領域（手作業で指定または物体検出モデルによる自動検出）のクリップに分割する。そして、クリップ単位で 3DCNN [15] や TimeSformer [16] などの深層学習モデルを用いて睡眠・覚醒状態を推定する。最後に学習した身体全体・顔領域の動きから睡眠・覚醒状態を推定する 2 つの深層学習モデルの推定結果とその時系列変化を考慮した推定結果統合手法を構築する。

第2章

準備

2.1 NBAS(Neonatal Behavioral Assessment Scale)

Brazelton らは、新生児の神経行動発達を評価するための基準となる NBAS を提案した [4]。本研究では、Brazelton の新生児行動評価の睡眠・覚醒状態を用いる。これは、体の動き、呼吸パターン、目の開閉、顔の動き、刺激に対する反応性に基づいて、睡眠・覚醒状態を 6 つの段階に定義するものである。Brazelton の評価尺度における 6 つの状態は以下のように定義されている。

- State1(睡眠)
深い睡眠状態である。体動はなく、稀に蹴り上げる。呼吸は深く規則的である。瞼は閉じていて眼球運動もなく、表情も変わらない。強い刺激にのみ反応する。
- State2 (睡眠)
浅い睡眠状態である。体動はわずかで不規則に運動し、呼吸は規則的である。瞼は閉じているが、急速な眼球運動をとめない、表情は時に微笑み、ぐずり泣く。外的・内的な刺激に反応する。
- State3(覚醒)
うとうとした状態である。体動は変化的で呼吸は不規則である。瞼は重く開眼・閉眼し、表情は稀に動く。反応は遅い。
- State4(覚醒)
微睡の状態である。体動は少なく、呼吸は規則的である。開眼し注視しており表情は目覚めた状態である。刺激に注意を向ける。
- State5(覚醒)
覚醒状態である。体動は活発で呼吸は不規則である。開眼しているが大きく見

開かず，活発な顔面運動がある．内的刺激に敏感に反応する．

- State6(覚醒)

啼泣状態である．体動は活発で呼吸は乱れる．目は開眼もしくは固く閉眼しており，表情は泣き顔である．不快な刺激に敏感に反応する．

このように，State1, 2 が睡眠状態を表し，State3, 4, 5, 6 が覚醒状態を表す．これらの State は，NICU 内の環境調査のみならず，実際の医療現場で新生児に対して医療行為を施す際の指標にも使われている．

2.2 データセット

本研究では，九州大学病院で撮影された NICU 内の 8 名の新生児の動画，合計 16 本をデータセットとして用いる．なお，本研究の実施については九州大学病院の倫理委員会で承認済みである．これらの動画は，新生児 1 人につき昼と夜にそれぞれ 2 時間程度撮影されたものである．動画は，新生児の全身が映るよう約 30cm の距離をとり，インキュベータの外から撮影されている．これらの動画には，NICU で働いている看護師が Brazelton の分類に基づいて判定した State が正解データとして付与されている．State は 1 分ごとにつけられており，本研究では動画データを State が付けられている 1 分単位の動画に分割し，クリップと呼ぶこととする．動画は毎秒 30 フレームで撮影されており，1 つのクリップには約 1,800 フレームの画像が含まれる．図 2.1 および図 2.2 にフレーム画像の例を示す．なお，フレーム画像の例は個人情報保護のため目元を隠してある．表 2.1 に各 State のクリップ数を示す．データセットには合計 2,287 個のクリップが含まれており，全ての State は同一の看護師によって付与されたものである．また，動画中にカメラの動きや看護師の介入が含まれているクリップは除外している．

表 2.1: データセットにおける各 State のクリップ数

State	1	2	3	4	5	6
クリップ数	763	1,205	199	52	46	22



図 2.1: フレーム画像の例 被験者#1 昼

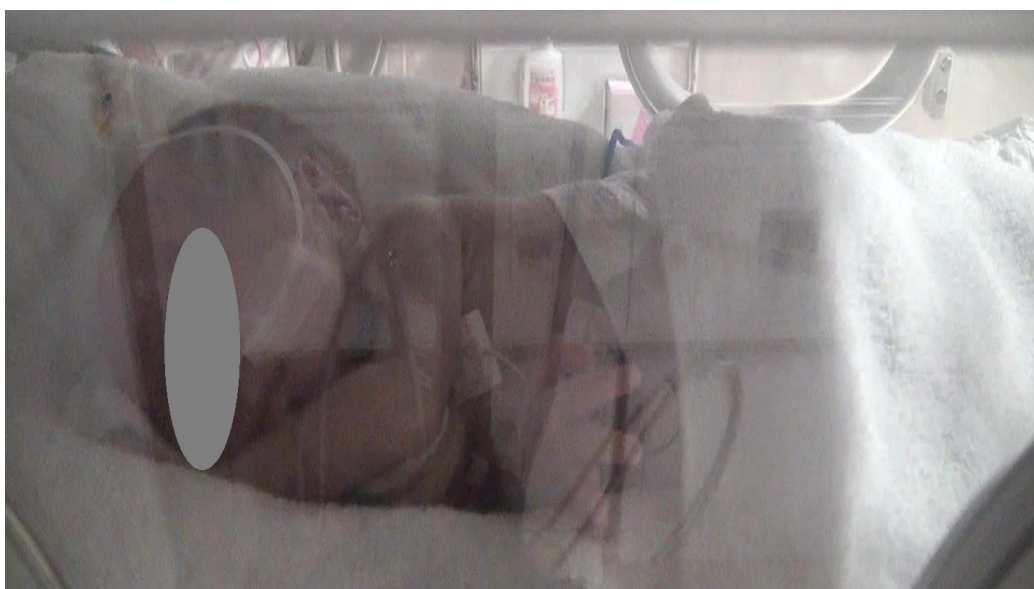


図 2.2: フレーム画像の例 被験者#1 夜

第 3 章

提案手法

3.1 睡眠・覚醒状態のクラス分類

睡眠・覚醒状態のクラス分類では, 新生児の動画から Brazelton の評価尺度で定義された 6 つの状態に自動で推定する手法を提案する. 本実験では深層学習モデルである 3DCNN と Timesformer を用いる.

3.1.1 3D Convolution Neural Network

3D Convolution Neural Network(3DCNN) [15] は, 現在, 画像認識の分野で流行中の 2DCNN を 3 次元方向に拡張したもので, 動画の時系列情報を考慮した学習を行うことができるモデルである. 2DCNN は図 3.1 のように画像の二次元配列に対して, フィルタをかけることにより畳み込みを行うのに対して, 3DCNN では図 3.2 のように 3 次元配列にフィルタをかけることにより, 畳み込みを行う. 今回は 3DCNN の一つである 3DResNet [17] を利用する. 3DResNet は CNN と同様に画像認識の分野で使われている ResNet を 3 次元方向に拡張したものである. ResNet [18] とはある層で求める最適な出力を学習するのではなく, 層の入力を参照した残差関数を学習することにより層を深くすることができるモデルで, 画像認識で高い精度を出している.

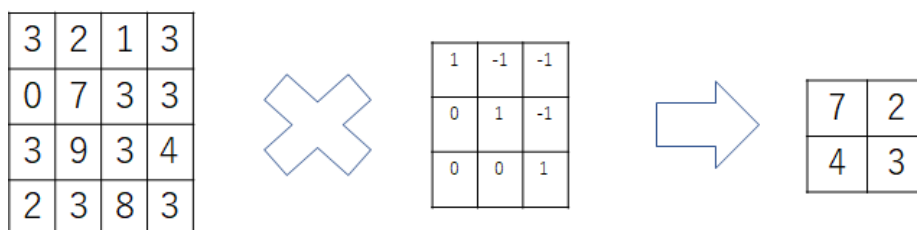


図 3.1: 2d convolution

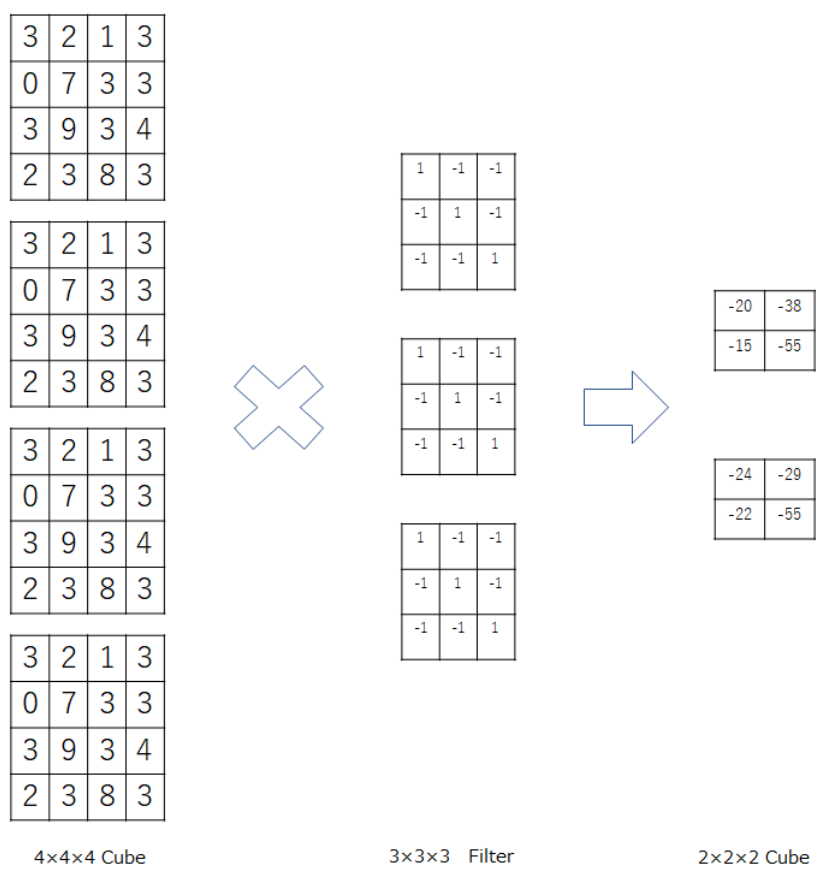


図 3.2: 3d convolution

3.1.2 TimeSformer

TimeSformer [16] は、時間系列データを処理するために設計された Transformer ベースのモデルである。Transformer [19] は自然言語処理や Vision Transformer [20] として画像認識などのタスクで成功を収めている。Transformer を動画の時系列情報を考慮した学習を行うことができるように改良したモデルが TimeSformer である。TimeSformer では、動画データの各フレームに対して、時間軸アテンションと空間軸アテンションと呼ばれる 2 つの異なるアテンションメカニズムを用いて情報を収集する。時間軸アテンションは、同じ場所にある別フレーム（図 3.3 の青と緑）のパッチとのみで比較し、動画内の時間的なパターンや動きを捉える。一方、空間軸アテンションは、同じフレーム内（図 3.3 の赤）の異なる領域の間での関連性を考慮し、各フレーム内の空間的な特徴を捉える。これらのアテンションメカニズムを組み合わせることで、TimeSformer は動画データから時間的なパターンと空間的な特徴を効果的に抽出し、動画分類タスクにおいて高い性能を発揮する。その結果、動画認識のための大規模なデータセットの Kinetics-400 と Kinetics-600 では、どちらも 3DCNN よりも高速で精度も良い結果を出している。

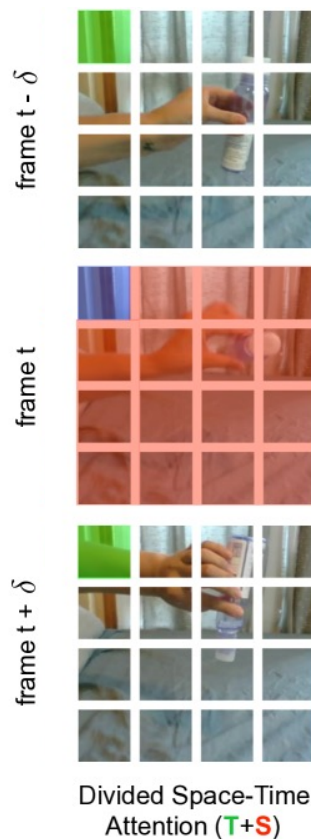


図 3.3: Divided Space-Time Attention [16]

3.2 顔領域の検出

睡眠覚醒状態の分類では表情による情報が重要であるため、顔領域を自動で推定する手法を提案する。本実験では物体検出モデルである YOLO と SSD を用いる。

3.2.1 YOLO

YOLO (You Only Look Once) [21] は、物体検出のためのディープラーニングアーキテクチャである。従来の手法では、画像を複数の領域に分割し、各領域内の物体の有無を判定する。しかし、YOLO は画像全体を一度に処理し、同時に物体の位置とクラスを予測する。そのため処理が非常に高速で、リアルタイムの物体検出に適している。YOLO のアーキテクチャは、図 3.4 のようにニューラルネットワークを使って画像を一定の大きさのグリッドに分割する。各グリッドセルは物体の存在を予測し、物体のバウンディングボックスの位置と対応する物体クラスの確率を出力する。これにより物体を高い精度で検出することができる。本研究では最新の YOLO-v7 [22,23] を使用する。

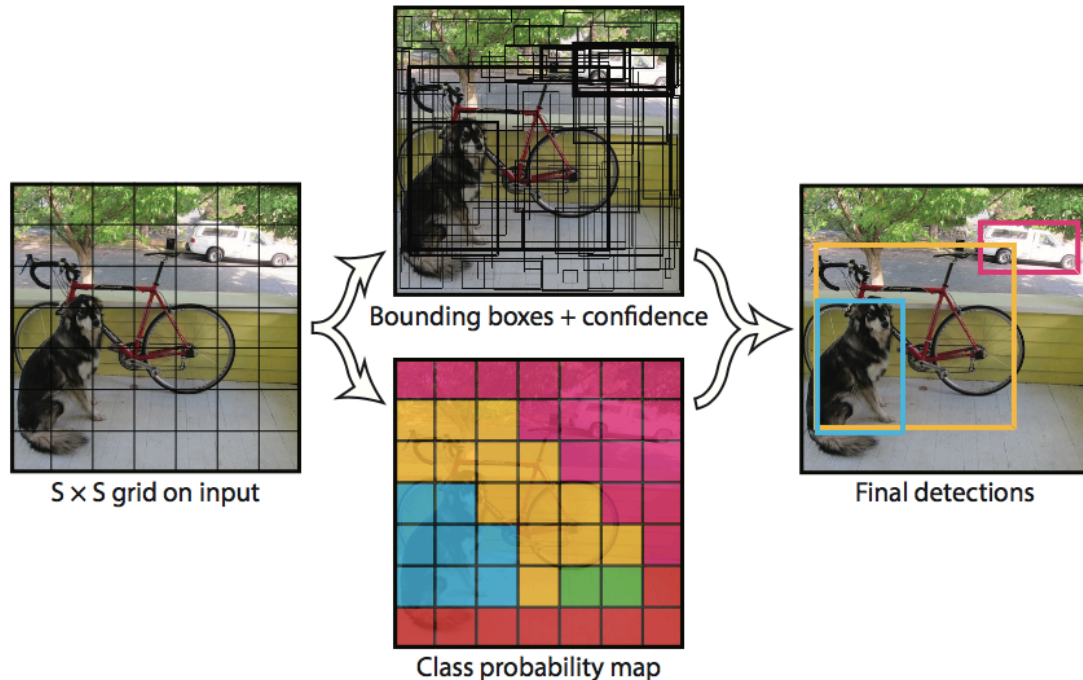


図 3.4: YOLO を用いた物体検出 [21]

3.2.2 SSD

SSD (Single Shot multibox Detector) [24] は物体検出のためのディープラーニングアーキテクチャである。SSD は、VGG16 [25] をベースネットワークとして用い、デフォルトボックスを用いて物体の位置を推定する。具体的には、図 3.5 のようにベースネットワークで畳み込みを行った後の特徴マップ上でデフォルトボックスを用いて物体検出を行う。そのため、畳み込み層によって大きさや形状の異なる特徴マップを用いて、複数または微細な物体を捉えることが可能である。

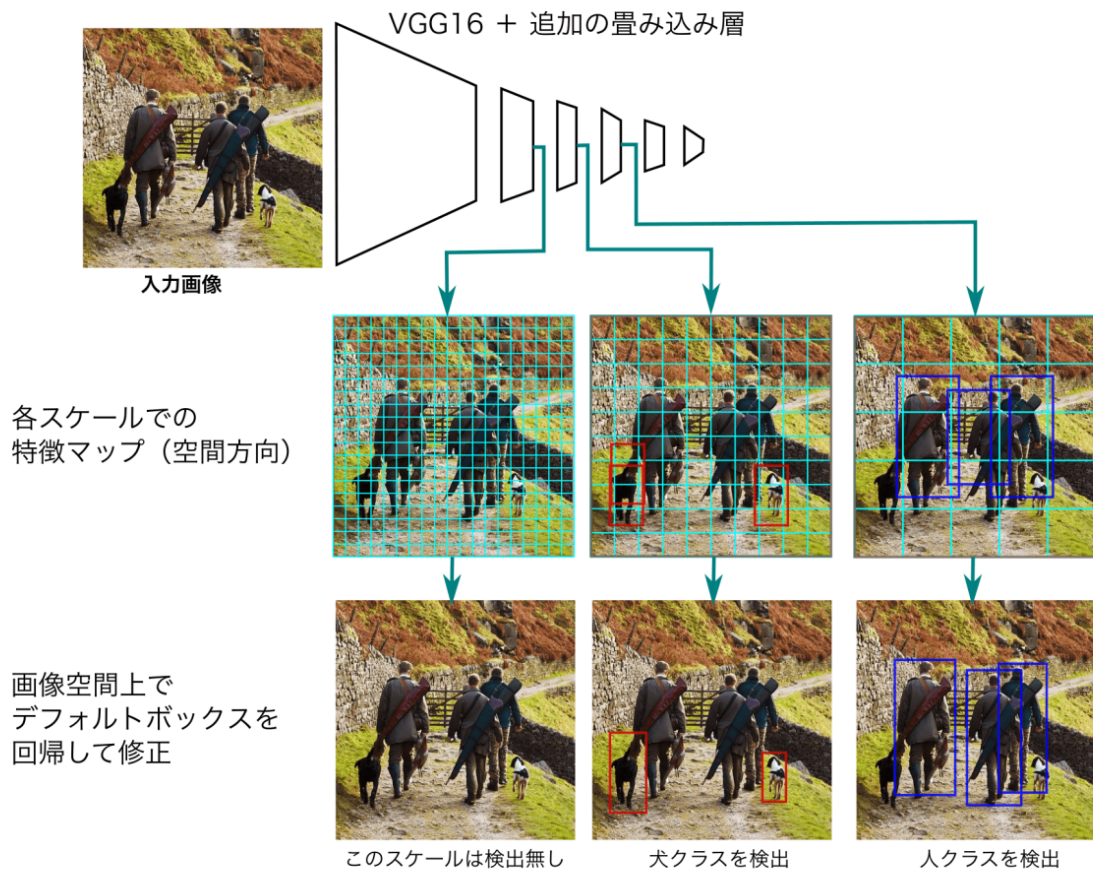


図 3.5: SSD を用いた物体検出 [24]

3.2.3 切り取り領域の正規化

物体検出によって得られる顔領域の座標は、フレームごとに異なる。そのため、その座標をそのまま顔領域の切り出しに使うと、同じクリップ内でも顔の位置がずれてしまうことがある。これを防ぐためには、切り出しに用いる座標をすべてのクリップで共通にする必要がある。この共通座標は以下の式で計算される。クリップ内の全ての出力顔領域の左上隅の x 座標と y 座標のリストを $P_{x,c}$ と $P_{y,c}$ とし、右下隅の x 座標と y 座標のリストを $Q_{x,c}$ と $Q_{y,c}$ とする。切り出す顔領域の左上隅の座標を $p_{x,c}$ と $p_{y,c}$ 、右下隅の座標を $q_{x,c}$ と $q_{y,c}$ とすると、計算は以下のようになる。

$$p_{x,c} = \min P_{x,c} \quad (3.1)$$

$$p_{y,c} = \min P_{y,c} \quad (3.2)$$

$$q_{x,c} = \max Q_{x,c} \quad (3.3)$$

$$q_{y,c} = \max Q_{y,c} \quad (3.4)$$

これにより、同一クリップ内で顔の位置がずれることなく、正確な顔領域を切り取ることができる。このとき、外れ値が含まれていると、検出座標がずれることがある。これを考慮し、同一クリップ内の画像について IoU を算出し、IoU が 0.3 未満であれば外れ値の座標とみなす。その後、外れ値を除去し、式 (3.1), (3.2), (3.3), (3.4) を算出する。顔が検出できないフレームがあっても、クリップ内の別のフレームで検出できれば、そのフレームの座標を用いて顔領域を切り出す。

3.3 時系列平滑化

新生児の睡眠覚醒状態は急激な状態変化を起こさないため、任意の時点の前後の State 推定値が推定誤差の検出・修正に役立つ可能性がある。各クリップの State 推定後、時系列平滑化により、以下の式で急激な状態変化の低減を試みる。まず、式 (3.5) は $2c + 1$ クリップの平均 State 値 f'_t を計算する。処理中の t クリップとその前の c クリップと次の c クリップ (実験では $c = 2$ を用いた)。

$$f'_t = \frac{\sum_{i=t-c}^{t+c} f_i}{2c + 1} \quad (3.5)$$

次に、State 変化を平滑化するために、推定状態と平滑化状態の差の絶対値が閾値 α (本実験では $\alpha = 0.5$ を採用) を超えたとき、式 (3.6) は推定値 f_t を四捨五入した値 f'_t に更新する。そうでなければ、推定値は更新されない。

$$f_t \leftarrow \begin{cases} \text{round}(f'_t) & |f_t - f'_t| > \alpha \\ f_t & \text{Otherwise} \end{cases} \quad (3.6)$$

3.4 身体全体・顔領域の推定結果の統合

3.4.1 特徴量結合

学習済みの2つの深層学習モデルの畳み込み層は、身体全体・顔領域のクリップの特徴を取得し、これら2つの特徴を足し合わせて1つの特徴に結合する。この結合された特徴量をSVMまたはMLP(Multi layer Perceptron)に入力し、これら2つの特徴に基づく分類を行う。このようにして、身体全体の情報と顔の情報が効果的に統合し、身体全体の動きと表情の情報を同時に捉えることが可能になる。

3.4.2 State 平均

身体全体・顔領域のクリップのそれぞれから深層学習モデルによる推定と時系列平滑化を行い、推定値の平均を新たな分類結果とする。

3.4.3 確率分布による重み付け

深層学習モデルを用いた State 推定では，あるクラスに属する確率を求め，これを推定の信頼度として用いることができる．以下の式は， t^{th} クリップにおける統合された State を，身体全体 f_b と顔領域の推定 f_f を，それぞれ対応する重み w_b と w_f を用いて加重平均したものである．

$$F(t) = \frac{w_b(t) f_b(t) + w_f(t) f_f(t)}{w_b(t) + w_f(t)} \quad (3.7)$$

ここで， f_b と f_f は時系列平滑化後の推定値である．重み w_b と w_f も同様に計算する．

$$w_*(t) = \frac{\sum_{i=t-c}^{t+c} p_*(i)}{2c+1} \quad (3.8)$$

ここで， $p_*(i)$ は，身体全体 (b) または顔領域クリップ (f) の i 番目のクリップにおける確率値である．具体的には，推定結果の組み合わせに対する信頼度を反映するために，信頼度の高い確率には大きな重みを，信頼度の低い確率には小さな重みを与える．このように重み付けを行うことで，身体全体と顔領域の結果の信頼性を適切に考慮しつつ，より信頼性の高い全体の推定結果を得ることができる．

第 4 章

実験

Scikit-learn 1.2.2, PyTorch 2.0.1, Python 3.11 を用いて提案手法を実装し, AMD Ryzen TR Pro 5965WX, 128GB RAM, NVIDIA RTX4090 GPU を用いて以下の実験を行った.

実験 1 では学習に用いるモデルの選定を行う. 実験 2 では時系列情報を考慮する手法と顔領域と身体全体の動画から出力された結果を組み合わせる手法を検討する. 実験 3 では顔領域の自動抽出の手法を検討する.

4.1 データセット

表 4.1 に新生児と State の組み合わせごとのビデオクリップ数を示す. 睡眠覚醒状態の分類には, Brazelton の睡眠覚醒状態をラベルとして用いる. しかし, 使用するデータセットには, State 6 のクリップを持つ新生児のデータが 2 人分しかないため, State 6 を State 5 に統合する. 実験では 8 人の新生児のデータを使用し, そのうち 7 人を学習用, 1 人をテスト用とし, 交差検証で行う.

表 4.1: 新生児ごとのクリップ数

State	No.2	No.3	No.4	No.5	No.6	No.7	No.8	No.9
1	75	68	69	108	145	107	98	93
2	149	159	127	127	100	174	174	195
3	6	47	66	15	34	9	12	10
4	8	11	9	21	2	0	1	0
5	7	1	11	15	10	0	2	0
6	19	0	0	3	0	0	0	0

4.2 評価方法

以下の実験では, Accuracy, Macro-F1, Kappa スコア, 二乗平均平方根誤差 (RMSE), IoU で提案手法の性能を評価する.

4.2.1 Accuracy

Accuracy とは, 正解数の割合を評価する指標である. 本研究では 5 クラス分類を行い, 正しく推定されたクリップの総数 TP , 誤分類されたクリップの総数 TN , 評価されたクリップの総数 N を用いて Accuracy を算出する.

$$Accuracy = \frac{TP + TN}{N} \quad (4.1)$$

4.2.2 Macro-F1

F 値は 2 クラス分類性能の評価に広く用いられている. 公平な評価を行うため, 5 クラス分類の性能評価には Macro-F1 (式 (4.2)) を用いた. これは, State のクリップ数に偏りがあるため, データ数の少ない State クラスを全て間違っても無視されることが無いようにするためである.

$$F_m = \frac{1}{S} \sum_{i=1}^S f_i \quad (4.2)$$

ここで S はクラス数 ($S = 5$), F_i は i 番目の State の F 値である.

4.2.3 重み付け Kappa スコア

Accuracy と Macro-F1 は順序分類性能を過小評価する可能性がある. 正解 State と予測 State の一致度を評価するために, 重み付け Kappa スコア [26] を採用する. これは, 混同行列の各セルに 2 次重み $w_{i,j} = (i - j)^2$ を定義するもので, 正解 State i と予測 State j が離れているときの誤差を重視する.

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad (4.3)$$

ここで $O_{i,j}$ は観察された一致度を表し, 評価されたクリップの総数に対するセル (i, j) 内のクリップ数の比として計算できる. 偶然の一致度 ($E_{i,j}$ と表記) を計算するには, 正解が State i にある確率と予測が State j にある確率を掛け合わせる. Kappa スコアは -1 から 1 の範囲をとる. 1 に近い値は完全な一致を示し, 0 に近い値は一致の度合いが偶然に近

いことを示す。負の値は、一致の度合いが偶然よりも悪いことを示す。これは正解ラベルと予測ラベルの一致度を評価し、State 間の距離が大きい間違いはより重いペナルティを受ける。具体的には、正解が State 1 の場合、予測が State 2 で間違っているよりも State 5 で間違っている方がスコアは悪くなる。

4.2.4 二乗平均平方根誤差 (RMSE)

二乗平均平方根誤差 (RMSE) は、回帰分析の評価指標の 1 つである。これは、測定値と予測値の差の 2 乗平均平方根を計算することによって、予測誤差の大きさを評価する。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2} \quad (4.4)$$

ここで、 y_i, \hat{y}_i, n はそれぞれ正解値、予測値、データ総数である。

4.2.5 IoU

IoU (Intersection over Union) は、画像処理や機械学習の分野で用いられる評価指標の一つである。主に物体検出やセグメンテーションなどのタスクの評価に使用され、IoU は予測領域と実際の領域の重なり度の度合いを評価するための指標である。

$$IoU = \frac{A_{overlap}}{A_{union}} \quad (4.5)$$

ここで $A_{overlap}$ は 2 つの領域が重なる面積である、 A_{union} は 2 つの領域の合計面積を示す。つまり、IoU は重なり部分の面積を領域全体の面積で割って算出される。

4.3 実験 1: モデル選定

4.3.1 実験条件

新生児の睡眠覚醒状態の推定を行うためのモデルを以下の 5 つの実験で比較する．

- 実験 1-0 : Optical Flow + SVM と身体全体の動画 (先行手法)
- 実験 1-1 : 3DCNN と身体全体の動画
- 実験 1-2 : TimeSformer と身体全体の動画
- 実験 1-3 : 3DCNN と顔領域の動画
- 実験 1-4 : TimeSformer と顔領域の動画

学習パラメータは表 4.2 のように設定する．GPU メモリサイズの制限により，図 4.1 のように 1 クリップ 1800 フレームあたり 8 フレームに間引いて入力フレームとする．顔領域の動画は手動で切り取った動画を使用する．

テストデータは No2, No4, No5 の 3 パターンで交差検証で行う．

表 4.2: 学習器のパラメータ

最適化アルゴリズム	SGD
loss 関数	Cross Entropy Loss
バッチサイズ	8
学習率	0.001
モーメンタム	0.9
エポック数	5000

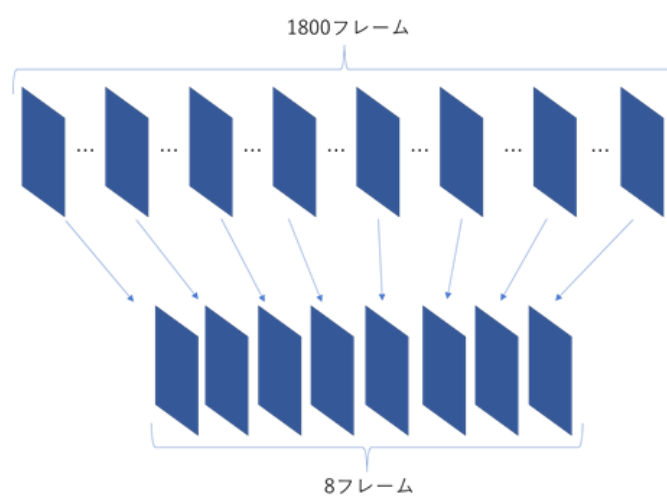


図 4.1: 入力フレーム

表 4.3: 実験 1 の分類精度

実験	手法	身体全体の動画	顔領域の動画	Accuracy	Macro-F1	Kaapa
1-0	先行手法	✓		0.408	0.166	0.045
1-1	3D CNN	✓		0.494	0.328	0.490
1-2	TimeSformer	✓		0.497	0.292	0.363
1-3	3D CNN		✓	0.541	0.374	0.641
1-4	TimeSformer		✓	0.462	0.379	0.466

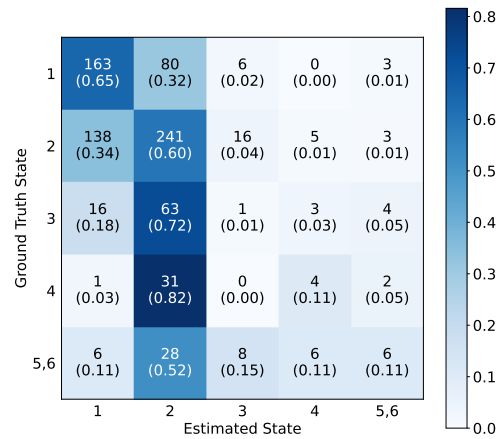
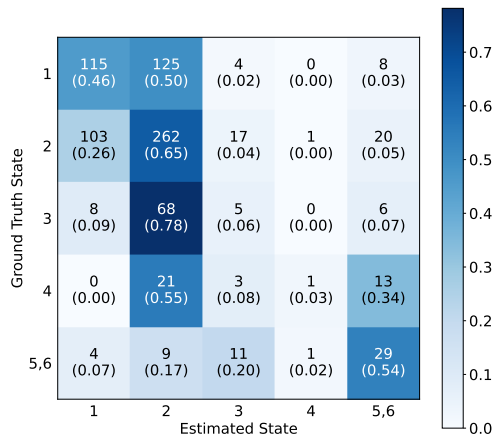


図 4.2: 実験 1-1(3D CNN+身体全体) 図 4.3: 実験 1-2(TimesFormer+身体全体)

4.3.2 実験結果と考察

実験結果を表 4.3 に示す。また各実験結果の混同行列を図 4.2～4.5 に示す。

先行手法の SVM を用いた実験 1-0 と 3DCNN, TimesFormer を比較すると Accuracy, Macro-F1, Kappa スコアすべての評価指数で大幅に向上している。この結果は, 3DCNN, TimesFormer のどちらも Optical Flow + SVM よりも時系列情報をより良く扱うことを表している。実験 1-1 と 1-2 を比較すると Accuracy ではわずかに TimeSformer のほうが高いが Macro-F1, Kappa スコアともに 3DCNN を用いたほうが高いことがわかる。

同様に実験 1-3 と 1-4 を比較すると Macro-F1 ではわずかに TimeSformer のほうが高いが Accuracy, Kappa スコアともに 3DCNN を用いたほうが高いことがわかる。このため, 新生児の睡眠覚醒状態の自動推定では 3DCNN が有効的であることがわかる。

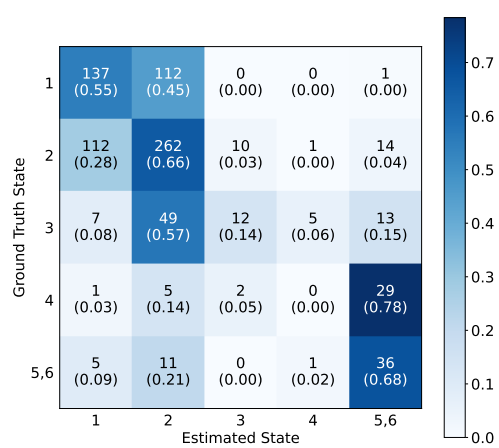


図 4.4: 実験 1-3(3D CNN+顔領域)

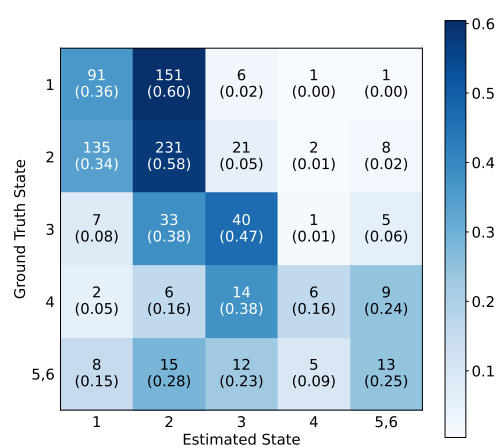


図 4.5: 実験 1-4(TimesFormer+顔領域)

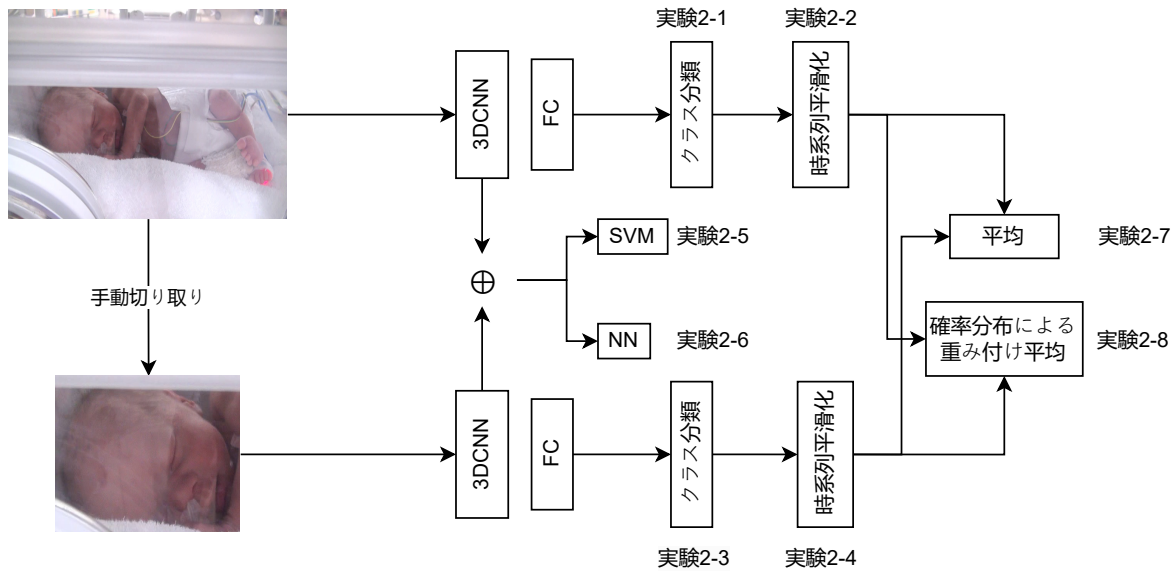


図 4.6: 実験 2 の概要

4.4 実験 2: 時系列平滑化

実験 1 で精度が優れていた 3DCNN を用いて実験 2 を行う。

4.4.1 実験条件

テストデータは No29 の 8 パターンで交差検証で学習モデルはすべて 3D CNN で行う。実験 2 の概要を図 4.6 に示す。

- 実験 2-1: 身体全体の動画
- 実験 2-2: 身体全体の動画と時系列平滑化
- 実験 2-3: 顔領域の動画
- 実験 2-4: 顔領域の動画と時系列平滑化
- 実験 2-5: 身体全体の動画と顔領域の動画の特徴抽出をし SVM を用いて学習
- 実験 2-6: 身体全体の動画と顔領域の動画の特徴抽出をし NN を用いて学習
- 実験 2-7: 身体全体の動画と顔領域の動画に時系列平滑化を行った平均
- 実験 2-8: 身体全体の動画と顔領域の動画に時系列平滑化を行い確率分布を考慮した推定

表 4.4: 実験 2 の結果

実験	身体全体の動画	顔領域の動画	平滑化	組み合わせ	Accuracy	Macro-F1	Kaapa	RMSE
2-1	✓				0.515	0.364	0.435	-
2-2	✓		✓		0.440	0.376	0.517	0.740
2-3		✓			0.570	0.430	0.550	-
2-4		✓	✓		0.599	0.450	0.624	0.679
2-5	✓	✓		SVM	0.534	0.260	0.237	-
2-6	✓	✓		NN	0.521	0.309	0.358	-
2-7	✓	✓	✓	Average	0.603	0.433	0.610	0.690
2-8	✓	✓	✓	Weighting	0.611	0.458	0.623	0.611

4.4.2 実験結果と考察

表 4.4 は、実験 2 の様々な組み合わせにおける分類性能の結果である。3 つの指標 (Accuracy, Macro-F1, Kappa スコア) はすべての手法で計算し、RMSE は連続値を推定できる時系列平滑化との組み合わせでのみ計算した。

3DCNN と身体全体または顔領域の動画を用いた結果、時系列平滑化により、身体全体の動画を用いた場合 (実験 2-1 と実験 2-2) には Kappa スコアで 0.082, 顔領域の動画を用いた場合 (実験 2-3 と実験 2-4) には Kappa スコアで 0.074 向上した。そのため、時系列平滑化により誤分類が減少し、睡眠覚醒状態を効果的に分類できることがわかった。

3DCNN から出力した身体全体と顔領域の特徴を連結した実験 (実験 2-5 と実験 2-6) では、NN 分類器は Macro-F1 と Kappa スコアで SVM を上回った。しかし、時系列平滑化の有無にかかわらず、顔領域の動画を用いた 3DCNN を上回ることとはできなかった。身体全体と顔領域の動画の分類結果を組み合わせる場合 (実験 2-7 と実験 2-8), 結果を平均化するよりも重み付けする方が、すべての指標でより高いスコアが得られることがわかった。

時系列平滑化による性能向上のさらなる分析として、図 4.7 と図 4.8 に、3DCNN と顔領域の映像を用いた分類結果と時系列平滑化を適用した混同行列を示す。これらの結果から、時系列平滑化は対角線付近により多くの推定結果を配置することに寄与していることが確認でき、時系列平滑化を行うことで、推定結果と正解値の一致度が向上し、一致しない場合でも Kappa スコアが向上することが確認された。

時系列平滑化と確率重み付けを行った 3DCNN を用いて、身体全体と顔領域の動画から

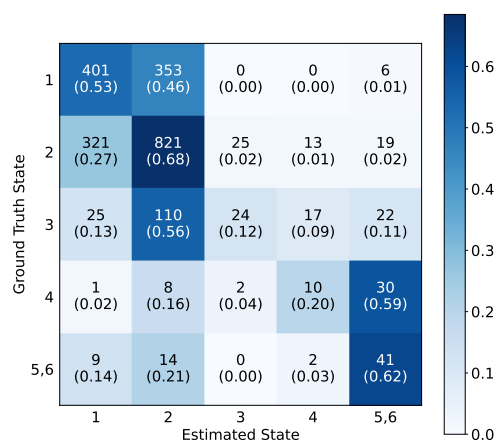


図 4.7: 実験 2-3(顔領域)

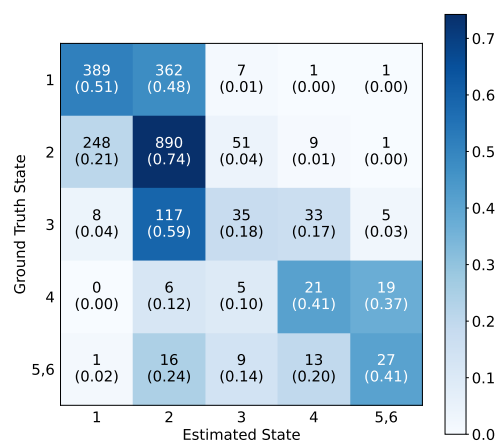


図 4.8: 実験 2-4(顔領域+時系列平滑化)

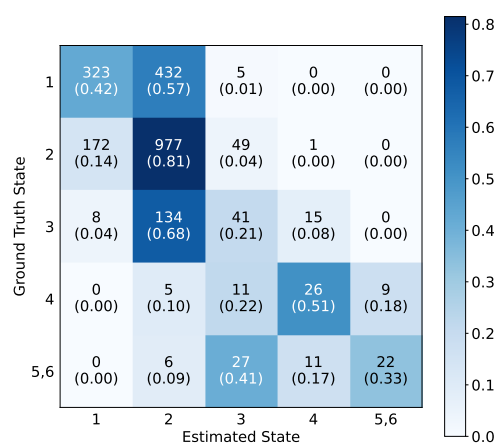


図 4.9: 実験 2-8(身体全体+顔領域+時系列平滑化+確率分布)

推定を行った結果(実験 2-8)の混同行列を図 4.9 に示す. 表 4.4 に示すように, RMSE 以外の数値性能は実験 2-4 と非常に近く, 最も低い RMSE(0.611) を達成しており, 誤分類の誤差が非常に小さいことがわかる.

以上の結果から, 実験 2-8 は, 新生児の睡眠覚醒状態を高い精度で推定でき, 分類結果の信頼性が従来手法よりも向上することがわかった.

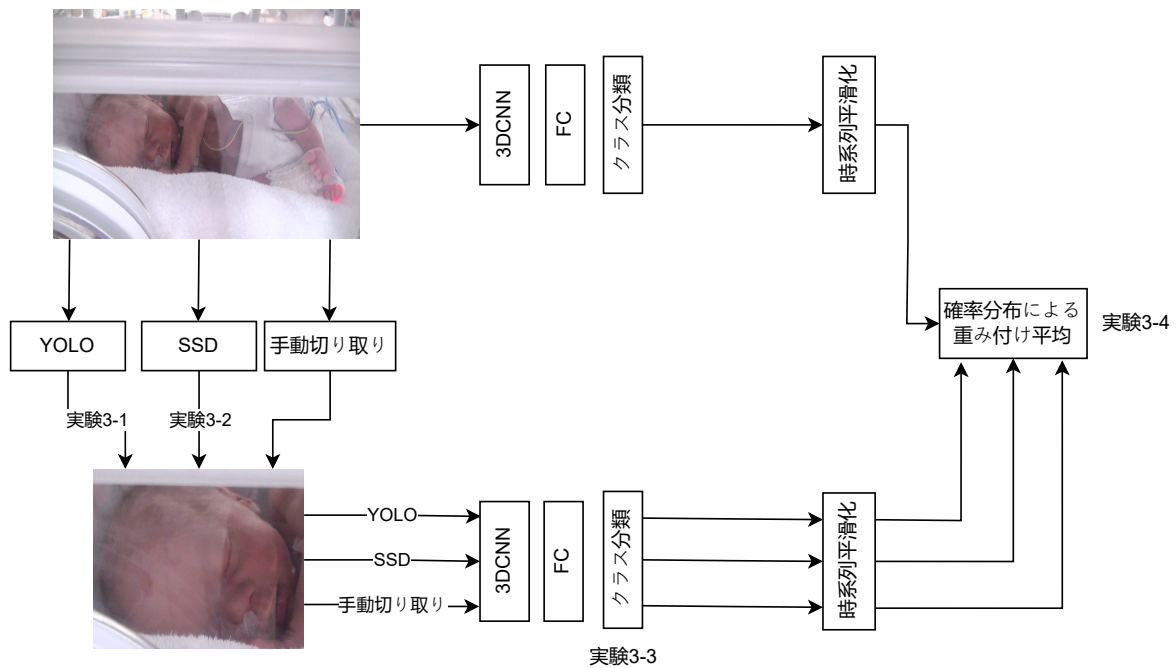


図 4.10: 実験 3 の概要

4.5 実験 3: 顔領域の自動抽出

4.5.1 実験条件

テストデータは No2, No4, No5 の 3 パターンで交差検証で行う。実験 3 の概要を図 4.10 に示す。

- 実験 3-1: YOLO を用いた顔領域の自動抽出
- 実験 3-2: SSD を用いた顔領域の自動抽出
- 実験 3-3: 顔領域の自動抽出と 3DCNN を用いた睡眠覚醒状態の推定
- 実験 3-4: 顔領域の自動抽出と身体全体による結果を用いて平滑化を行い確率分布を考慮した推定

表 4.5: 顔領域検出の実験結果 (IoU)

	YOLO(実験 3-1)	SSD(実験 3-2)
No2	0.845	0.731
No4	0.826	0.799
No5	0.753	0.639
平均	0.808	0.723



図 4.11: YOLO からの出力結果

4.5.2 実験結果と考察

表 4.9 は、実験 3-1, 3-2 における新生児の顔領域の IoU を YOLO と SSD で比較した結果である。すべての被験者において、SSD よりも YOLO の方が IoU が高い。YOLO から出力された顔領域の実際の画像を図 4.11 に示す。

表 4.6 は、実験 3-1, 3-2 における YOLO と SSD の出力結果に対して、それぞれ切り取り領域の正規化を行っても顔領域が出力されなかったクリップ数を示している。

この結果は、YOLO が SSD よりも顔領域を検出できなかったクリップが少ないことを示している。表 4.7 は、実験 3-3 による、手動切り取り、YOLO、SSD の 3 つの異なる方法で切り出した顔領域の動画から、3DCNN を用いて新生児の睡眠覚醒状態を分類したときの Kappa スコアである。

Kappa スコアにおいて YOLO は手動切り取りに比べて精度が若干低下している。一方、

表 4.6: 切り取れなかったクリップ数

クリップ数		YOLO(実験 3-1)	SSD(実験 3-2)
No2	264	0	19
No4	282	2	13
No5	289	5	2

表 4.7: 実験 3-3 の結果 (Kappa スコア)

	YOLO	SSD	手動切り取り
No2	0.465	0.505	0.510
No4	0.682	0.667	0.696
No5	0.619	0.625	0.685
平均	0.589	0.599	0.631

SSD は No.2 と No.5 で手動切り取りより精度が向上している。これは SSD が顔検出ができずに切り取ることができなかったクリップが多いためである。

表 4.8 は、実験 3-4 における実験 2 で行った時系列平滑化と確率分布を考慮した推定と実験 3 で行った顔領域の自動抽出を組み合わせた結果の比較である。YOLO と手動切り取りを比較すると Kappa スコアで 0.043 の低下となっている。また、SSD と手動切り取りを比較すると Kappa スコアで 0.104 の低下となっている。よって、身体全体の動画と組み合わせることで SSD よりも YOLO のほうが優れていることがわかる。

図 4.12, 4.13, 4.14 ではそれぞれ実験 3-4 の YOLO, SSD, 手動切り取りの結果の混同行列を示す。図 4.12 と図 4.13 を比較すると YOLO のほうが State 1 を多く捉えること

表 4.8: 実験 3-4 の結果

身体全体の動画	顔領域の動画	平滑化	組み合わせ	Accuracy	Macro-F1	Kaapa
✓	YOLO	✓	Weighting	0.562	0.426	0.684
✓	SSD	✓	Weighting	0.514	0.369	0.623
✓	手動切り取り	✓	Weighting	0.602	0.518	0.727

表 4.9: 顔領域検出の実験結果 (IoU)

	IOU		切り取り失敗数		睡眠覚醒分類 (Kappa)		
	YOLO	SSD	YOLO	SSD	YOLO	SSD	手動
No2	0.845	0.731	0 / 264	19 / 264	0.638	0.624	0.618
No4	0.826	0.799	2 / 282	13 / 282	0.672	0.611	0.703
No5	0.753	0.639	5 / 289	2 / 289	0.714	0.632	0.811
平均	0.808	0.723	-	-	0.684	0.623	0.727

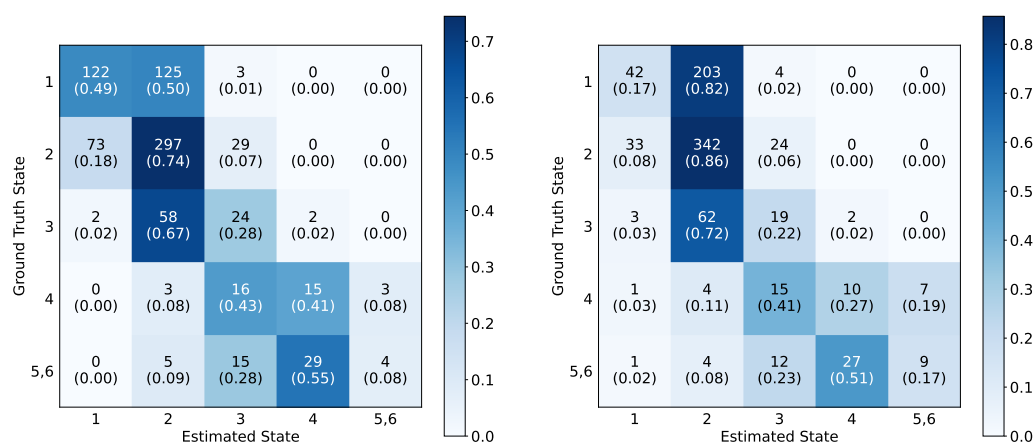


図 4.12: 身体全体+顔領域 (YOLO)+時系 図 4.13: 身体全体+顔領域 (SSD)+時系列
列平滑化+確率分布 平滑化+確率分布

ができています。図 4.14 と比較すると顔領域の自動抽出を行った図 4.12 と図 4.13 どちらも State5,6 を捉えられていないことがわかる。

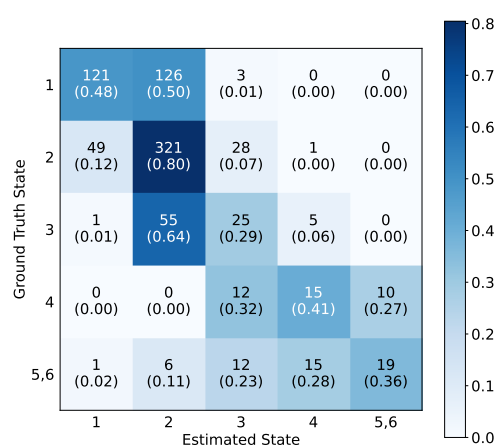


図 4.14: 身体全体+顔領域 (手動切り取り)+時系列平滑化+確率分布

第 5 章

結言

5.1 本研究のまとめ

本研究では動画中の新生児の睡眠・覚醒状態を自動推定する手法を提案した。深層学習モデルの 3DCNN と TimeSformer を用いて比較した結果、睡眠覚醒状態の自動推定では、3DCNN のほうが優れていることがわかった。時系列変化を考慮した推定結果統合手法では時系列平滑化と確率分布による重み付けを行うことで精度向上につながることがわかった。顔領域の自動抽出では身体全体の動画と組み合わせる最終結果では YOLO のほうが SSD よりも高い精度となった。YOLO を用いた顔領域の動画と身体全体の動画のそれぞれから 3 DCNN を用いて出力された結果に時系列平滑化と確率分布による重み付けを行うことで Kappa スコアで 0.684 となった。

5.2 今後の課題

今後の課題としては、実用レベルの推定精度ではないため推定精度向上を目指す。そのために、身体と顔の動画の 2 つを組み合わせる手法を検討していく必要がある。また、データ拡張を行い、学習データ不足を補い、学習データ数の不均衡を是正する。さらに、日本赤十字伊勢病院で新たに収集された動画データを用いた State 判定実験も今後の課題として残されている。

付録 A

付録

本研究に関するプログラムはすべて以下のディレクトリ

- /net/nfs2/home/yuki/workspace/mie-neonatal

本研究に関するデータセットはすべて以下のディレクトリ

- /net/nfs2/home/yuki/data

A.1 プログラムの詳細

/net/nfs2/home/yuki/workspace/mie-neonatal/body/3dcnn

詳しい実行方法は README を参照してください.

謝辞

本研究の全過程において、終始懇切丁寧なご指導を賜りました本学若林哲史教授、盛田健人准教授、また、中間発表などの際には有意義な御助言・ご検討を頂きました白井伸宙助教、また本研究を進めるにあたり、新生児の動画を提供していただいた三重大学医学部看護学科元教授の新小田春美先生をはじめ九州大学病院の方々に深く感謝いたします。

参考文献

- [1] Majid Mirmiran and Simone Lunshof. Perinatal development of human circadian rhythms. *Progress in brain research*, Vol. 111, pp. 217–26, 1996.
- [2] Majid Mirmiran and Ronald L. Ariagno. Influence of light in the nicu on the development of circadian rhythms in preterm infants. *Semin. Perinatol.*, Vol. 24, pp. 247–257, 2000.
- [3] H. Shinkoda, Y. Kinoshita, R. Mitsutake, F. Ueno, H. Arata, C. Kiyohara, Y. Suetsugu, Y. Koga, K. Anai, M. Shiramizu, M. Ochiai, and T. Kaku. The influence of premature infants / sleep and physiological response under nicu environment (illuminance, noise) - seen from circadian variation and comparison of day and night -. *Mie Nursing Journal*, Vol. 17, No. 1, pp. 35–44, 2015.
- [4] T. Berry Brazelton. *Neonatal Behavioral Assessment Scale*. Pastics International Medical Publications, 1973.
- [5] Wei Chen Saadullah Farooq Abbasi, Harun Jamil. Eeg-based neonatal sleep stage classification using ensemble learning. *Computers, Materials & Continua*, Vol. 70, No. 3, pp. 4619–4633, 2022.
- [6] Ninah Koolen, Lisa Oberdorfer, Zsófia Róna, Vito Giordano, Tobias Werther, Katrin Klebermass-Schrehof, Nathan J. Stevenson, and Sampsa Vanhatalo. Automated classification of neonatal sleep states using eeg. *Clinical Neurophysiology*, Vol. 128, pp. 1100–1108, 2017.
- [7] Jan Werth, Mustafa Radha, Peter Andriessen, Ronald M. Aarts, and Xi Long. Deep learning approach for ecg-based automatic sleep state classification in preterm infants. *Biomedical Signal Processing and Control*, Vol. 56, p. 101663, 2020.
- [8] Sandie Cabon, Fabienne Poree, Antoine Simon, Bertille Met-Montot, Patrick Pladys, Olivier Rosec, Nicolas Nardi, and Guy Carrault. Audio- and video-based estimation of the sleep stages of newborns in neonatal intensive care unit. *Biomedical Signal Processing and Control*, Vol. 52, , 05 2019.

- [9] H F Prechtl. The behavioral states of the newborn infant (a review). *Brain Res*, Vol. 76, pp. 185–212, 1974 Aug 16 1974.
- [10] Waseem Rawat and Zenghui Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, Vol. 29, No. 9, pp. 2352–2449, 09 2017.
- [11] Steve R Gunn, et al. Support vector machines for classification and regression. *ISIS technical report*, Vol. 14, No. 1, pp. 5–16, 1998.
- [12] Masashi Hattori, Kento Morita, Tetsushi Wakabayashi, Harumi Shinkoda, Asami Matsumoto, Yukari Noguchi, and Masako Shiramizu. Neonatal sleeping state estimation by body movement detection using optical flow. *Proceedings of the Annual Conference of Biomedical Fuzzy Systems Association*, Vol. 33, pp. 56–59, 10 2020.
- [13] Kento Morita, Nobu C. Shirai, Harumi Shinkoda, Asami Matsumoto, Yukari Noguchi, Masako Shiramizu, and Tetsushi Wakabayashi. Automatic neonatal alertness state classification based on facial expression recognition. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 26, No. 2, pp. 188–195, 2022.
- [14] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1, pp. 886–893 vol. 1, 2005.
- [15] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221–231, 2013.
- [16] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.
- [17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, Vol. abs/1706.03762, , 2017.
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

- [21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.
- [22] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.
- [23] Chien-Yao Wang, Hong-Yuan Mark Liao, and I-Hau Yeh. Designing network design strategies through gradient path analysis. *arXiv preprint arXiv:2211.04800*, 2022.
- [24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, Vol. abs/1512.02325, , 2015.
- [25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [26] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, Vol. 70, No. 4, pp. 213–220, 1968.