高速かつ省メモリなドライバ姿勢推定

Fast and Compact Driver Pose Estimation

西行 健太
Kenta Nishiyuki

オムロン株式会社 技術・知財本部 研究開発センタ 画像センシング研究室
Vision Sensing Lab., Technology Research Center, Technology and Intellectual Property H.Q., OMRON Corporation
kenta.nishiyuki@omron.com

日向 匡史
Tadashi Hyuga

(同 上)
tadashi.hyuga@omron.com

田崎 博
Hiroshi Tasaki

(同 上)
hiroshi.tasaki@omron.com

木下 航一
Koichi Kinoshita

(同 上)
koichi.kinoshita@omron.com

長谷川 友紀
Yuki Hasegawa

(同 上)
yuki.hasegawa@omron.com

山下 隆義
Takayoshi Yamashita

中部大学工学部情報工学科
Department of Information Engineering, Chubu University
takayoshi@isc.chubu.ac.jp

藤吉 弘亘
Hironobu Fujiyoshi

中部大学工学部ロボット理工学科
Department of Robotic Science and Technology, Chubu University
fujiyoshi@isc.chubu.ac.jp

keywords: driver monitoring, autonomous driving, fast and compact pose estimation, embedded system, deep learning

Summary

Driver pose estimation is a key component in driver monitoring systems, which is helpful for driver anomaly detection. Compared with traditional human pose estimation, driver pose estimation is required to be fast and compact for embedded systems. We propose fast and compact driver pose estimation that is composed of ShuffleNet V2 and integral regression. ShuffleNet V2 can reduce computational expense, and integral regression reduce quantization error of heat maps. If a driver suddenly gets seriously ill, the head of the driver is out of view. Therefore, in addition to localizing body parts, classifying whether each body part is out of view is also crucial for driver anomaly detection. We also propose a novel model which can localize and detect each body part of the driver at once. Extensive experiments have been conducted on a driver pose estimation dataset recorded with near infrared camera which can capture a driver at night. Our method achieves large improvement compared to the state-of-the-art human pose estimation methods with limited computation resources. Furthermore, We perform an ablation study of our method which composed of ShuffleNet V2, integral regression, and driver body parts detection. Finally, we show experimental results of each driver action for driver monitoring systems.

1. は じ め に

自動車の安全な運転を実現するために、ドライバモニタリングシステム (DMS) の導入が必要とされている。DMS は眠気や姿勢などのドライバ状態を推定することで、ドライバの異常状態を検知するシステムである。国土交通省が作成したドライバ異常自動検知システム基本設計書 [国土交通省自動車局 18] で示されている通り、ドライバ姿勢の崩れを検知することはドライバの異常検知につながる。[国土交通省自動車局 18] では、ドライバ姿勢情報は物理量 (角度や関節位置) への置き換えが比較的行いやすく、数値でしきい値を定義することが可能であると言及されている。そのため、ドライバ異常を直接推定す

る方式と比べて、ドライバ姿勢推定を用いた異常検知では、誤識別の要因が姿勢推定結果にあるのか、しきい値などの異常判定方法にあるのかを解析しやすい。本論文では、ドライバの異常検知を目的として、高速かつ省メモリなドライバ姿勢推定を提案する。

従来の DMS の多くはドライバの顔部分のみに着目し、脇見などを検知する。しかし、飲食や通話など多様なドライバの状態を検知するためには、頭部の姿勢だけではなく、手や首元などのより多くの関節点を捉えることが重要となる。本論文では、頭部だけではなく、首元、両手といった上半身全体のドライバ姿勢を推定する。

人物の関節点座標を推定する人物姿勢推定は、コンピュータビジョンの分野で活発に研究されている。近年、

OpenPose [Cao 17] など Deep Neural Network (DNN) を用いた高精度な人物姿勢推定が提案されている。これらの関連研究では、MS-COCO [Lin 14], MPII [Andriluka 14], LSP [Johnson 10], FLIC [Sapp 13] などのデータセットが使用されている。これらのデータセットを用いて評価していた関連研究の人物姿勢推定と、ドライバ姿勢推定では消費リソースやデータセットの条件などが異なる。

デスクトップ PC やサーバーなどと比べて、DMS は電力量が制限されるため、消費リソースが限られた組込み機器での動作が求められる。DNN を用いた姿勢推定は高精度であるが、消費リソースが多くなる傾向にある。DNN を用いた姿勢推定を DMS に搭載するためには、ネットワークモデルを軽量化し、高速かつ省メモリなモデルを構築する必要がある。本論文では、ShuffleNet V2 と Integral Regression をベースとし、高速かつ省メモリなドライバ姿勢推定を提案する。

DNN を用いた姿勢推定に用いられる MS-COCO などのデータセットは、被写体とカメラ間の距離が遠いため、被写体の全身が映るシーンが多く、ほとんどの関節点が画像中に映る。そのため、人物姿勢推定で一般的に行われる評価では、推定した関節点座標と正解の誤差のみを評価し、関節点が画像中に映っているかどうかは評価しない。一方、DMS では車内にカメラを設置するため、被写体とカメラ間の距離が近く、上半身など一部の関節点しか映らないことが多い。更に、ドライバが急病などで倒れた場合には、頭部が画像に映らないため、画像中に関節点が映っているかどうか（関節点有無）の判定が、ドライバの異常検知に役立つ。しかし、従来の人物姿勢推定は関節点有無を評価しないため、ドライバ姿勢推定に適さない。本論文では、関節点座標だけでなく関節点有無も推定するモデルを提案する。

人物姿勢推定のデータセットは、可視光カメラで撮影されたものが多い。一方、DMS では夜間でもドライバを撮影する必要があるため、近赤外カメラで撮影されることが多く、色情報が失われる。本論文では、人物が上半身など一部の関節点のみしか映らない、かつ、色情報がないデータセットを作成し、DMS での利用を想定した評価実験を行い、提案手法の効果を示す。

2. 関 連 研 究

人物姿勢推定はトップダウン型とボトムアップ型に分けられる。トップダウン型では、人検出や顔検出を用いて人物領域を特定し、特定した人物領域に対して姿勢推定を行う。一方、ボトムアップ型では、画像から人物の関節候補点を抽出し、それらの関節候補点を関節間の関係性などにに基づき、つないでいくことで複数人物の姿勢推定を一度に行う。

2.1 トップダウン型姿勢推定

DNN を用いた姿勢推定に関する初期の研究では、画像から関節点座標を直接推定する回帰を用いた手法が提案されていた。Toshev らは、段階的に関節点座標を推定する DeepPose を提案した [Toshev 14]。DeepPose では入力画像から関節点座標を推定した後、推定した座標付近の画像を切り出し、その画像から再度関節点座標を推定する。DeepPose では、このような処理を繰り返すことで関節点座標の推定精度を高める。

人物の関節点座標を一意に決定することは難しく、回帰を用いたネットワークの学習は難しい。そこで、Tompson らは関節点座標をヒートマップとして表現することで DeepPose よりも高精度な姿勢推定を実現した [Tompson 14]。ヒートマップでは学習に使用する関節点座標の正解を、正解の関節点座標を中心として広がる 2 次元の正規分布として表現する。ヒートマップでは人物の関節点座標を一意に決定しないため、学習が安定し、精度が高くなる。以降、ヒートマップを用いた姿勢推定が主流となり、ヒートマップを出力するネットワーク構造を工夫した手法が提案された。

Wei らは、段階的に姿勢推定の精度を向上させていく Convolutional Pose Machines (CPMs) を提案した [Wei 16]。最初のステージでは画像を入力として、ヒートマップを出力する。以降のステージでは、前ステージの出力結果であるヒートマップと、画像に畳み込みとプーリングを適用した特徴マップを入力として、関節点座標を表すヒートマップを出力する。各ステージでヒートマップと正解の誤差を計算して学習することで、勾配消失問題を回避している。また、各ステージにはプーリング層があるため、ステージを経るごとに受容野が大きくなり、より大局的な特徴を考慮した姿勢推定が行える。しかし、CPMs では、ステージを経るごとに、高解像度の情報が失われる。

Newell らは、高解像度から低解像度にするダウンサンプリング処理と、低解像度から高解像度にするアップサンプリング処理を行う Hourglass Module を提案した [Newell 16]。ダウンサンプリング処理とアップサンプリング処理を行うことで様々な空間解像度の情報を抽出できる。また、Hourglass Module には skip connection があり、ダウンサンプリング処理でも高解像度の情報が保持される。また、CPMs と同様に、各 Hourglass module で正解の関節点座標を用いて学習するため、Hourglass module を重ねても勾配消失を抑制できる。

Yan らは、空間解像度の変化に頑健な Pyramid Residual Module を提案した [Yang 17]。Hourglass module では、Skip connection を一つの空間解像度情報を保持するために使用する。一方、Pyramid Residual Module では、異なる空間解像度間で Bottom-up と Top-down 処理を並列に行うことで多様な空間解像度の特徴を捉えられる。

Hourglass や Pyramid Residual Module などは高精度

である一方、ネットワーク構造が複雑である。そのため、Xiao らは、ResNet と少数の Deconvolutional 層を用いた単純なネットワーク構造を用いてベースラインのモデルを提案した [Xiao 18]。提案手法であるベースラインモデルは、単純な構造で既存手法と同等以上の精度を達成した。

Sun らは [Xiao 18] をベースとしてモデル構造を改良し、空間解像度ごとにネットワークを分岐させて処理する HRNet を提案した [Sun 19]。HRNet では、高解像度の情報は解像度を落とすことなく、下位層から上位層へ伝えられる。そのため、Pyramid Residual Module と同様に、高解像度の特徴を抽出することができ、高精度な姿勢推定を実現している。

2.2 ボトムアップ型姿勢推定

Pishchulin らは、ヒートマップで推定した関節候補点同士をつなぎ合わせてグラフとみなし、各人物の正しい関節点の対応付けを整数線形計画問題として最適化する DeepCut を提案した [Insafutdinov 16, Pishchulin 16]。DeepCut は整数線形計画問題による最適化の処理時間が多く、リアルタイムの姿勢推定が難しかった。

Cao らは関節点間の関係をベクトル場で表現した Part Affinity Field (PAF) を用いて関節候補点をつなげることで、複数人の姿勢をリアルタイムに推定する OpenPose を提案した [Cao 17]。DeepCut に比べて、PAF を用いた関節候補点の対応付けは、画像中に映る人物数が増えた場合でも、一定の処理時間で複数人物の姿勢を推定できる。

Kreiss らは、遠方に映る低解像度の人物や、混雑しているシーンであっても高精度な姿勢推定を行うため、高解像度なヒートマップを生成する Part Intensity Fields (PIF) と、混雑しているシーンでも正確に関節同士のつながりを表現できる Part Association Fields (PAF) を提案した [Kreiss 19]。

2.3 関連研究のドライバ姿勢推定への適用

ボトムアップ型の姿勢推定は、関節点間の関係を用いて関節候補点をつなげることで各人物の関節点の対応付けを行うが、画像中に一部の関節点しか映ることのないドライバ姿勢推定では、多くのシーンで関節点間の関係を利用することができない。そのため、関節点間の関係を表すヒートマップを出力する畳み込み層は、不要な演算量の増大につながる。従って、本論文ではトップダウン型の姿勢推定を対象とする。トップダウン型では顔検出や人物検出を行い、人物領域を抽出したのち、姿勢を推定するが、本論文で扱うドライバ姿勢推定ではドライバのみが映るシーンを想定しているため、入力画像から直接ドライバの姿勢を推定する。

2.4 人物姿勢推定のデータセット

人物姿勢推定に用いるデータセットには、Leads Sports Pose Dataset (LSP) [Johnson 10], Frames Labeled In Cinema (FLIC) [Sapp 13], MPII Human Pose Dataset (MPII) [Andriluka 14], Microsoft Common Objects in Context (COCO) [Lin 14] がある。これらのデータセットは、被写体とカメラ間の距離が比較的遠く、多くのシーンで被写体の全身が映る。更にこれらのデータセットは、可視光カメラで撮影されたものであり、色情報が利用できる。一方、ドライバ姿勢推定では、被写体とカメラ間の距離が近い場合、多くのシーンで上半身など一部の関節点しか映らない。また、夜間でもドライバを監視する必要があるため、近赤外カメラで撮影されたグレイスケール映像となる。そのため、人物姿勢推定で用いられるこれらのデータセットは、ドライバ姿勢推定には適さない。

2.5 人物姿勢推定の評価方法

人物姿勢推定の評価指標では、Percentage of Correct Parts (PCP), Percentage of Correct Keypoints (PCK), Percentage of Detected Joints (PDJ), Average Precision (AP), Average Recall (AR) が用いられる。PCP, PCK, PDJ は、単一の人物姿勢推定に用いられる指標であり、推定した関節点座標と正解の誤差がしきい値よりも小さいときに推定結果が正しく行われたとし、その正解率を評価指標とする。AP と AR は、COCO で用いられる評価指標であり、推定した関節点と正解の類似度を示す尺度である Object Keypoint Similarity (OKS) を用いて算出される。これらの評価尺度は、画像中に対象の関節点映っていない場合には、評価尺度に反映されない。従って、人物姿勢推定の評価指標は関節点有無の評価には適さない。

2.6 ドライバ姿勢推定

DMS に関する研究の多くが、ドライバの顔部分のみに着目したドライバ姿勢推定を行っていた [Borghini 17, Jha 17]。しかし、飲食や通話など多様なドライバの状態を検知するためには、頭部の姿勢だけではなく、手や首元などの関節点座標を推定する必要がある。Eshed らは2台のカメラを用いて、手と顔の姿勢を捉えてドライバがどの部分に注目しているかを推定した [Ohn-Bar 14]。しかし、Eshed らは可視光カメラで撮影した画像を用いており、夜間のシーンには対応していない。また、手の姿勢はハンドル、ギア、車載機器のどの領域にあるかのみを判定しているため、飲食や通話などの多様なドライバの状態を推定することは難しい。

3. 提案手法

3.1 概要

ドライバ姿勢推定では組込機器で動作可能な、高速かつ省メモリなネットワークモデルが求められる。消費リ

ソースを削減するためには、精度を低下させることなく、ヒートマップの解像度を落としたり、演算量の多い畳み込み処理を減らす必要がある。本論文では、これらを実現するために、Integral Regression [Sun 18] と ShuffleNet V2 [Ma 18] をベースとしたネットワークモデルを提案する。また、関節点座標の推定と同時に、画像中に各関節点が映っているかどうか（関節点有無）を判定することで、一部の関節点しか映らないような DMS 特有のシーンに対応した手法を提案する。

提案手法のネットワークモデルを図 1 に示す。HRNet [Sun 19] は高精度であるが、複数の解像度を処理するため畳み込み処理の演算量が増大する傾向にあるため、提案手法は Hourglass [Newell 16] をベースとしたモデルとする。入力画像から 2 つの Hourglass Module により特徴抽出を行った後、各関節点に対して座標と有無を推定する。関節点座標は特徴マップに対して畳み込みと Softmax を適用してヒートマップを生成し、ヒートマップに対して Integral Regression を適用して座標を推定する。関節点有無は特徴マップをヒートマップで重み付けし、Global Average Pooling を適用する。その後、全結合層を適用して関節点有無を表す信頼度を出力する。

3.2 高速かつ省メモリなドライバ姿勢推定

高速かつ省メモリなドライバ姿勢推定を実現するため、精度低下を抑制しつつ、入力画像の低解像度化と畳み込み処理の高速化を行う。

入力画像の低解像度化は、消費リソースの削減につながるが、関節点座標を表すヒートマップも低解像度となるため、量子化誤差を招く。そのため、提案手法では Sun らが提案した Integral Regression [Sun 18] を用いて、量子化誤差を抑制しつつ、入力画像を低解像度にする。また、人物姿勢推定では Convolutional Neural Network (CNN) が一般的に良く使用されるが、CNN は畳み込みの演算量が多く、処理時間が増大する傾向にある。本論文では、Ma らが提案した ShuffleNet V2 [Ma 18] を用いて、畳み込みの消費リソースを削減する。

§1 Integral Regression

ヒートマップを用いた人物姿勢推定では、ヒートマップ上の最大値を取る座標を関節点座標とする。そのため、ヒートマップの解像度が低くなるほど、関節点座標の量子化誤差が大きくなる。一方、ヒートマップの解像度を高くすると消費リソースが増加する。そこで、Sun らは Integral Regression を用いてヒートマップ分布の重心位置を関節点座標として出力することで、サブピクセル単位での座標推定を可能にした [Sun 18]。ヒートマップの解像度が入力解像度より低い場合でも、Integral Regression では高精度な関節点座標の推定が可能であり、消費リソースの削減につながる。また、Integral regression は最大値に基づく関節点座標の推定と異なり、ヒートマップ分布全体に関して微分可能であるため End-to-end で学習で

きる。

提案手法では Integral Regression を用いて、量子化誤差を抑制しながら、入力画像とヒートマップの解像度を小さくする。従来のヒートマップを用いた関節点座標の推定では、式 (1) の $\arg\max$ を用いてヒートマップの最大値から関節点座標を推定する。 H_k は k 番目の関節点のヒートマップ、 p はヒートマップ上の座標、 J_k は k 番目の関節点の座標を示す。Integral Regression では式 (2) のように、ヒートマップの期待値から関節点座標を算出する。 Ω は座標の集合を示す。なお、Integral Regression で使用されるヒートマップ $\hat{H}_k$ は、式 (3) のような softmax を用いて、全ての要素が非負であり、全ての要素を足すと 1 になるように正規化される。式 (2) は Soft Argmax としても知られる。

$$J_k = \arg\max_p H_k(p) \quad (1)$$

$$J_k = \int_{p \in \Omega} p \cdot \hat{H}_k(p) \quad (2)$$

$$\hat{H}_k(p) = \frac{e^{H_k(p)}}{\int_{q \in \Omega} e^{H_k(q)}} \quad (3)$$

§2 ShuffleNet V2

畳み込み処理の演算量を減らすため、提案手法では通常の畳み込み層の代わりに ShuffleNet V2 で提案されたモジュールを用いる。ShuffleNet V2 モジュールの構成を図 1 の最下部に示す。このモジュールでは、初めにチャンネルを分割し、一方では畳み込み処理を行い、もう一方では畳み込み処理を行わない。これらのチャンネルを結合した後、チャンネルの順序を入れ替えることで畳み込み処理の回数を削減しながら、全てのチャンネルに畳み込み処理を適用する。また、ShuffleNet V2 は 3 つの畳み込み層から構成される。1 つ目は、 1×1 のフィルタを用いた畳み込み処理、2 つ目は Depthwise の畳み込み処理、3 つ目は 1 つ目と同じく 1×1 のフィルタを用いた畳み込み処理である。2 つ目の Depthwise の畳み込み処理ではチャンネル方向の畳み込み処理を行わず、空間方向に対してのみ畳み込み処理を行うため、通常の畳み込み処理よりも計算量が少ない。

3.3 関節点有無の判定

人物姿勢推定では、多くの関節点が画像中に映るシーンで評価を行っていた。そのため、関節点座標と正解座標の誤差のみを評価し、関節点有無は評価していなかった。一方、ドライバ姿勢推定では、上半身など人物の一部分しか映らないため、画角内に関節点が入らないシーンが多い。例えば、意識を失っており、頭部が映っていない場合には関節点有無がドライバの異常検知に役立つ。また、人物姿勢推定では関節点座標をヒートマップから推定しているが、ヒートマップ上の最大値がしきい値以上かどうかで、関節点有無を推定できる。しかし、その

図1 ネットワークモデル全体 (提案手法, 入力画像 96×64, 54ch)

ような場合は関節点有無の学習を行っていないため、関節点毎にヒートマップの最大値が大きく異なり、適切なしきい値の設定が難しい。提案手法では関節点座標の推定と同時に関節点有無を推定するため、しきい値の設定が不要となる。

3.4 学 習

提案手法では入力画像1枚に対して、ヒートマップ、関節点座標、関節点有無の3つの損失を計算し、一度にネットワークモデルを学習する。ヒートマップの損失関数はL2-lossを用いる。ヒートマップのGround Truth (GT)は正解の関節点座標を中心として広がる正規分布を用いて作成する。ヒートマップの損失 L_H は、式(4)となる。 H_k は k 番目の関節点のヒートマップ、 p はヒートマップ上の座標、 H_k^* はヒートマップのGT、 K は関節点の数となる。

$$L_H = \sum_{k=1}^K \sum_p \|H_k(p) - H_k^*(p)\|_2^2 \quad (4)$$

提案手法ではIntegral Regressionを用いるため、関節点座標についても損失を計算できる。関節点座標のGTは入力画像サイズを基準とした座標とする。関節点座標の損失関数 L_C は、L1-lossを用いて、式(5)となる。 C は関節点座標、 C^* は関節点座標のGT、 j は j 番目の関節点を示す。

$$L_C = \sum_{j=1}^J \|C(j) - C^*(j)\| \quad (5)$$

関節点有無の損失関数 L_D は、クラス間のデータの偏りに対応するため、Focal loss [Lin 17]を用いて、式(8)となる。 x は関節点有無を示す2値の出力値、 t は関節

点有無のGT、 CE はクロスエントロピー、 γ は損失を調整するパラメータを示す。本論文では、 γ を2とする。 t は関節点が画像中に映る場合に1、画像に映らない場合に0とする。

$$CE(x_t) = -\log(x_t) \quad (6)$$

$$x_t = \begin{cases} x & t = 1 \\ 1 - x & \text{otherwise} \end{cases} \quad (7)$$

$$L_D = -(1 - x_t)^\gamma \log(x_t) \quad (8)$$

関節点有無のGTは関節点が画像中に映る場合に1、映らない場合に0とする。

提案手法のネットワークモデルは、これら3つの損失をあわせた式(9)を用いて学習する。

$$L = L_H + L_C + L_D \quad (9)$$

4. 実 験

提案手法の有効性を確認するため、既存手法との精度比較を行う。DMSでは一般的な人物姿勢推定と比較して、消費リソースが限られるため、提案手法と既存手法の計算量が同等になるようパラメータ数を揃えたネットワークモデルを用いる。また、本論文ではAblation Studyとして、提案手法を構成するIntegral Regression, ShuffleNet V2, 関節点有無の判定の3つの要素を除いた際の精度比較を行う。最後にドライバの行動パターン毎の評価を行い、行動パターンの傾向を確認する。

図2 データセットの画像例 (同意書により画像利用許諾確認済み)

4.1 実験データ

本論文ではドライビングシミュレータを用いて、16 パターンの動作を 100 人の被験者で撮影し、ドライバ姿勢推定の評価データセットを作成した。16 パターンの動作は、前方注視、脇見、もたれ、ストレッチ、振り返り、眠気、スマホ操作、通話、カメラ撮影、読書、飲食、着替え、突っ伏し、パニック、居眠り、赤ちゃんを抱く、となる。100 人の被験者のデータのうち、50 人分を学習に、残りの 50 人分を評価用に用いる。外乱光の影響を軽減するため、我々が試作した近赤外線カメラを用いて撮影した。撮影した動画の解像度は 752×480 であり、それを 96×64 に縮小して使用する。また、計算量を減らすため、10fps で撮影した動画を 1fps に間引いて使用する。540,434 枚を学習用、528,124 枚を評価用として用いる。撮影した映像にはドライバの上半身のみが映るため、関節点は頭部中心、首元、左手中心、右手中心の計 4 点とする。図 2 に撮影した画像の例を示す。

4.2 評価実験のパラメータ

§1 入力画像、出力ヒートマップ

本実験では入力画像を 96×64 にリサイズしたのち、Global Contrast Normalization [Jarrett 09] を用いて正規化を行う。また、本実験で用いるヒートマップのサイズは 24×16 とする。評価は入力画像と同じ 96×64 の解像度に対して行う。

§2 ネットワーク

本実験では提案手法と既存研究の HRNet [Sun 19], Hourglass [Newell 16], OpenPose [Cao 17] を比較する。また、消費リソースが制限された組込み機器への搭載を想定して、1 回の姿勢推定にかかる演算量が 100, 50, 25MFLOPs 程度になるモデルを用いて実験を行う。各モデルの消費リソースを表 1 に示す。畳み込み層の出力チャンネル数を減らすことで演算量を調整する。提案手法のネットワーク構成を、図 1 に示す。本実験で用いる提案手法の 3 つのモデルは、Hourglass Module 内の畳み込み層での出力チャンネル数が異なる。図 1 に示すモデルは、54ch モデルであり、Hourglass Module に含まれる畳み込み層の出力チャンネル数は 54ch, 108ch, 216ch, 432ch となる。32ch モデルの出力チャンネル数は 32ch, 64ch, 128ch, 256ch, 24ch モデルは 24ch, 48ch, 96ch, 192ch となる。本実験に用いる HRNet は、Sun らが提案したモデルと同様に 4 つのステージから構成される。演算量を調整するため、HRNet でも畳み込み層の出力チャネル

ル数を削減する。本実験では、1 つ目のステージの出力チャンネル数を 16ch, 12ch, 8ch に減らした HRNet のモデルを用いる。2 つ目以降のステージも、出力チャンネル数は 1 つ目のチャンネル数に合わせて削減する。本実験に用いる Hourglass は、提案手法で用いる Hourglass Module と同じものを使用する。ただし、ShuffleNet V2 は通常の畳み込み層に置き換え、Integral Regression や関節点有無の出力は行わない。Hourglass でも畳み込み層の出力チャンネル数を削減して演算量を調整する。本実験に用いる OpenPose では、学習を安定させるため、Cao らが提案したモデルに対して、Batch Normalization(BN) 層 [Ioffe 15] を追加する。また、Cao らのモデルでは複数ステージから構成されているが、処理量削減のため、本実験では 1 つ目のステージのみ使用する。更に 1 層目の畳み込み層の出力チャンネル数を、4ch, 6ch, 8ch に減らし、以降の畳み込み層の出力チャンネル数も、1 層目に合わせて削減する。

表 1 ネットワークの演算量及びパラメータ数: * OpenPose の演算量には PAF による関節点同士の関連付け処理部は含まない。

手法	MFLOPs	パラメータ数 [M]
100 MFLOPs		
HRNet [Sun 19], 16ch	113.0	3.90
Hourglass [Newell 16], 20ch	109.2	1.43
OpenPose [Cao 17], 8ch	111.2*	0.13
提案手法, 54ch	102.16	1.14
50 MFLOPs		
HRNet [Sun 19], 12ch	65.3	2.18
Hourglass [Newell 16], 15ch	61.6	0.81
OpenPose [Cao 17], 6ch	62.7*	0.07
提案手法, 32ch	43.0	0.51
25 MFLOPs		
HRNet [Sun 19], 8ch	28.3	0.97
Hourglass [Newell 16], 10ch	27.6	0.36
OpenPose [Cao 17], 4ch	28.0*	0.03
提案手法, 24ch	28.0	0.36

§3 ハイパーパラメータ

初期の学習率は 0.001 とする。300iteration の間に 0.004 まで増加させた後、5000iteration ごとに学習率を半減し、合計で 30000iteration の学習を行う。重み減衰は 0.0001, 勾配のクリッピングは 5.0 とする。バッチサイズは 1GPU あたり 96 に設定し、4 つの GPU を用いる。

§4 学習の詳細

i. Data Augmentation

手が映っているデータと映っていないデータが同等程度になるよう over sampling を行う。また、Data Augmentation として、下記 4 つの処理を行う。

- 画像の高さと幅の $\pm 25\%$ の範囲でランダムに並行移動
- 0.8~1.75 倍の範囲でランダムに拡大縮小

- 50%の確率で左右反転
- 0.5～1.5 倍の範囲でランダムに明度調整

ii. Ground Truth

ヒートマップの学習に用いる Ground Truth (GT) は、正解の関節点座標から σ が 2 ピクセルの正規分布に従って作成したヒートマップを用いる。また、関節点有無の GT は関節点が画像に映る場合に 1、画像に映らない場合に 0 とする。

§5 評価方法

i. 評価指標

本実験では関節点座標の評価指標は Probability of Correct Keypoints (PCK) を用いる。本実験では解像度が 96×64 の入力画像に対して、PCK のしきい値を 6 ピクセルとする。関節点有無の評価指標は、mean Average Precision (mAP) を用いる。

ii. 関節点座標、関節点有無の推定

提案手法では Integral Regression を用いるため、関節点座標はモデルの出力をそのまま用いる。一方、既存研究ではヒートマップ上で最大値を取る座標を関節点座標とする。また、提案手法では関節点有無の判定もモデルの出力を用いる。既存研究ではヒートマップの最大値がしきい値以上かどうかで関節点有無を判定する。

4.3 精度比較

関節点座標の精度比較結果を表 2 に示す。なお、OpenPose では PAF を用いて、関節点の関連付けを行うが、本実験では単一関節点のみ出現するシーンが多いため、関節点の出力のみを用いて、評価する。全ての演算量において、既存研究よりも提案手法の方が精度が高い。特に両手の精度では、提案手法は既存研究よりも 5～10% 程度改善している。従って、提案手法は消費リソースを削減した時の精度低下を抑制できる。OpenPose はパラメータ数が少ないが、50MFLOPs、25MFLOPs のように演算量を大幅に削減した場合に、提案手法よりも精度が大きく低下する。そのため、OpenPose では省メモリ化と高速化を両立しようとした場合に精度低下が起こる。

関節点有無の精度比較結果を表 3 に示す。100MFLOPs と 50MFLOPs のモデルでは、モデルによる精度の差がほとんどない。一方、25MFLOPs のモデルでは、HRNet の精度が Hourglass、OpenPose、提案手法と比べて低くなっている。Hourglass は提案手法よりも mAP が 1% 程度高い。これより、関節点有無の精度について、提案手法は既存研究の Hourglass と同等程度に、消費リソースを大幅に削減した時の精度低下を抑制できる。

§1 実験結果画像

提案手法と既存研究の姿勢推定結果画像を図 3 に示す。既存手法の関節点有無のしきい値は 0.5 とした。図 3 より、提案手法は既存研究よりも関節点座標の推定結果が正解座標に近く、未検出も少ないことがわかる。飲み物

表 2 精度比較 (関節点座標)

手法	PCK-6px (頭部)	PCK-6px (首元)	PCK-6px (右手)	PCK-6px (左手)
100 MFLOPs				
HRNet [Sun 19], 16ch	90.4%	91.5%	82.9%	72.2%
Hourglass [Newell 16], 20ch	89.9%	91.43%	82.5%	74.3%
OpenPose [Cao 17], 8ch	89.6%	88.2%	83.5%	74.9%
提案手法, 54ch	91.9%	94.5%	90.5%	84.0%
50 MFLOPs				
HRNet [Sun 19], 12ch	89.7%	91.1%	82.3%	71.7%
Hourglass [Newell 16], 15ch	89.2%	91.1%	83.0%	73.1%
OpenPose [Cao 17], 6ch	88.8%	89.5%	73.9%	72.7%
提案手法, 32ch	91.8%	94.0%	89.8%	82.3%
25 MFLOPs				
HRNet [Sun 19], 8ch	88.6%	90.3%	78.6%	64.3%
Hourglass [Newell 16], 10ch	89.5%	90.0%	80.6%	69.1%
OpenPose [Cao 17], 4ch	88.0%	87.4%	79.5%	67.2%
提案手法, 24ch	90.9%	93.5%	88.1%	79.1%

を把持している左手のように遮蔽があるような場合でも、提案手法では手の中心座標を正確に推定できる。

パニック動作時のヒートマップを図 4 に示す。ヒートマップは出力値が 0.5 の場合は赤色、0.0 は青色とした。図 4 より、既存研究の Hourglass では左手のヒートマップに出力があるが、関節点によりヒートマップの最大値が大きく異なるため、ヒートマップの最大値に対して同一のしきい値を用いると、未検出や誤検出が起きる。既存手法では、関節点座標のみを正解値として学習するが、提案手法では関節点座標と同時に関節点有無も学習する。そのため、提案手法ではしきい値を調整することなく、関節点有無を推定できる。

4.4 Ablation Study

提案手法は ShuffleNet V2, Integral Regression, 関節点有無の判定、の 3 つの要素から構成される。本実験では Ablation Study として、提案手法から各要素を除いた実験を行い、どの要素が演算量低下時の精度低下抑制に寄与するかを確認する。ShuffleNet V2 は通常の畳み込み層に置き換える。ただし、畳み込み層は ShuffleNet V2 に比べて演算量が多いため、出力チャンネル数を 20ch に減らして演算量が同程度になるよう調整する。Integral Regression は argmax に置き換え、ヒートマップ上で最大値を取る座標を関節点座標とする。関節点有無の判定は単純なしきい値処理に置き換え、推定した関節点座標のヒートマップの値がしきい値以上かどうかで関節点有無を判定する。

Ablation Study の関節点座標精度を表 4 に示す。上から提案手法 (Proposed)、提案手法から ShuffleNet V2 を除いたモデル (No-SNV2)、Integral Regression を除いた

図3 姿勢推定結果：青色が頭部中心，緑色が首元，赤色が右手中心，水色が左手中心の座標を示す。○は推定結果，×は正解座標を示す。

図4 ヒートマップ(パニック)

モデル (No-IR), 関節点有無の判定を除いたモデル (No-Det) の結果である。ShuffleNet V2 を除いたモデルでは提案手法に比べて 1% 程度精度が低下している。Integral Regression を除いたモデルでは，関節点座標の推定精度が大幅に低下している。関節点有無の判定は関節点座標の推定精度にはほとんど影響しない。

Ablation Study の関節点有無の精度を表 5 に示す。関節点有無については，ShuffleNet V2 と Integral Regression を除いた場合でも精度はあまり変わらない。一方，関節点有無の判定を除いたモデルでは精度が低下する。図 4 にて示す通り，既存研究の Hourglass では関節点によりヒートマップの最大値が大きく異なるため，適切なしきい値を設定することが難しい。一方，関節点有無の判定ではしきい値を調整することなく，関節点有無を判定できるため，精度に影響を与えたと考えられる。

4.5 動作パターン毎の評価

動作パターン毎の評価結果を表 6，表 7 に示す。動作によっては部位が全く映らないことがあるため，関節点有無の評価指標は正解率 (正解フレーム数 / 全フレーム数) を用いる。関節点座標では，多くの動作が高精度であるが，振り返り動作の両手，読書動作の左手，突っ伏し動作の頭部及び両手については精度が低い。これらの動作は手が映るシーンが少ないため，精度が低下したと考えられる。しかし，これらの動作は，手の座標よりも頭部座標が行動の推定に重要だと考えられるため，ドライバの異常検知には大きく影響しない。一方，通話，カメラ撮影，飲食といった動作は物体を把持することが行動の推定に重要であるため，手の位置の推定精度が重要となる。これらの動作パターンでは，多くのシーンで手が映るため，手の推定精度が 90% 以上と精度が高くなったと

表 3 精度比較 (関節点有無)

手法	AP (頭部)	AP (首元)	AP (右手)	AP (左手)	mAP
100 MFLOPs					
HRNet [Sun 19], 16ch	99.9%	98.4%	92.1%	83.9%	93.6%
Hourglass [Newell 16], 20ch	99.9%	98.3%	91.8%	84.7%	93.7%
OpenPose [Cao 17], 8ch	99.9%	98.7%	89.4%	81.2%	92.3%
提案手法, 54ch	99.9%	98.4%	91.6%	84.2%	93.5%
50 MFLOPs					
HRNet [Sun 19], 12ch	99.9%	98.3%	92.0%	84.2%	93.6%
Hourglass [Newell 16], 15ch	99.9%	98.5%	91.6%	83.9%	93.5%
OpenPose [Cao 17], 6ch	99.9%	98.4%	89.6%	86.1%	93.5%
提案手法, 32ch	99.9%	98.6%	90.6%	84.5%	93.4%
25 MFLOPs					
HRNet [Sun 19], 8ch	99.9%	98.3%	88.5%	74.7%	90.4%
Hourglass [Newell 16], 10ch	99.9%	98.4%	91.9%	83.1%	93.4%
OpenPose [Cao 17], 4ch	99.9%	98.3%	90.9%	81.4%	92.6%
提案手法, 24ch	99.9%	98.2%	89.3%	81.7%	92.3%

表 4 Ablation Study(関節点座標)

手法	PCK-6px (頭部)	PCK-6px (首元)	PCK-6px (右手)	PCK-6px (左手)
Proposed, 54ch	91.9%	94.5%	90.5%	84.0%
No-SNV2, 20ch	91.8%	94.4%	89.7%	81.3%
No-IR, 54ch	86.8%	82.4%	71.9%	60.9%
No-Det, 54ch	91.8%	94.7%	90.7%	83.6%

考えられる。また、突っ伏し動作では、ドライバがハンドル方向に突っ伏す。そのため、頭部が大きく映り、頭部の正確な中心座標は人目でも判別が難しい。突っ伏し動作は、画面下部に頭部位置が移動したことにより、ドライバ異常の検知が可能であるため、大まかな頭部中心が推定できれば問題ない。

関節点有無については、スマホ操作の首元と読書時の首元の精度が低い。スマホ操作と読書時の見え方は同じであり、どちらの動作もドライバが下を向いている。そのため、首元の多くが隠れてしまい、誤判定が多くなったと考えられる。しかし、どちらの動作も頭部位置に特徴があるため、首元の有無はドライバ異常の検知には影響しない。

5. ま と め

本論文では、DMSのためのドライバ姿勢推定手法を提案した。提案手法では、ShuffleNet V2とIntegral Regressionを用いることで高速で省メモリな姿勢推定を実現できることを示した。DMS特有の関節点が見えないシーンに対応するため、関節点有無の判定を提案した。また、ドライビングシミュレータで撮影した評価データを用いて、提案手法が既存手法に比べて、モデルサイズ低減時の精度低下を抑制できることを示した。本論文ではドラ

表 5 Ablation Study(関節点有無)

手法	AP (頭部)	AP (首元)	AP (右手)	AP (左手)	mAP
Proposed, 54ch	99.9%	98.4%	91.6%	84.2%	93.5%
No-SNV2, 20ch	99.9%	98.3%	90.8%	82.7%	92.9%
No-IR, 54ch	99.9%	98.5%	92.5%	85.9%	94.2%
No-Det, 54ch	99.9%	91.6%	90.2%	84.8%	91.6%

表 6 行動パターン別評価 (関節点座標)

手法	PCK-6px (頭部)	PCK-6px (首元)	PCK-6px (右手)	PCK-6px (左手)
前方注視	99.9%	100.0%	96.6%	96.6%
脇見	99.2%	99.3%	94.7%	93.8%
もたれ	99.3%	96.3%	97.7%	92.6%
ストレッチ	94.5%	92.1%	75.9%	75.9%
振り返り	83.6%	81.2%	54.4%	32.5%
眠気	99.1%	99.3%	86.7%	99.3%
スマホ操作	97.5%	99.2%	100.0%	100.0%
通話	99.9%	99.8%	98.7%	97.6%
カメラ撮影	92.7%	90.8%	94.4%	93.6%
読書	97.8%	98.4%	100.0%	0.0%
飲食	98.9%	97.7%	93.3%	91.5%
着替え	85.8%	88.2%	74.5%	78.3%
突っ伏し	43.9%	98.0%	44.2%	49.4%
パニック	89.1%	85.4%	83.1%	80.8%
居眠り	98.6%	96.3%	88.9%	100.0%
赤ちゃんを抱く	92.7%	93.6%	75.3%	73.4%

イビングシミュレータのデータを用いて評価を行ったが、更なる実用性向上のため、今後は実写映像での評価を実施していく必要がある。本論文では精度低下の抑制に注力したが、今後は実用性向上のため更なる精度向上に取り組む。

◇ 参 考 文 献 ◇

- [Andriluka 14] Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, in *CVPR*, pp. 3686–3693 (2014)
- [Borghi 17] Borghi, G., Venturelli, M., Vezzani, R., and Cucchiara, R.: Poseidon: Face-from-depth for driver pose estimation, in *CVPR*, pp. 4661–4670 (2017)
- [Cao 17] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y.: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields, in *CVPR*, pp. 7291–7299 (2017)
- [Insafutdinov 16] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B.: Deeppose: A deeper, stronger, and faster multi-person pose estimation model, in *ECCV*, pp. 34–50 (2016)
- [Ioffe 15] Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in *ICML*, pp. 448–456 (2015)
- [Jarrett 09] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y.: What is the best multi-stage architecture for object recognition?, in *ICCV*, pp. 2146–2153 (2009)
- [Jha 17] Jha, S. and Busso, C.: Challenges in head pose estimation of drivers in naturalistic recordings using existing tools, in *ITSC*, pp. 1–6 (2017)
- [Johnson 10] Johnson, S. and Everingham, M.: Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, in *BMVC*, pp. 12.1–12.11 (2010)
- [Kreiss 19] Kreiss, S., Bertoni, L., and Alahi, A.: Pifpaf: Composite fields for human pose estimation, in *CVPR*, pp. 11977–11986 (2019)

表 7 行動パターン別評価 (関節点有無)

手法	正解率 (頭部)	正解率 (首元)	正解率 (右手)	正解率 (左手)
前方注視	100.0%	96.3%	99.9%	99.9%
脇見	100.0%	91.4%	99.5%	99.5%
もたれ	98.1%	85.9%	91.5%	99.9%
ストレッチ	99.3%	87.6%	82.4%	82.3%
振り返り	95.9%	86.7%	99.5%	98.6%
眠気	100.0%	87.3%	99.5%	99.8%
スマホ操作	100.0%	61.3%	99.9%	99.9%
通話	99.9%	85.7%	91.4%	89.1%
カメラ撮影	99.1%	82.4%	89.0%	89.1%
読書	99.9%	63.8%	99.6%	99.7%
飲食	99.9%	81.3%	88.5%	95.0%
着替え	96.3%	82.2%	91.4%	91.5%
突っ伏し	85.1%	97.5%	91.1%	92.2%
パニック	96.1%	82.0%	90.3%	90.7%
居眠り	99.2%	85.9%	99.9%	99.9%
赤ちゃんを抱く	98.8%	70.1%	97.2%	95.0%

- [Lin 14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft coco: Common objects in context, in *ECCV*, pp. 740–755 (2014)
- [Lin 17] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal loss for dense object detection, in *ICCV*, pp. 2980–2988 (2017)
- [Ma 18] Ma, N., Zhang, X., Zheng, H.-T., and Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design, in *ECCV*, pp. 116–131 (2018)
- [Newell 16] Newell, A., Yang, K., and Deng, J.: Stacked hourglass networks for human pose estimation, in *ECCV*, pp. 483–499 (2016)
- [Ohn-Bar 14] Ohn-Bar, E., Martin, S., Tawari, A., and Trivedi, M. M.: Head, eye, and hand patterns for driver activity recognition, in *ICPR*, pp. 660–665 (2014)
- [Pishchulin 16] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., and Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation, in *CVPR*, pp. 4929–4937 (2016)
- [Sapp 13] Sapp, B. and Taskar, B.: MODEC: Multimodal Decomposable Models for Human Pose Estimation, in *CVPR* (2013)
- [Sun 18] Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y.: Integral human pose regression, in *ECCV*, pp. 529–545 (2018)
- [Sun 19] Sun, K., Xiao, B., Liu, D., and Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, in *CVPR* (2019)
- [Tompson 14] Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation, in *NIPS*, pp. 1799–1807 (2014)
- [Toshev 14] Toshev, A. and Szegeedy, C.: Deeppose: Human pose estimation via deep neural networks, in *CVPR*, pp. 1653–1660 (2014)
- [Wei 16] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y.: Convolutional pose machines, in *CVPR*, pp. 4724–4732 (2016)
- [Xiao 18] Xiao, B., Wu, H., and Wei, Y.: Simple baselines for human pose estimation and tracking, in *ECCV*, pp. 466–481 (2018)
- [Yang 17] Yang, W., Li, S., Ouyang, W., Li, H., and Wang, X.: Learning feature pyramids for human pose estimation, in *ICCV*, pp. 1281–1290 (2017)
- [国土交通省自動車局 18] 国土交通省自動車局 先進安全自動車推進検討会: ドライバー異常自動検知システム基本設計書 (2018)

〔担当委員: 吉川 友也〕

2020 年 5 月 12 日 受理

著者紹介

西行 健太

2010 年 奈良先端科学技術大学院大学博士前期課程修了。同年、株式会社メガチップス入社。2016 年 株式会社 ECC 入社。2017 年 オムロン株式会社入社、現在に至る。現在、中部大学博士後期課程 (社会人ドクター) に在学中。コンピュータビジョン、機械学習の研究に従事。

日向 匡史

2006 年 大阪大学大学院基礎工学研究科博士後期課程修了。同年、オムロン株式会社入社、現在に至る。画像認識の研究開発に従事。画像センシングシンポジウム オーディエンス賞 (2008 年)、同デモンストレーション賞 (2010 年、2016 年) 受賞。

田崎 博

2003 年 岡山大学大学院自然科学研究科機械システム工学専攻修了。同年、オムロン株式会社入社、現在に至る。人間-機械協調システムの研究開発に従事。

木下 航一

1998 年 神戸大学大学院自然科学研究科博士前期課程修了。同年、オムロン株式会社入社、現在に至る。2013 年 名古屋大学大学院情報科学研究科博士後期課程修了、博士 (情報科学)。主として画像認識の研究開発に従事。画像の認識・理解シンポジウム ベストインタラクティブセッション賞 (2008 年)。画像センシングシンポジウム優秀論文賞 (2008 年)。画像センシングシンポジウム高木賞 (2009 年) 受賞。

長谷川 友紀

1997 年 筑波大学第 3 学群情報学類卒業。同年、オムロン株式会社入社。2017 年 同社 画像センシング研究室 室長、現在に至る。電気学会産業応用部門 優秀論文発表賞 (2008 年) 受賞。

山下 隆義 (正会員)

2002 年 奈良先端科学技術大学院大学博士前期課程修了。同年、オムロン株式会社入社。2011 年 中部大学大学院博士後期課程修了 (社会人ドクター)。2014 年 中部大学講師。2017 年 中部大学准教授。人の理解に向けた動画処理、パターン認識・機械学習の研究に従事。画像センシングシンポジウム高木賞 (2009 年)、電子情報通信学会 情報・システムソサイエティ論文賞 (2013 年)、電子情報通信学会 PRMU 研究会研究奨励賞 (2013 年) 受賞。

藤吉 弘亘

1997 年 中部大学大学院博士後期課程修了。同年、米カーネギーメロン大学ロボット工学研究所 Postdoctoral Fellow。2000 年 中部大学工学部情報工学科講師。2004 年 中部大学准教授。2005 年 米カーネギーメロン大学ロボット工学研究所客員研究員 (2006 年)。2010 年 中部大学教授。2014 年 名古屋大学客員教授。計算機視覚、動画処理、パターン認識・理解の研究に従事。ロボカップ研究賞 (2005 年)、情報処理学会論文誌 CVIM 優秀論文賞 (2009 年)、情報処理学会山下記念研究賞 (2009 年)、画像センシングシンポジウム優秀学術賞 (2010、2013、2014 年)、電子情報通信学会 情報・システムソサイエティ論文賞 (2013 年) 他