ドライバ姿勢と動作のマルチタスク学習による高速かつ高精度なドライバ動作認識

Fast and Accurate Driver Action Recognition with Multi-Task Learning of Driver Pose and Action

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, action recognition, multi-task learning, pose estimation

Summary

Driver Action Recognition is a key component in driver monitoring systems, which is helpful for the safety management of commercial vehicles. Compared with traditional human action recognition tasks, driver action recognition is required to be fast and accurate on embedded systems. We propose a fast and accurate driver action recognition method that is composed of CNN based driver pose estimation and RNN based driver action recognition. We train our network model with multi-task learning includes localizing and detecting each body part of the driver, classifying state of each body part, and recognizing driver action at once. Our multi-task learning for the proposed model achieves a significant improvement compared to state-of-the-art human action recognition methods with limited computational resources. We also perform ablation study of our methods which composed of the driver pose localization, detection, and classification.

1. はじめに

保有台数1万台当たりの人身事故発生件数を車種別にみると、バスやタクシーなどの事業用乗用車が最も多い[交通20]。事業用乗用車の事故を減らすため、営業車5台以上を保有する事業者は安全運転管理者を選任する義務が生じる。営業車が増えるほど安全運転管理者の管理負担は増大するため、運転動作評価用ドライバモニタリングシステム(DMS)を活用し、管理負担を軽減することが期待されている。

従来の実用化されているDMSではドライバの顔向きや目開閉度からドライバが運転に集中しているかどうかを推定しているが、運転動作評価用DMSでは、ドライ

バの飲食など、運転中にドライバが取りうる多様な動作を認識することでより詳細な評価を行うことが求められる。本論文では、運転動作評価用DMSのためのドライバ動作認識を提案する。

DMSでは自動車の組込機器での動作が求められるため、サーバやデスクトップPCと比べて、消費リソースに限られる。近年提案されているConvolutional Neural Network(CNN)を用いた人物動作認識は高精度であるが、演算量などの消費リソースが大きいため、DMSにそのまま搭載することが難しい。本論文では、軽量な姿勢推定モデルをベースとした高精度で演算量の少ないドライバ動作認識を提案する。

CNN を用いた動作認識で用いられる Kinetics [Carreira 18, Kay 17] などのデータセットは、様々な種類の動作が含まれているが、前方注視や余所見などの運転動作の評価に役立つ動作は含まれていない。本論文では、ドライバの運転への復帰時間を基に、運転中に取りうるドライバ動作を 7 つの動作として定義し、それらの動作を含んだデータセットを構築する。また、そのデータセットを用いて提案手法の評価実験を行う。

CNN を用いた従来の人物動作認識の多くは、入力動画から直接人物の動作を推定する。これらの手法は高精度であるが、本論文ではドライバの動作だけでなく、ドライバの姿勢も同時に学習するマルチタスク学習を行うことで、演算量が制限された条件で従来の人物動作認識よりも高い精度を達成できることを示す。本論文では、ドライバ動作だけでなく、ドライバの関節がどこに映っているか（関節点座標）、ドライバの関節が画像中に映っているかどうか（関節点有無）、ドライバの関節がどのような状態にあるか（関節点状態）、の 3 つのドライバ姿勢を同時に学習するマルチタスク学習を提案する。

2. 関連研究

2.1 動作認識

人物動作認識はコンピュータビジョンの分野で活発に研究されており、近年は CNN を用いた高精度な手法が数多く提案されている。

Karpathy らは、動画に対して時系列データを結合する 3 つの Fusion model (Early-fusion, Late-fusion, Slow-fusion) を提案した [Karpathy 14]。Early-fusion では、複数フレームの画像を深さ方向に結合し、複数チャンネルの入力画像として扱う。Late-fusion では、複数フレームの画像を別々に CNN に与え、それらの出力である特徴マップを結合する。Slow-fusion は、Early-fusion と Late-fusion の中間に相当し、入力から出力に向かって、徐々に時間方向の情報を結合する。

画像から特徴抽出を行う CNN と時系列データを扱う Recurrent Neural Network (RNN) を組み合わせた動作認識が提案されている。Donahue らは、Long Short Term Memory (LSTM) [Hochreiter 97] を CNN と組み合わせた Long-term Recurrent Convolutional Networks (LRCN) を提案した [Donahue 15]。LSTM は長期的な時間依存を学習できる RNN であり、CNN と組み合わせることで、長時間の動作を認識できる。

Simonyan らは、時間情報と空間情報を並列に処理する 2-stream CNN を提案した [Simonyan 14]。2-stream CNN では、時間情報のオプティカルフロー画像と、空間情報の RGB カラー画像が入力となる。時間情報と空間情報を並列に扱うことで、高精度な動作認識を実現した。

空間方向に対して畳み込みを行う CNN を時間方向に拡張した 3D-CNN [Taylor 10, Tran 15] を動作認識に適用

した手法が提案されている。Feichtenhofer らは、3D の畳み込み層で、時間方向のスライドを小さくした Slow Pathway とスライドを大きくした Fast Pathway を組み合わせた SlowFast Networks を提案した [Feichtenhofer 19]。SlowFast Networks は、異なるスライドを利用することで、様々な時間解像度の特徴を抽出できる。

2.2 マルチタスク学習

2.1 節で言及した動作認識モデルは人物動作を高精度に認識できるが、マルチタスク学習を用いて人物姿勢も同時に学習することで、更なる高精度化が期待できる。マルチタスク学習 [Caruana 97] は、複数のタスクを学習することで、各タスクの精度を向上させる手法である。[Caruana 97] では、複数のタスクを同時に学習することで、各タスクで共通して有効な特徴が選ばれやすくなり、精度が向上すると言及されている。人物動作認識でも同様に、人物姿勢と共通した有用な特徴を抽出することで精度が向上することが期待できる。

Gkioxari らは物体検出やセマンティックセグメンテーションなどで使用される R-CNN [Girshick 14] を用いて、姿勢推定と動作認識のマルチタスク学習を提案した [Gkioxari 14]。[Gkioxari 14] では、人物動作だけでなく、人物の関節点座標も学習することで、動作認識の精度が向上することを示した。

本論文では演算量が制限された条件でドライバ動作認識の精度を向上させるため、ドライバの関節点座標だけでなく、関節点有無や関節点状態も同時に学習するマルチタスク学習を提案する。

3. 提案手法

本論文では、ドライバの姿勢と動作のマルチタスク学習を用いた高精度かつ演算量の少ないドライバ動作認識を提案する。提案手法のネットワークモデルの概要を図 1 に示す。

提案手法のネットワークモデルは、姿勢推定部と動作認識部に分かれる。姿勢推定部では CNN を用いて、ドライバの関節点があるか（関節点座標）、関節点が画像中に映っているか（関節点有無）、関節点がどのような状態にあるか（関節点状態）、の 3 つのドライバ姿勢を出力する。動作認識部では姿勢推定部が出力したドライバ姿勢と特徴マップを入力として、RNN を用いてドライバ動作を認識する。

3.1 運転動作の定義

Zhao らが提案したドライバ動作認識のデータセット Southeast University Driving-posture Dataset (SEU dataset) [Zhao 12] では、電話、食事、ブレーキ、運転、スマートフォンの操作、タバコの 6 動作が含まれる。SEU dataset の動作は、比較的容易に正常な運転に戻れる動作である。

図1 ネットワークモデル全体 (提案手法、入力画像 96×64 , 34ch) : () 内は畳み込み層では出力チャンネル数, FC 層では出力ノード数, Feature Map ではマップの解像度を示す. SNV2 の Conv(1x1) はカーネルサイズが 1x1 の畳み込みを示す. 各出力結果では, 出力結果の次元数を示す. FC は全結合層, BN は Batch Normalization 層を示す.

しかし、運転時は居眠りや発作など、意識を失うような深刻な状態が起こりうるため、これらの動作の認識が深刻な事故の抑制につながる。また、SEU dataset は手動運転時のドライバ動作を対象としているが、今後は自動運転車の急速な普及が想定されており [高度 20]、自動運転中のドライバ動作の認識も重要な課題である。

本論文では、SEU dataset で取り扱っている軽度な不注意状態に加えて、意識を失うような深刻な状態、さらに自動運転中にドライバが取りうる動作も検討の対象とする。そのため、本論文では安全性に問題がない状態から運転復帰が困難な状態までを幅広くカバーする 7 種類のドライバ動作（前方注視、余所見、眠気、物体把持、下向き、意識不明、パニック）を代表的な動作として取り上げる。

正常な運転状態は、前方を注視しており、手はハンドルを握っているか、何も持っていない状態である。本論文では、そのような正常な運転状態を示す動作として、“前方注視”を認識対象とする。SEU dataset のプレーキと運転の 2 動作は、本論文の“前方注視”に該当する。

正常な運転状態への復帰に少し時間がかかる危険な状態は、余所見や短時間の閉眼など前方を向いていない状態や、スマートフォンや飲食物など手で物体を掴んでいる状態である。本論文では、そのような危険な状態として、横や後ろを向いている“余所見”，うつらうつらしている“眠気”，手で物体を持っている“物体把持”，スマートフォンや本などを見ている“下向き”の 4 つの動作を認識対象とする。SEU dataset の電話、食事、スマー

トフォンの操作、タバコの 4 動作は、本論文の“物体把持”に含まれる。自動運転時にはカメラなど多様な物体を持っている可能性が高いため、本論文では特定の物体を対象としない“物体把持”を認識対象動作とする。

正常な運転状態への復帰が困難な状態は意識を失っていたり、パニックになっているような状態である。本論文では、そのような復帰が困難な状態として、眠っていたり、急病で意識を失っていたりする“意識不明”，蜂が入ってきており運転に集中できないような状態である“パニック”の 2 つの動作を認識対象とする。現在導入が進められている自動運転車では、一般的には起こりにくい居眠りなどの運転状態への復帰が困難な動作が起こる可能性が高い。また、完全な自動運転車の開発は難しく、今後も緊急時には手動運転に切り替わるような半自動運転が主流になると考えられるため、上記のような運転動作の評価は必要となると考え、本論文では動作認識の対象とする。

以上より、本論文では、前方注視、余所見、眠気、物体把持、下向き、意識不明、パニック、の 7 動作を認識するネットワークモデルを提案する。

3.2 姿勢推定部

姿勢推定部では関節点座標、関節点有無、関節点状態の 3 つのドライバ姿勢を出力する。本論文で用いる姿勢推定部を図 1 の上部に示す。姿勢推定部には、人物姿勢推定用ネットワークモデルである Hourglass Module [Newell 16] を用いる。Hourglass Module は高解像度から低解像度

情報を圧縮していくダウンサンプリング処理を行った後、低解像度から高解像度に変換していくアップサンプリング処理を行うことで、様々な解像度の人物姿勢を推定するネットワークモデルである。本論文では Hourglass Module に、ShuffleNet V2 [Ma 18] と Integral Regression [Sun 17] を導入することで、演算量の少ない姿勢推定を提案する。

§1 ShuffleNet V2

本論文では通常の畳み込み層の代わりに Ma らが提案した ShuffleNet V2 [Ma 18] を用いて、畳み込みの演算量を減らす。ShuffleNet V2 の構成を図 1 の最下部に示す。ShuffleNet V2 ではチャンネルを分割したのち、一方のみに畳み込みを行う。分割したチャンネルを結合する際に、チャンネルの順序を入れ替えることで畳み込みの回数を減らしつつ、全てのチャンネルに対して畳み込みを行う。ShuffleNet V2 では 1×1 のフィルタを用いた畳み込み、Depth-wise の畳み込み、再び 1×1 のフィルタを用いた畳み込みを順に行う。Depth-wise の畳み込みはチャンネル方向の畳み込みを行わず、空間方向に対してのみ畳み込みを行うため、通常の畳み込みよりも演算量が少ない。

§2 Integral Regression

多くの人物姿勢推定では、ヒートマップを出力し、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。ヒートマップの解像度を小さくすると演算量が減るが、関節点座標の量子化誤差が増える。Sun らはヒートマップの重心位置を関節点座標として出力する Integral Regression [Sun 17] を提案した。Integral Regression は、サブピクセル単位で関節点座標を推定するため、ヒートマップの解像度を小さくした際の量子化誤差を抑制できる。本論文では Integral Regression を用いて、量子化誤差を抑制しつつ、ヒートマップの解像度を小さくすることで、演算量を減らす。従来のヒートマップの最大値を用いた関節点座標の推定は、式 (1) 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、 I_k と J_k は k 番目の関節点の座標を示す。Integral Regression を用いたヒートマップの重心位置の推定は、式 (2) となる。 Ω は座標の集合を示す。Integral Regression で使用されるヒートマップ $\hat{H}_k$ は、式 (3) 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。

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

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

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

§3 関節点座標及び関節点有無の推定

本論文の姿勢推定部では、頭部、首元、右手、左手の 4 つの関節点座標を出力する。各関節点の座標は 2 次元、

関節点座標の次元数は 8 となる。また、DMS ではカメラとドライバ間の距離が狭く、関節点が画像中に映らないことが頻繁にあるため、提案手法では関節点が画像中に映るかどうか（関節点有無）も推定する。関節点有無はドライバが急病で倒れ、頭部が画面中に映らないような動作の認識に役立つ。関節点有無は画像中に映る場合に 1、映らない場合に 0 となる 2 値の出力とする。関節点有無の次元数は 4 となる。

§4 関節点状態の推定

高精度なドライバ動作認識には、ドライバの関節点座標や関節点有無に加えて、各関節点の詳細な状態が役立つ。飲み物を顔に近づける動作と、手で顔を搔く動作ではいずれの動作も頭部付近に手があるため、関節点座標と関節点有無だけでは区別が難しい。それらの動作を区別するためには、手で物体を把持しているかどうかといった関節点の詳細な状態を把握することが必要である。本論文では、[国土交通省自動車局 18] で示されるドライバ観測指標を基に、ドライバ動作認識に必要なドライバ状態として、右手状態、左手状態、目状態、顔向きピッチ、ヨー、ロールの 6 つの関節点状態を定義し、それらを推定するモデルを提案する。各関節点状態のクラスを表 1 に示す。右手・左手状態は上から順に、手が画面に映っていない、物体を持っている、何も持っていない、ハンドルを握っている、の 4 つとする。目状態は、目が画面に映っていない、前方を注視している、余所見をしている、目を閉じている、の 4 つとする。顔向きピッチは、顔が映っていない、下を向いている、少し下を向いている、前を向いている、少し上を向いている、上を向いている、の 6 つとする。顔向きヨーは、顔が映っていない、右後ろを向いている、右を向いている、右前を向いている、正面を向いている、左前を向いている、左を向いている、左後ろを向いている、の 8 つとする。顔向きロールは、顔が映っていない、 -90° 、 -45° 、 0° 、 45° 、 90° の 6 つとする。顔向きロールの角度は時計回りの方向が正の値、反時計回りの方向が負の値とする。

3.3 動作認識部

提案手法では、動作認識部のネットワークモデルとして、LSTM [Hochreiter 97] を用いた RNN を提案する。動作認識部を図 1 の右下部に示す。動作認識部では本論文で定義した 7 動作の認識結果を出力する。動作認識部の入力、姿勢推定部が出力する関節点座標、関節点有無、関節点状態の 3 つのドライバ姿勢と、Hourglass Module が出力する特徴マップとなる。関節点座標、関節点有無、関節点状態はそれぞれ、8、4、32 次元の特徴ベクトルとなる。また、Hourglass Module から出力される特徴マップは Global Average Pooling (GAP) を経て、68 次元の特徴ベクトルとなる。それらを結合した 112 次元の特徴ベクトルが、動作認識部の入力となる。動作認識部は全結合層により特徴抽出を行った後、LSTM 層で時系列の情

表 1 関節点状態

右手/左手	目	顔向きピッチ	顔向きヨー	顔向きロール
<ul style="list-style-type: none"> • Unknown • Hand-on • Hand-off • Handle 	<ul style="list-style-type: none"> • Unknown • Open • Side • Close 	<ul style="list-style-type: none"> • Unknown • Down • DownFront • Front • UpFront • Up 	<ul style="list-style-type: none"> • Unknown • RightBack • Right • RightFront • Front • LeftFront • Left • LeftBack 	<ul style="list-style-type: none"> • Unknown • -90° • -45° • 0° • 45° • 90°

報を処理する．最後に全結合層を用いて，7 動作の認識結果を出力する．また，学習を安定させるため，フレーム単位での動作認識結果も出力する．姿勢推定部で得られた特徴ベクトルを入力として，1 層の全結合層を用いてフレーム単位の動作認識結果を出力する．フレーム単位の動作認識結果は，学習にのみ使用され，テスト時の最終的な出力結果は LSTM の出力結果とする．

3.4 学 習

提案手法では入力画像 1 枚に対して，姿勢推定部の関節点座標，関節点有無，関節点状態の 3 つのドライバ姿勢と，動作認識部のドライバ動作の計 4 つの損失を計算し，ネットワークモデルを学習する．

関節点座標を示す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．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．

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

提案手法では微分可能な Integral Regression を用いるため，関節点座標に対しても損失を計算する．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，式 (8) となる．クラス間のデータの偏りに対応するため，Focal loss [Lin 17] を用いる． x は関節点有無の 2 値出力， 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．

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

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

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

関節点状態の損失関数 L_S も，クラス間のデータの偏りに対応するため，Focal loss を用いた式 (8) を用いる．

動作認識結果はフレーム単位の認識結果と，LSTM による時間的な成分を考慮した認識結果の 2 つとなる．フレーム単位の動作認識結果の損失関数 L_F ，時間的な成分を考慮した認識結果の損失関数 L_A は式 (10) となる．いずれの損失関数も Softmax Cross Entropy を用いる．

$$y_c = \frac{e^{z_c}}{\sum_{d=1}^C e^{z_d}} \quad (9)$$

$$L_A = \sum_{c=1}^C t_c \log(y_c) \quad (10)$$

提案手法のネットワークモデルは，これらの損失をあわせた式 (11) を用いて学習する． w は各損失を調整するための重みである．本論文では，全ての損失を同等に扱うため，重み w を全て 1.0 とする．

$$L = w_H L_H + w_C L_C + w_D L_D + w_S L_S + w_F L_F + w_A L_A \quad (11)$$

3.5 ドライバ動作認識用データセット

動作認識用の評価データセットとして，Kinetics, Charades, AVA dataset [Kay 17, Carreira 18, Sigurdsson 16, Gu 18] がある．これらのデータセットは，可視光カメラで撮影された映像である．DMS では昼だけでなく，夜間もドライバを撮影する必要があるため，可視光カメラではなく，近赤外線カメラを用いることが多い．また，データセットに含まれる動作も DMS での利用を想定していない．そのため，本論文では，DMS での利用を想定した近赤外線カメラで撮影したドライバ動作認識用データセットを構築した．

4. 実 験

提案手法の有効性を確認するため、既存手法と精度を比較する。DMS では一般的な人物動作認識よりも、消費リソースが限られるため、提案手法と既存手法の演算量が同程度になるようパラメータ数を調整し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、提案手法の 3 つのドライバ姿勢を除いたモデルで精度を比較する。最後に提案手法で推定したドライバ姿勢と動作を示し、ドライバ姿勢が動作認識の誤認識の要因を解析するのに役立つことを示す。

4.1 実験データ

本論文では運転中にドライバが取りうる 7 つの動作を近赤外線カメラで撮影した。100 人の被験者を撮影し、50 人分を学習用、残り 50 人分を評価用として用いる。7 つの動作は、前方注視、余所見、眠気、物体把持、下向き、意識不明、パニック、である。撮影した動画のフレームレートは 10fps、解像度は 752×480 である。演算量を減らすため、撮影した動画を 96×64 に縮小し、データセットを構築した。1 動画当たりの撮影時間は 30 秒程度であり、各動作を被験者一人当たり 4~12 回程度撮影した。動画は合計で 6353 本となる。

提案手法はドライバ姿勢と動作のマルチタスク学習を行うため、ドライバ姿勢用の学習及び評価用データも構築した。ドライバ動作の撮影動画と同じ映像を使用し、頭部、首元、右手、左手の 4 つの関節点座標と関節点有無のアノテーションを行った。更に、関節点状態として、右手状態、左手状態、目状態、顔向きのピッチ、ヨー、ロールの 6 つの状態のアノテーションを行った。

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

入力画像は 96×64 に縮小したのち、Global Contrast Normalization [Jarrett 09] による正規化を行う。提案手法の姿勢推定部で出力するヒートマップのサイズは 24×16 とする。本実験で用いる 3D-CNN と Fusion-model では、複数フレームの画像が入力として必要なため、全て 32 フレームを入力として用いる。また、2-stream CNN のオプティカルフロー画像は 31 フレームを入力として用いる。

§2 ネットワーク

本論文では、比較対象の既存手法を、CNN と LSTM を用いた手法として LRCN [Donahue 15]、CNN と時間方向の結合を用いた手法として Fusion model (Early-fusion, Late-fusion, Slow-fusion) [Karpathy 14]、3D-CNN として SlowFast [Feichtenhofer 19]、2-stream CNN として [Simonyan 14] とする。本研究で組込を想定している車載情報端末向け統合 SoC は、CPU に Arm Cortex-A9 533MHz を利用している。該当 CPU の演算性能は 8.5GFLOPs である。複数のシステムの演算リソースを考慮し全演算

性能の 20% を本提案手法に利用できると仮定した場合、フレームレート 30FPS における提案手法の利用可能最大演算性能は 56MFLOPs となる。従って、各手法は演算量が 50MFLOPs、25MFLOPs 程度になるようパラメータ数を調整する。各モデルの消費リソースを表 2 に示す。

先行研究で提案されている LRCN や Fusion model、2-stream CNN では、学習を安定させるため、カラー画像の ImageNet [Deng 09] を用いて事前学習を行っている。しかし、本実験では近赤外カメラで撮影したグレースケール画像のドライバ動作認識を対象としているため、ImageNet による事前学習は行わない。本実験では、学習を安定させるため、Batch Normalization 層 [Ioffe 15] を導入する。

LRCN, Fusion model, 2-stream CNN は演算量を調整するため、3 層の畳み込み層を用いる。LRCN は畳み込み層の出力チャンネル数を 14, 30, 48 に減らした 50MFLOPs のモデルと、12, 20, 30 に減らした 25MFLOPs のモデルを用いる。Early-fusion モデルは畳み込み層の出力チャンネル数を 16, 16, 32 に減らした 50MFLOPs のモデルと、8, 14, 18 に減らした 25MFLOPs のモデルを用いる。Late-fusion モデルは畳み込み層の出力チャンネル数を 12, 20, 32 に減らした 50MFLOPs のモデルと、8, 16, 18 に減らした 25MFLOPs のモデルを用いる。Slow-fusion モデルは畳み込み層の出力チャンネル数を 8, 12, 20 に減らした 50MFLOPs のモデルと、4, 10, 16 に減らした 25MFLOPs のモデルを用いる。2-stream CNN は画像を入力とする CNN とオプティカルフロー画像を入力とする CNN のいずれも畳み込み層の出力チャンネル数を 8, 12, 22 に減らした 50MFLOPs のモデルと、4, 8, 18 に減らした 25MFLOPs のモデルを用いる。2 層目以降の畳み込み層では、畳み込み層の後に 2×2 の Average Pooling 層を用いる。また、これらのモデルでは 3 層の畳み込み層の後に、2 層の全結合層をつなげる。全結合層の出力ノード数は 256, 7 とする。SlowFast は [Feichtenhofer 19] で提案されているモデルと同様に、4 つの residual stage を用いる。各 stage の層数は 3, 4, 6, 3 とする。また、本実験では入力画像の解像度が小さいため、畳み込みのカーネルサイズを 3 以下とする。各ステージの出力チャンネル数は 4, 8, 16, 32 の 50MFLOPs のモデルと、1, 2, 4, 8 の 25MFLOPs のモデルを用いる。SlowFast は大きな stride や、1×1 のカーネルを用いた畳み込み、下位層に Global Average Pooling があるため、層数や出力チャンネル数が比較的多い場合でも、消費リソースが少ない。提案手法の 50MFLOPs モデルのパラメータは図 1 で示す通りである。25MFLOPs で用いるモデルでは、Hourglass Module の ShuffleNet V2 の最初の出力チャンネル数を 34ch から 22ch に減らし、以降の層のチャンネル数も 44ch, 88ch, 176ch に減らす。

§3 ハイパーパラメータ

学習率は 0.001 とし、300iteration までに徐々に 0.004 まで上昇させる。その後 5000iteration 毎に学習率を半減

表 2 ネットワークの演算量及びパラメータ数

手法	MFLOPs	パラメータ数 [M]
50 MFLOPs		
LRCN [Donahue 15] 14-30-48ch	49.5	5.73
Early-fusion [Karpathy 14] 16-16-32ch	52.7	3.16
Late-fusion [Karpathy 14] 12-20-32ch	51.9	6.30
Slow-fusion [Karpathy 14] 8-12-20ch	51.1	1.98
SlowFast [Feichtenhofer 19] 4-8-16-32ch	55.9	0.05
2-stream [Simonyan 14] 8-12-22ch	50.1	4.34
提案手法, 34ch	47.20	0.98
25 MFLOPs		
LRCN [Donahue 15] 12-20-30ch	25.7	3.62
Early-fusion [Karpathy 14] 8-14-18ch	25.6	1.78
Late-fusion [Karpathy 14] 8-16-18ch	26.5	3.55
Slow-fusion [Karpathy 14] 4-10-16ch	25.5	1.58
SlowFast [Feichtenhofer 19] 1-2-4-8ch	28.1	0.02
2-stream [Simonyan 14] 4-8-18ch	25.0	3.55
提案手法, 22ch	25.3	0.75

させ、30000iteration で学習終了とする。学習率を徐々に上昇させた後に、減少させる方法は Goyal らの Gradual Warmup [Goyal 17] を参考にした。重み減衰は 0.0001 とし、重みのクリッピングは 5.0 とする。また、バッチサイズは 32 とし、4 つの GPU を用いて学習する。学習時には RNN の状態を 1iteration 毎に 25% の確率でリセットする。

§4 学習・評価方法の詳細

各動作の動画数が同程度になるように over sampling を行う。また、Data Augmentation として、下記 4 つの処理を実施する。

- ランダムにデータ長の調整（動画のフレーム数を 1～500 フレームの範囲でランダムにカット）
- ランダムに平行移動（画像の高さ・幅の $\pm 25\%$ ）
- ランダムに拡大縮小（0.8～1.75 倍）
- 50% の確率で左右反転

本実験では動作認識の評価指標として、正解率 (Accuracy) を用いる。正解率は正解したフレーム数を合計フレーム数で除算した値となる。また、提案手法では関節点座標、関節点有無、関節点状態を出力するため、それらも評価する。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 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。関節点状態はクラス数に偏りがあるため、評価指標に正解率のマクロ平均（マクロ正解率）を用いる。マクロ正解率はクラス毎に正解率を算出したのち、全クラスの正解率を平均した値となる。

4.3 精度比較

提案手法と既存手法について、動作認識の精度を比較した結果を表 3 に示す。また、提案手法の各ドライバ姿勢の精度を表 4 に示す。50MFLOPs, 25MFLOPs のいずれのモデルも、既存手法よりも提案手法の方が精度が高い。また、25MFLOPs のモデルでは、既存手法で最も精度の高い SlowFast よりも提案手法の精度が 5% 以上高い。従って、提案手法は、演算量を大幅に削減した時のドライバ動作認識の精度低下を抑制できる。2-stream CNN は入力画像の解像度が小さく、動きの特徴を上手くとらえられなかったため、その他の手法よりも低い精度になったと考えられる。

表 3 精度比較（動作推定）

手法	正解率
50 MFLOPs	
LRCN [Donahue 15], 14-30-48ch	68.3%
Early-fusion [Karpathy 14], 16-16-32ch	70.6%
Late-fusion [Karpathy 14], 12-20-32ch	69.1%
Slow-fusion [Karpathy 14], 8-12-20ch	70.7%
2-stream [Simonyan 14], 8-12-22ch	42.3%
SlowFast [Feichtenhofer 19], 4-8-16-32ch	83.52%
提案手法, 34ch	83.84%
25 MFLOPs	
LRCN [Donahue 15], 12-20-30ch	67.9%
Early-fusion [Karpathy 14], 8-14-18ch	66.8%
Late-fusion [Karpathy 14], 8-16-18ch	66.2%
Slow-fusion [Karpathy 14], 4-10-16ch	65.6%
2-stream [Simonyan 14], 4-8-18ch	41.8%
SlowFast [Feichtenhofer 19], 1-2-4-8ch	76.25%
提案手法, 22ch	82.17%

4.4 Ablation Study

本実験では Ablation Study として、提案手法から 3 つのドライバ姿勢を除いた実験を行い、各ドライバ姿勢が動作認識精度にどの程度寄与しているかを確認する。表 5 に Ablation Study の実験結果を示す。提案手法と、動作認識と関節点座標の推定を行うモデル (Only-Position)、動作認識と関節点有無の推定を行うモデル (Only-Detection)、動作認識と関節点状態の推定を行うモデル (Only-State)、提案手法から姿勢推定部（関節点座標、関節点有無、関節点状態の全て）を除いたモデル (No-Pose) を比較する。各モデルは演算量が 50MFLOPs, 25MFLOPs 程度になるようパラメータ数を調整する。No-Pose では、動作認識部の入力 Hourglass Module が出力した特徴マップのみとなる。また、No-Pose では関節点座標、関節点有無、関節点状態を出力するために必要な畳み込み層や全結合層などを取り除く。

提案手法から姿勢推定部を除いたモデルでは、精度が 4% 以上低下する。従って、ドライバの姿勢と動作のマルチタスク学習は、演算量を減らした際のドライバ動作認

表 4 姿勢推定精度 (関節点座標, 関節点有無, 関節点状態)

手法	関節点座標 PCK-6px				関節点有無 mAP				関節点状態 マクロ正解率					
	頭部	首元	右手	左手	頭部	首元	右手	左手	右手	左手	目	ピッチ	ヨー	ロール
提案手法, 34ch	91.8%	94.0%	89.1%	76.6%	98.3%	73.9%	92.4%	91.1%	85.4%	83.6%	71.6%	74.7%	79.6%	81.3%
提案手法, 22ch	90.8%	93.2%	85.7%	73.2%	98.2%	73.1%	91.7%	90.3%	82.6%	82.0%	70.5%	72.7%	79.2%	78.0%

識の精度低下を抑制する。提案手法を除いた比較では、関節点状態の推定を行う Only-State が最も高精度である。ドライバ動作認識では、既存手法の [Gkioxari 14] のような関節点座標と動作のマルチタスク学習よりも関節点状態と動作のマルチタスク学習の方が高精度であることを示している。

表 5 Ablation Study

手法	正解率	MFLOPs	パラメータ数 [M]
50 MFLOPs			
提案手法, 34ch	83.84%	47.20	0.98
Only-Position, 34ch	81.91%	46.86	0.64
Only-Detection, 34ch	80.30%	46.99	0.77
Only-State, 38ch	81.99%	49.76	0.93
No-Pose, 40ch	76.42%	50.53	0.73
25 MFLOPs			
提案手法, 22ch	82.17%	25.3	0.75
Only-Position, 22ch	80.82%	24.92	0.41
Only-Detection, 22ch	78.17%	25.05	0.54
Only-State, 26ch	81.70%	25.01	0.67
No-Pose, 28ch	77.23%	25.6	0.45

4.5 実験結果画像

提案手法のドライバ姿勢と動作の出力結果例を図 2 に示す。上段、中段が正しく動作を認識できた結果、下段が誤った認識結果となる。

例 3 (手元のスマートフォンを操作しているシーン)、例 5 (カメラで撮影しているシーン) では、関節点座標は正解座標とのズレが少しあるが、ほとんどのシーンで正確に推定している。動作や関節点状態についても、正確に推定している。例 6 (急病で意識を失っているシーン) では、関節点状態の目状態が Side となっている。また、顔向きヨーも RightFront と誤認識が起きている。

例 9 (前方を注視しているシーン) では少し上を向いているため、顔向きロールが 45° と誤認識しており、目状態も Close と誤認識している。例 10 (パニックのシーン) では、瞬間的に前方を向いたシーンがあり、一時的に前方注視と誤認識している。例 11 (余所見のシーン) では、目状態を Close、顔向きピッチを UpFront と誤認識している。例 12 (下向きのシーン) では、右下に映っている人形の一部をドライバの手と誤認識したため、関節点座標や左手状態を誤認識している。

4.6 考察

図 2 下段の誤認識結果を確認すると、誤認識の要因を推測することができる。例 9 (前方を注視しているシーン) では、ドライバが上を向いており、人目でも目を開けているかどうか分からない状態であるため、目状態を Close と誤認識したと考えられる。目を閉じている状態は眠気の特徴であるため、行動を眠気と誤認識したと考えられる。例 11 (パニックのシーン) では、メガネのフレームにより目の状態を確認することが難しく、目状態を Close と誤認識し、結果として行動もパニックと誤認識したと考えられる。例 12 (下向きのシーン) では、右下の人形をドライバの手と誤認識しているため、手を振り回しているパニックのシーンと誤認識したと考えられる。このようにドライバ動作の誤認識が生じた場合、中間出力結果であるドライバ姿勢が誤認識の要因の解析に役立つ。そのため、提案手法のドライバ姿勢と動作のマルチタスク学習は精度向上だけでなく、誤認識時の解析にも有用である。

5. まとめ

本論文では運転動作評価のためのドライバ認識を提案した。提案手法ではドライバ姿勢と動作のマルチタスク学習を行うことで既存手法と比べて演算量を減らした際の精度低下を抑制できることを示した。本論文ではドライビングシミュレータのデータを用いて評価したが、今後は実車での評価や、より多くの被験者で交差検定を実施することで、更なる実用性向上に取り組む。

◇ 参考文献 ◇

- [Carreira 18] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A.: A short note about kinetics-600, *arXiv preprint arXiv:1808.01340* (2018)
- [Caruana 97] Caruana, R.: Multitask learning, *Machine Learning*, Vol. 28, No. 1, pp. 41–75 (1997)
- [Deng 09] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, in *CVPR*, pp. 248–255 (2009)
- [Donahue 15] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description, in *CVPR*, pp. 2625–2634 (2015)
- [Feichtenhofer 19] Feichtenhofer, C., Fan, H., Malik, J., and He, K.: Slowfast networks for video recognition, in *ICCV*, pp. 6202–6211 (2019)
- [Girshick 14] Girshick, R., Donahue, J., Darrell, T., and Malik, J.: Rich feature hierarchies for accurate object detection and semantic

種類	正解	推定	種類	正解	推定	種類	正解	推定	種類	正解	推定
行動	前方注視	前方注視	行動	余所見	余所見	行動	下向き	下向き	行動	眠気	眠気
右手状態	Unknown	Unknown	右手状態	Unknown	Unknown	右手状態	Unknown	Unknown	右手状態	Unknown	Unknown
左手状態	Unknown	Unknown	左手状態	Unknown	Unknown	左手状態	Unknown	Unknown	左手状態	Unknown	Unknown
目状態	Open	Open	目状態	Side	Side	目状態	Side	Side	目状態	Close	Close
ピッチ	Front	Front	ピッチ	Front	Front	ピッチ	DownFront	DownFront	ピッチ	Front	Front
ヨー	Front	Front	ヨー	Right	Right	ヨー	Front	Front	ヨー	Front	Front
ロール	0°	0°	ロール	0°	0°	ロール	0°	0°	ロール	0°	0°

種類	正解	推定	種類	正解	推定	種類	正解	推定	種類	正解	推定
行動	物体把持	物体把持	行動	意識不明	意識不明	行動	パニック	パニック	行動	意識不明	意識不明
右手状態	Hand-on	Hand-on	右手状態	Unknown	Unknown	右手状態	Unknown	Unknown	右手状態	Unknown	Unknown
左手状態	Hand-on	Hand-on	左手状態	Unknown	Unknown	左手状態	Hand-off	Hand-off	左手状態	Unknown	Unknown
目状態	Unknown	Unknown	目状態	Unknown	Side	目状態	Side	Side	目状態	Close	Close
ピッチ	Front	Front	ピッチ	Up	UpFront	ピッチ	Front	Front	ピッチ	Front	Front
ヨー	Front	Front	ヨー	LeftBack	RightFront	ヨー	LeftFront	LeftFront	ヨー	Front	RightFront
ロール	0°	0°	ロール	90°	Unknown	ロール	0°	45°	ロール	-45°	-45°

種類	正解	推定	種類	正解	推定	種類	正解	推定	種類	正解	推定
行動	前方注視	眠気	行動	パニック	前方注視	行動	余所見	パニック	行動	下向き	パニック
右手状態	Unknown	Unknown	右手状態	Unknown	Unknown	右手状態	Unknown	Unknown	右手状態	Unknown	Unknown
左手状態	Unknown	Unknown	左手状態	Unknown	Unknown	左手状態	Unknown	Unknown	左手状態	Unknown	Hand-off
目状態	Open	Close	目状態	Open	Open	目状態	Side	Close	目状態	Side	Close
ピッチ	UpFront	UpFront	ピッチ	Front	Front	ピッチ	Front	UpFront	ピッチ	Front	DownFront
ヨー	Front	Front	ヨー	Front	Front	ヨー	Right	Right	ヨー	LeftFront	LeftFront
ロール	0°	45°	ロール	0°	0°	ロール	0°	0°	ロール	0°	-45°

図2 実験結果画像:青色が頭部中心, 緑色が首元, 赤色が右手中心, 水色が左手中心の関節点座標を示す。○は推定結果, ×は正解座標を示す。誤った推定結果を赤字で示す。上段と中段は正しく動作を認識できたシーン, 下段は誤って認識したシーンとなる。上段左から順に, “正面を向いている”, “余所見”, “手元のスマートフォンを操作”, “うつらうつら”となる。中段左から順に, “カメラで写真撮影”, “急病で意識を失い, 前方に倒れている”, “蜂が車内に入り, パニック”, “眠っている”のシーンとなる。下段左から順に, “少し上を向きながら前方を注視”, “蜂が車内に入り, パニック”, “余所見”, “下を向いている”のシーンとなる。

- segmentation, in *CVPR*, pp. 580–587 (2014)
- [Gkioxari 14] Gkioxari, G., Hariharan, B., Girshick, R., and Malik, J.: R-cnns for pose estimation and action detection, *arXiv preprint arXiv:1406.5212* (2014)
- [Goyal 17] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour, *arXiv preprint arXiv:1706.02677* (2017)
- [Gu 18] Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions, in *CVPR*, pp. 6047–6056 (2018)
- [Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997)
- [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)
- [Karpathy 14] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L.: Large-scale video classification with convolutional neural networks, in *CVPR*, pp. 1725–1732 (2014)
- [Kay 17] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset, *arXiv preprint arXiv:1705.06950* (2017)
- [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)
- [Sigurdsson 16] Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding, in *ECCV*, pp. 510–526 (2016)
- [Simonyan 14] Simonyan, K. and Zisserman, A.: Two-stream convo-

- lutional networks for action recognition in videos, in *NeurIPS*, pp. 568–576 (2014)
- [Sun 17] Sun, X., Xiao, B., Liang, S., and Wei, Y.: Integral human pose regression, *arXiv preprint arXiv:1711.08229* (2017)
- [Taylor 10] Taylor, G. W., Fergus, R., LeCun, Y., and Bregler, C.: Convolutional learning of spatio-temporal features, in *ECCV*, pp. 140–153 (2010)
- [Tran 15] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M.: Learning spatiotemporal features with 3d convolutional networks, in *ICCV*, pp. 4489–4497 (2015)
- [Zhao 12] Zhao, C., Zhang, B., He, J., and Lian, J.: Recognition of driving postures by contourlet transform and random forests, *IET Intelligent Transport Systems*, Vol. 6, No. 2, pp. 161–168 (2012)
- [交通 20] 交通事故総合分析センター: 交通統計 平成 30 年版 (2020)
- [高度 20] 高度情報通信ネットワーク社会推進戦略本部・官民データ活用推進戦略会議: 官民 ITS 構想・ロードマップ 2020 (2020)
- [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: 村田 真悟〕

2020 年 9 月 28 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。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 弘亙

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。