

# 危険運転シーン抽出のためのマルチモーダル深層学習技術

丹野 良介<sup>†</sup> 小澤 暖<sup>†</sup> 伊藤 浩二<sup>†</sup>

<sup>†</sup> NTT コミュニケーションズ株式会社 〒 105-0023 東京都港区芝浦

E-mail: †{r.tanno,d.ozawa,kj.ito}@ntt.com

あらまし 近年のドライブレコーダーの急速な普及に伴い、運転ドライバーの安全運転意識の改善や危険運転への気付きを促進するといった、安全運転指導のためにドライブレコーダー映像を活用するといった例が多く存在する。しかし、そのためには記録された大量のドライブレコーダーの映像の中から、危険運転を抽出し、分類をする必要があり、その作業には多くの時間を要するといった問題があった。またそれらの作業は専任のスタッフが行うが、事故や危険なシーンなどのセンセーショナルな内容が映像中に含まれており、長時間の作業は困難であることから、大量の映像を短時間で正確に行うためにも、AIで危険運転シーンの自動検知を実現することが求められている。本研究では、日本カーソリューションズ株式会社様から提供頂いたドライブレコーダーデータを用いて、映像やセンサデータ、音声情報からなる時系列マルチモーダルデータを抽出し、それらの特徴量を組合せたマルチモーダル深層学習を利用した危険運転（ヒヤリハットや事故）の検知を行った研究について報告する。

キーワード マルチモーダル、深層学習、ドライブレコーダー

## 1 はじめに

近年、危険運転や煽り運転に遭遇し、巻き込まれ事故の発生件数がドライブレコーダーの普及とともに表面化し始め、大きな社会問題となってきている。ドライブレコーダー映像は、交通事故時における捜査や過失割合の判断に利用される他、あおり運転や車上荒らしなどの抑止力も期待できる。一方で、運転ドライバーの安全運転意識の改善や危険運転への気付きを促進するといった、安全運転指導のためにドライブレコーダー映像を活用するといった例も多く存在する。

しかし、そのためには記録された大量のドライブレコーダーの映像の中から、危険運転を抽出し、分類をする必要があり、その作業には多くの時間を要するといった問題があった。またそれらの作業は専任のスタッフが行うが、事故や危険なシーンなどのセンセーショナルな内容が映像中に含まれており、長時間の作業は困難であることから、大量の映像を短時間で正確に行うためにも、AIで危険運転シーンの自動検知を実現することが求められている。

本研究では日本カーソリューションズ株式会社様から提供して頂いたドライブレコーダーデータを用いて、映像やセンサデータ、音声情報からなる時系列マルチモーダルデータを抽出し、それらの特徴量を組合せたマルチモーダル深層学習を利用した危険運転（ヒヤリハット、事故）の検知を行った研究について報告する。

## 2 関連研究

車載映像に関する研究は自動運転の技術の発展と共に増加し

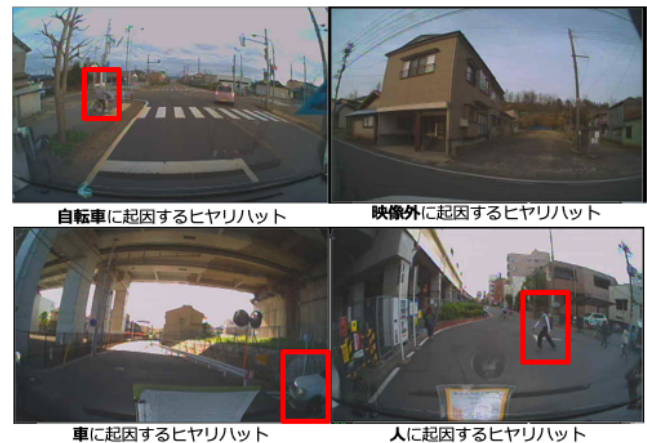


図1 様々なヒヤリハット映像の例

てる。運転手が取る行動である運転モデルを利用した研究としては Xu らの研究がある [1]。この研究では end-to-end に運転モデルを FCN-LSTM で学習することで、次を取る行動（右折、停止など）を予測する。また本研究と同様に危険運転や交通事故を扱う研究として [2] や [3] などがある。また、映像だけではなくセンサ情報を利用したヒヤリハット分類分析に関する研究として [4] などが存在する。

一方で、映像センサ単体だけではなく各モーダル情報の組合せを利用した研究もある。山本ら [5] は映像に加えてセンサ情報を組合せたヒヤリハット検出手法を提案し、センサか映像いずれかの情報を欠損させた手法や、時系列を考慮しない手法に比べて高い検出性能を示すことを明らかにした。しかし、一般的な車両に取り付けられているドライブレコーダーは車両の前方映像のみを記録するため、図1中の「映像外に起因するヒヤリハット」のように、特に後方や横方向といった危険運転の因

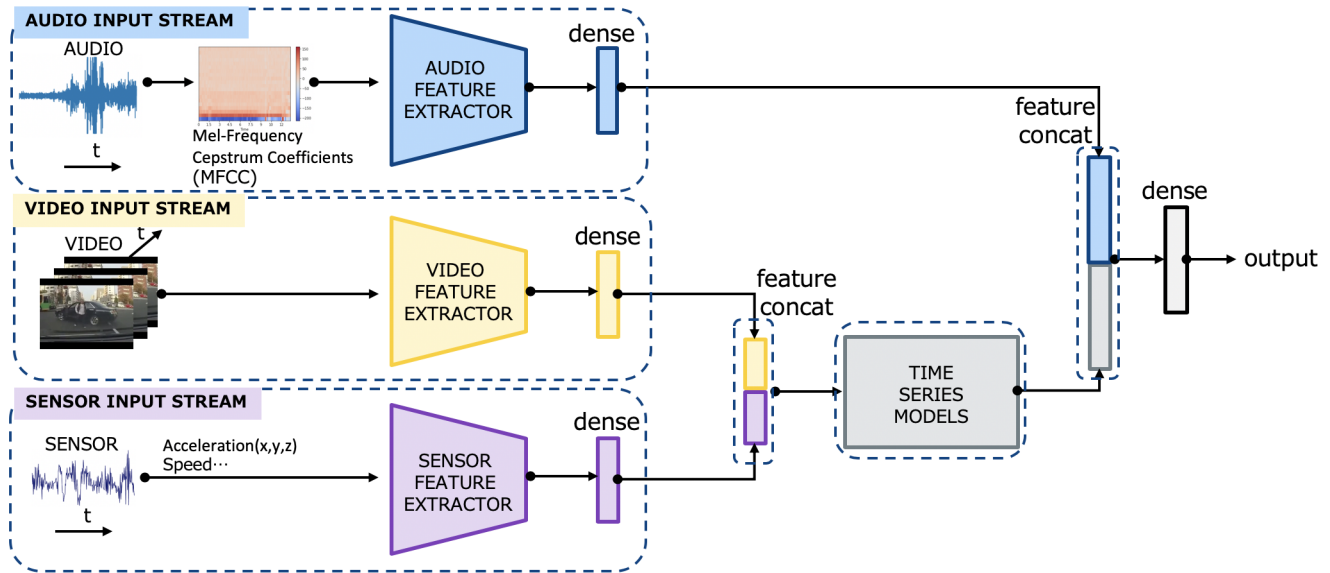


図2 今回実験に用いたネットワーク構成の概略図

子が存在する場合に発生するヒヤリハット事象に対して、危険運転シーンの判別は困難であると考えられる。

そこで、本研究では、映像とセンサに加え、環境音や車内の会話といった音声情報も組合せて学習することで、危険運転シーンの検出精度の向上を目指す。

### 3 提案手法

今回、実験に用いたネットワーク構造の概略を図2に示す。本研究で提案するマルチモーダル深層学習ネットワークへの入力にはドライブレコーダーに記録されている映像・センサ・音声の3つのモーダル情報から構成されるため、用いるニューラルネットワークも3つのストリームをもつ構成とした。また、各入力ストリームから得られる複数の特徴量のconcat方式は、入力の時点で複数のデータを深さ方向にconcatし、複数の特徴を1つのデータとして扱うEarly Fusionの形式をとるのではなく、複数のデータを単独でそれぞれのニューラルネットワークで学習し、最終的に最後のFC層で特徴をconcatするLate Fusionの形式を取る。

まず、映像を入力とするネットワークについては、ドライブレコーダーに記録されている映像から数十フレーム分切り出した画像をネットワークの入力とした。一般的に、動画の特徴抽出の手法は大きく分けて2つあり、1つ目は映像をあらかじめ複数枚の画像に分割する手法、2つ目は動画を直接入力する方法である。この時、後者はニューラルネットワークの入力に動画を直接用いるため、フレーム間の情報を考慮することが可能であるが、1つの動画を一度に読み込む必要があるなど、学習のためのコストが非常に大きい。一方で、前者の映像をフレーム分割する方法の場合は、既存の画像認識の手法を容易に

流用可能であること、また、分割した単一の画像をネットワークの入力とすることが多く、学習コストが小さいなどのメリットがあるため、今回は前者の方法を採用した。

センサ情報に関しては $x, y, z$ 方向の加速度と速度から成る4次元のデータを入力とするが、それぞれの次元に対して、平均0、標準偏差1になるように正規化を行ったデータをセンサ用ネットワークの入力とした。

音声については、音声認識の特徴量として良く用いられるメル周波数ケプストラム係数(MFCC: Mel-Frequency Cepstrum Coefficients)を抽出する前処理を施し、求めたMFCC特徴量を音声用ネットワークの入力とした。

映像・センサ・音声の各モーダル情報から特徴量を抽出後、RNNやLSTMなどから構成されるネットワークにおいて時系列モデリングを行い、最終的に危険運転ラベルの出力が得られるように学習をする。

## 4 実験

### 4.1 データセット

各データには映像の他に、 $x, y, z$ 方向の加速度や速度といったセンサ情報、環境音や車内の会話から成る音声が含まれており、1つのドライブレコーダーあたり約15秒程の動画で構成され、イベント検出時の映像を基にして、「事故(Accident)」、「ヒヤリハット(Near Miss)」、「正常(No-Near-Miss)」の中のいずれかのラベルが各動画にアノテーションされている。

今回の実験で学習および評価に用いたラベルとデータ数は表1の通りであり、各ラベルのデータを学習用8割、評価用2割として学習および評価を行った。

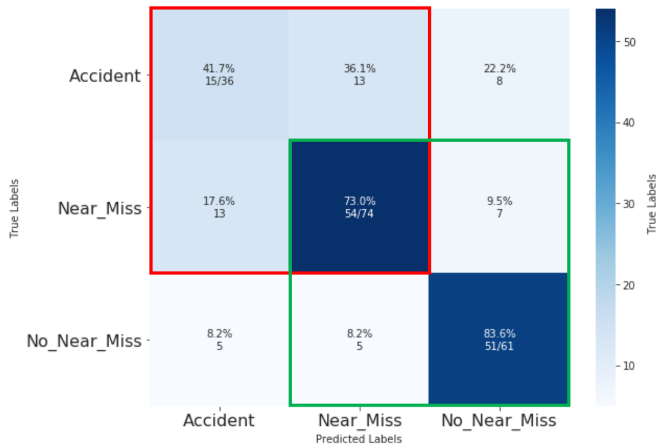


図 3 比較手法 [5] の混合行列 (映像+センサ)

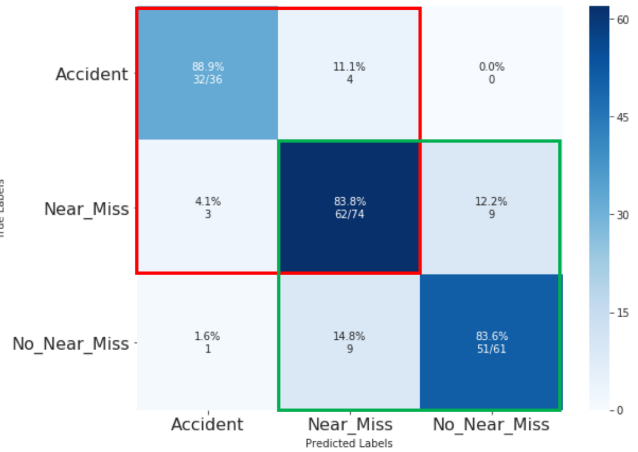


図 4 提案手法の混合行列 (映像+センサ+音声)

表 1 学習/評価用データ

	事故	ヒヤリハット	正常
データ数	190 件	364 件	298 件

表 2 比較手法 [5](映像+センサ)

	Precision	Recall	F1-Score
Accident	0.51	0.50	0.51
s Near-Miss	0.76	0.73	0.74
No-Near-Miss	0.78	0.72	0.72
Avg/Total	0.72	0.72	0.72

## 4.2 結果と考察

各手法により学習したモデルを評価用のデータに適用し算出した混合行列を図 3, 図 4 に示す。「事故 (Accident)」と「ヒヤリハット (Near-Miss)」を誤認識 (図中の赤囲み部分) していた従来手法 (図 3) と比較して提案手法 (図 4) では改善されていることがわかる。例えば, 低速度での衝突による事故の事故の場合, 映像とセンサのみでは判別しづらい事象であったが, 音声も用いることで事故時に発生する衝突音などの環境音, また, 人の声なども学習データとして含まれているからだと考える。

一方で, 提案手法により「ヒヤリハット (Near-Miss)」と「正常 (No-Near-Miss)」を誤認識 (図中の緑囲み) する割合が増加している部分が存在する。例えば, 「ヒヤリハット (Near-Miss)」を「正常 (No-Near-Miss)」と誤認識したデータ群を参照すると, 特徴的な音声やセンサのブレが無く, 直線道路を走行中に車の真横を自転車と並走するといったデータが多く見受けられた。また, 「正常 (No-Near-Miss)」を「ヒヤリハット (Near-Miss)」と誤認識したデータ群については, 道路上の起伏を原因として発生する鈍い音やセンサのブレ, また, 雪道シーンにおける映像など他の映像データと比較すると特異なシーンが多く含まれていることがわかった。

また, 評価指標として Precision, Recall, F1-Score の結果を表 2, 表 3 に示すが, 全指標において改善されていることから, 音声を考慮した本手法が危険運転シーンの判別において有効であることがわかった。

## 5 おわりに

ドライブレコーダーデータを用いて, 映像やセンサデータ,

表 3 提案手法 (映像+センサ+音声)

	Precision	Recall	F1-Score
Accident	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
Near-Miss	<b>0.83</b>	<b>0.84</b>	<b>0.83</b>
No-Near-Miss	<b>0.85</b>	<b>0.84</b>	<b>0.84</b>
Avg/Total	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>

音声情報からなる時系列マルチモーダルデータを抽出し, それらの特徴量を組合せたマルチモーダル深層学習を利用した危険運転 (ヒヤリハット, 事故) の検知を行った。

現状, 各モーダル情報から特徴量を抽出する 3 つのネットワーク部分について, 単純な CNN 層や LSTM 層, Fully Connected 層などの積層から構成されている。しかし, 特に映像部分の特徴を抽出するネットワークについては, 例えば 1 つの動画から通常の RGB 画像と Optical Flow を抽出した画像に分割し, 両者を入力とする Simonyan らの Two-Stream [6] による動き特徴の学習, また, 動画を直接入力する手法として Tranらによる 3D Convolution [7] を用いることで, 通常, CNN の 2 次元のフィルターを 3 次元形状に拡張し, 縦横の空間以外の時間方向への広がりを持つ特徴を抽出可能となり, さらなる精度向上を期待する。また, 音声用ネットワークについては, Sound Net [8] などのネットワーク構成の考慮, また, 映像中のどの部分に危険運転の因子となりえる要因が存在するかなどの Attention 機構の追加などを今後の課題とする。

## 文 献

- [1] F. Yu H. Xu, Y. Gao and T. Darrell. End-to-end Learning of Driving Models from Large-scale Video Datasets. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2017.
- [2] J. Canny J. Kim. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [3] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating Accidents in Dashcam Videos. In *Proc. of Asian Conference on Computer Vision*, 2016.
- [4] 菊池理人, 日景由華, 御室哲志. ドライブレコーダデータの自動分別の試み. 計測自動制御学会東北支部 290 回研究集会, 2014.
- [5] 山本修平, 結城遠藤, 戸田浩之. 映像とセンサ信号を用いたドライブレコーダデータからのヒヤリハット検出手法. 情報処理学会論文誌データベース (TOD), 第 10 巻, pp. 26–30, 2017.
- [6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, 2014.
- [7] R. Fergus L. Torresani D. Tran, L. Bourdev and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.
- [8] C. Vondrick Y. Aytar and A.A. Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. In *In Advances in Neural Information Processing Systems*, 2016.