

ドライブレコーダーデータを対象とした マルチモーダル深層学習

丹野 良介^{1,a)} 小澤 暖^{1,b)} 伊藤 浩二^{1,c)}

概要：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。

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

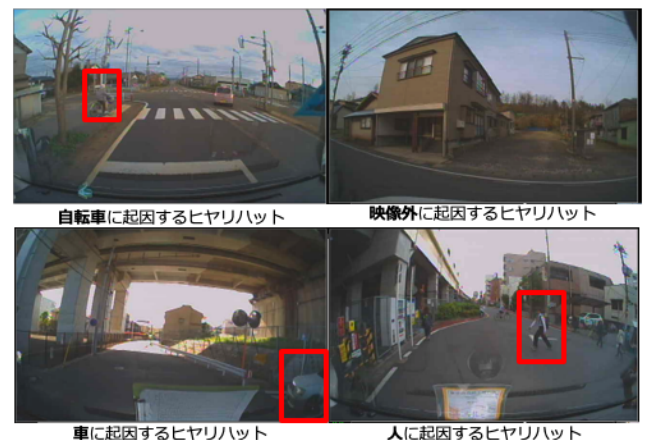


図 1 様々なヒヤリハット映像の例

Fig. 1 Examples of various near-miss video

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てる。運転手が行う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 [2]。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、次に行う行動（右折、停止など）を予測する。また本研

¹ NTT コミュニケーションズ株式会社
NTT Communications Corporation

a) r.tanno@ntt.com

b) d.ozawa@ntt.com

c) kj.ito@ntt.com

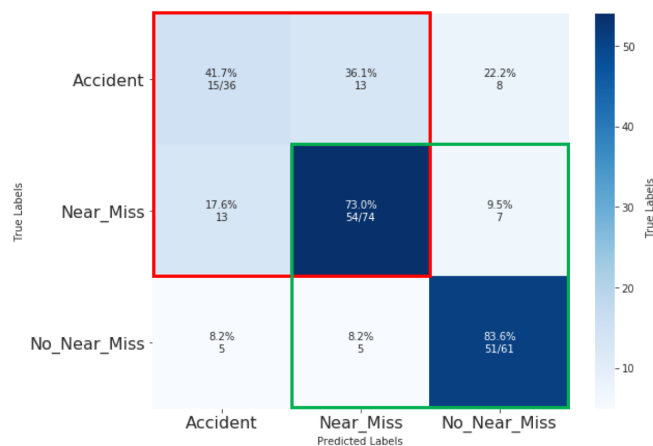


図 2 比較手法 [1] の混合行列 (映像+センサ)
Fig. 2 Confusion matrix of comparison method (Video + Sensor)

表 1 学習/評価用データ
Table 1 Train/Test data

	事故	ヒヤリハット	正常
データ数	190 件	364 件	298 件

究と同様に危険運転や交通事故を扱う研究として [3] や [4] 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 [5] 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ダリティ情報の組合せを利用した研究もある。山本ら [1] 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

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 つのモダリティ情報から構成される。映像を入力とする部分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 10 フレーム分切り出した画像を CNN の入力とする。

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

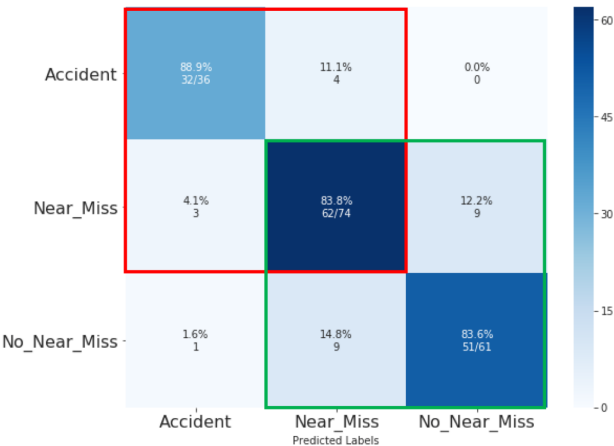


図 3 提案手法の混合行列 (映像+センサ+音声)
Fig. 3 Confusion matrix of proposed method (video + Sensor + Audio)

表 2 比較手法 [1](映像+センサ)
Table 2 Metrics of comparison method (Video + Sensor)

	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

表 3 提案手法 (映像+センサ+音声)
Table 3 Metrics of proposed method (video + Sensor + Audio)

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

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入力とする。

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 特徴量をネットワークへの入力とする。

各モダリティ情報は CNN 層や Fully Connected 層などから構成されるネットワークへと入力し、その後、各モダリティ情報間のネットワークのスケールの差異を吸収する目的として Batch Normalization 層へと入力される。

その後、RNN や LSTM などから構成される RNN Unit を介して時系列モデリングを行い、それぞれの特徴を Late Fusion の形で concat し、最終的に危険運転ラベルの予測を行う。概略図については発表当日に言及する。

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。

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2、図3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2) と比較して提案手法 (図3) 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や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、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部分につ

いて、単純な CNN 層や LSTM 層、Fully Connected 層の積層から構成されているが、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 Two-Stream [6] による動き特徴の学習や Sound Net [7] 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。

参考文献

- [1] 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ー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), Vol. 10, No. 4, pp. 26–30 (2017).
- [2] H. Xu, Y. Gao, F. Y. and Darrell, T.: End-to-end Learning of Driving Models from Large-scale Video Datasets, *Proc. of IEEE Computer Vision and Pattern Recognition* (2017).
- [3] J. Kim, J. C.: Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention, *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2017).
- [4] Chan, F.-H., Chen, Y.-T., Xiang, Y. and Sun, M.: Anticipating Accidents in Dashcam Videos, *Proc. of Asian Conference on Computer Vision* (2016).
- [5] 菊池理人, 日景由華, 御室哲志: ドライブレコーダー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).
- [6] Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, *Advances in Neural Information Processing Systems* (2014).
- [7] Y. Aytar, C. V. and Torralba, A.: SoundNet: Learning Sound Representations from Unlabeled Video, *In Advances in Neural Information Processing Systems*, (2016).