The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

運転データによる大規模ドライバー識別

Large Scale Driver Identification Using Driving Data

田中 大貴 *1*2 Daiki Tanaka　馬場 雪乃 *3*2 Yukino Baba　鹿島 久嗣 *1*2 Hisashi Kashima　齋藤朋也 *4*2 Tomoya Saito　大久保雄太 *5*2 Yuta Okubo

*1京都大学 Kyoto University　*2理化学研究所 革新知能統合研究センター RIKEN Center for AIP　*3筑波大学 University of Tsukuba

*4SOMPO リスケアマネジメント株式会社 SOMPO RISK MANAGEMENT & HEALTH CARE　*5損害保険ジャパン日本興亜株式会社 Sompo Japan Nipponkoa Insurance Inc.

We address a large-scale driver identification problem whose goal is to predict the driver who drives a car from driving data collected using GPS devices. In contrast with the prior research considering at most a few dozen of drivers, we try to identify a huge number of drivers up to 10,000 drivers. The experiment show our method can identify drivers more precisely than baseline method. We also show that temporal features are quite effective in the large scale driver identification and that speed and acceleration features also contribute to driver identification.

1. はじめに

ドライバー識別問題とは、自動車を運転しているドライバーに関する様々な情報をもとに、ドライバー個人を識別する問題である。その応用としては、複数人が同じ車を運転する可能性がある場合に、各ドライバーに合ったミラー角度の設定、運転席の高さの設定、エアコンの温度の設定などの適切な車両設定の自動化や、ドライバーの好みに合わせた情報推薦、ドライバーの運転経歴の自動作成、車両盗難の検知など多岐にわたる。

近年、機械学習をはじめとするデータ解析技術は様々な分野においてその利用が検討されており、ドライバー識別もその例外ではない。機械学習を用いてドライバー識別に取り組む研究は過去にも試みられてきたものの [Fung 17, Hallac 16, Van Ly 13, Miyajima 07]、データの大規模収集の困難さから、識別対象とするドライバー数は高々十数人の規模であった。一方、GPS を搭載したスマートフォン等の普及に伴い、これらを用いたアプリケーションを車上で用いることによって、運転行動に関連するデータが比較的簡単に大規模に収集できるようになってきた現在、大規模なドライバー識別の実現可能性が高まってきた。

本研究ではスマートフォン上で動作する運転診断アプリケーションを用いて収集された運転データをもとに、数千人規模のドライバー識別の実現可能性について検証する。最大 1 万人を対象としたドライバー識別実験では、ベースライン手法と比較して高い精度でドライバーを識別できることが示された。また、位置や時間に関する特徴量、特に位置による特徴量が大規模なドライバー識別において極めて有効であることを確認するとともに、速度や加速度情報もドライバー識別に一定の寄与があることを示した。

本研究の貢献は以下のように纏められる：

- 実データに対して数千人規模という大規模なドライバー識別を行なった。
- ドライバー識別に有効な特徴量について検討した。

連絡先: 田中 大貴, daiki_yosky@ml.ist.i.kyoto-u.ac.jp

2. 関連研究

これまでの研究においても、ドライバー識別問題は比較的少人数を対象とした設定で取り組まれてきた。Fung らは、14 人の熟練ドライバーの加速時、減速時の速度、加速度、躍度、車体角度から特徴抽出を行った [Fung 17]。運転データに含まれる加速、減速それぞれの識別結果を多数決を用いて統合し、識別器には多クラス線形判別分析を用いて、2 人でのドライバー識別において約 80%、14 人でのドライバー識別において約 50%の精度を得た。

Hallac らは、車両がコーナーを曲がる数秒の短い区間の情報のみからドライバー識別に取り組んだ [Hallac 16]。ハンドル角度、ステアリング速度・加速度、車体速度・加速度、車体の向き、エンジン回転数、アクセルペダル・ブレーキペダルの位置、トルク、スロットル位置から特徴量を抽出し、分類を行なった。ウェーブレット変換による特徴生成も行った。識別器にランダムフォレストを使い、2 人でのドライバー識別において 76.9%、5 人でのドライバー識別において 50.1%の精度を得た。

Van Ly らは、加速、減速、コーナリングという 3 つのイベントを用いて 2 人ドライバー識別に取り組んだ [Van Ly 13]。彼らは運転データからこれら 3 種のイベントを抽出し、イベント毎に特徴抽出を行い、ドライバー識別を行なった。彼らは CAN バスから得られる加速度とジャイロ情報から特徴抽出を行なった。識別器には SVM と k-means クラスタリングを用いて、約 60%の精度を得た。

Miyajima らは、ガウス混合モデルを用いてアクセルペダルとブレーキペダルの踏み具合をモデル化することで、2 人ドライバー識別に取り組んだ [Miyajima 07]。時系列データに対してケプストラム解析を行い、その有用性について示した。精度は 76.8%を得た。

表 1 に纏めるように、過去にもドライバー識別に取り組んだ研究はいくつかあるが、いずれも高々数十人の識別を対象としている。本研究では、これらの研究が取り組んでいない数千人規模でのドライバーの識別に取り組んだ。

表 1: ドライバー識別の関連研究の概要

著者	年	分類器	ドライバー数
Fung et al. [Fung 17]	2017	多クラス線形判別	14
Hallac et al. [Hallac 16]	2016	ランダムフォレスト	5
Van et al. [Van Ly 13]	2013	SVM, k-means	2
Miyajima et al. [Miyajima 07]	2007	ガウス混合モデル	2

3. データセット

本研究で使用するデータセットは、損害保険ジャパン日本興亜株式会社によって収集された約 36,000 人分の運転データである。なお、位置情報を含む運転データについては、個人情報の削除ならびに変換を行い、データから個人特定が困難であるように加工したのちに用いた。

3.1 データ取得

本研究で用いたデータセットは、自動車内にドライバーが持ち込んだスマートフォン上で動作する運転診断アプリケーションから収集されたものである。ドライバーはアプリケーションを起動させ、運転データの送信を開始し、運転を開始する。運転が終了したら運転データの送信を終了する。運転データのサンプリング頻度は 1Hz である。

3.2 運転データ

運転データは、ドライバーが行なう運転に関するデータを提供する。運転データはスマートフォン ID に紐づけられている。運転データには 1Hz 毎に記録された時間 (タイムスタンプ)、緯度、経度、速度、前後加速度、左右加速度が含まれている。運転データに含まれる総走行時間は約 70 万時間で、欠損データを除いた総走行時間は約 58 万時間である。総運転データ数は 1,575,841 であり、36,725 人の運転データが含まれている。また、運転データの時系列長はデータにより様々である。本研究で行なう実験にはこの運転データのサブセットを用いる。

4. 問題設定

本研究で扱うドライバー識別問題とは、運転中のタイムスタンプ、経度、緯度、前後加速度、左右加速度、速度のデータから、その運転を行なったドライバーを識別することである。

入力として、各ドライバーの各運転に対して、運転データが与えられる。なお、各ドライバーに対して、1 つ以上の運転データが存在する。運転データとは、1 回のアプリケーションの開始から終了までの走行時の情報を記録したものであり、具体的にはタイムスタンプ、位置情報、加速度データ、速度データが時系列で記録されている。すなわち、i 番めの走行データは、タイムスタンプ $\mathbf{time}_i$ と緯度 $\mathbf{lat}_i$ と経度 $\mathbf{lon}_i$ と前後加速度データ $\mathbf{acc}_i^{\mathrm{FB}}$ と左右加速度データ $\mathbf{acc}_i^{\mathrm{LR}}$ と速度データ $\mathbf{v}_i$ で与えられる。なお、$\mathbf{time}_i$, $\mathbf{lat}_i$, $\mathbf{lon}_i$, $\mathbf{acc}_i^{\mathrm{FB}}$, $\mathbf{acc}_i^{\mathrm{LR}}$, $\mathbf{v}_i$ は時系列データである。時系列データ $\mathbf{time}_i$, $\mathbf{lat}_i$, $\mathbf{lon}_i$, $\mathbf{acc}_i^{\mathrm{FB}}$, $\mathbf{acc}_i^{\mathrm{LR}}$, $\mathbf{v}_i$ から i 番目の走行に対する特徴ベクトル $\boldsymbol{x}_i = (x_i^1, x_i^2, ..., x_i^n)$ を構成し、その特徴ベクトル集合ならびにドライバーの識別子から分類器を訓練する。

なお、本研究では 3 種類の特徴ベクトルを設計し、それらからそれぞれドライバー識別を行う。特徴量と特徴ベクトルの構成については次の章で述べる。

表 2: 実験で用いる特徴量の種類

データ	特徴量の種類
速度、加速度	前後加速度の最大値、最小値、平均値、分散
	左右加速度の最大値、最小値
	速度の最大値、平均値、分散
位置、時間	アプリ開始・終了時の緯度、緯度の平均・分散
	アプリ開始・終了時の経度、経度の平均・分散
	運転開始時の曜日、時間

5. 特徴量の設計

本研究でドライバー識別に用いる特徴量は、速度と加速度に関する特徴量、位置に関する特徴量、時間に関する特徴量の 3 種類である。それぞれの特徴量について、以下で説明するとともに表 2 にまとめる。

5.1 速度と加速度に関する特徴量

運転データに含まれる速度データと加速度データから、9 つの特徴量を設計する。速度については、最大値、平均値、分散を、加速度については、前後加速度の最大値、最小値、平均値、分散ならびに左右加速度の最大値、最小値を用いる。

なお、信号待ちなどによる停車中のデータや徒歩によるデータを除外するため、速度に関する特徴量を抽出する際には時速 7km/h 以下のデータを排除し、加速度に関する特徴量を設計する際には −0.001G 以上 0.001G 以下のデータを排除する。

5.2 位置に関する特徴量

運転データに含まれる緯度データと経度データから、アプリケーション開始時と終了時の緯度経度ならびに緯度経度の平均値と分散の 8 つの特徴量を用いる。

5.3 時間に関する特徴量

運転データに含まれるタイムスタンプから、アプリケーションが起動された時間と その日の曜日の 2 つの特徴量を設計する。なお、曜日は月曜日から日曜日まで曜日ごとに、0 から 6 までの異なる数を割り当て、それらを one-hot-vector として特徴量とする。得られた one-hot-vector を他の特徴量から構成された特徴ベクトルに連結する。

6. 評価実験

前述のデータを用いて、多人数でのドライバー識別の実現可能性を検討する。まずは大規模ドライバー識別で用いる識別器を決定するための小規模での予備実験を行った後、全データを用いた本実験を行った。

6.1 評価指標

数千人を超える多数のドライバー識別を確実に行うのは困難である。識別器はより本人であると思われる順位を出力するため、上位に予測されるドライバーに本人が含まれていれば良

いとする予測精度指標として、情報検索の精度評価にしばしば用いられる Recall@k と平均逆順位を用いる。

Recall@k 上位 k 件の候補中に含まれる正解ラベル数を全正解ラベル数で割ったものの平均値である。本研究では正解ラベルが運転をしたドライバーのラベル 1 件であるので、Recall@k は以下の式で定義される：

$$\text{Recall@k} = \frac{1}{|M|} \sum_{m \in M} t_{m,k} \tag{1}$$

ただし、$t_{m,k}$ はサンプル $m \in M$ の予測結果の上位 k 位に正解ラベルが含まれていれば 1、含まれていなければ 0 をとる。実験では $k = 1, 3, 5, 10$ の場合について Recall@k を求めた。

平均逆順位 正解ドライバーの順位の逆数の平均値であり、以下の式で定義される：

$$\text{MRR} = \frac{1}{|M|} \sum_{m \in M} \frac{1}{r_m} \tag{2}$$

ただし、r_m はサンプル $m \in M$ の予測結果中の正解ラベルの順位である。

6.2 予備実験

大規模ドライバー識別問題で用いる識別器を決定するため、まずは予備実験として小規模なドライバー識別問題によって各識別器の性能を確認する。

識別対象とするドライバーは、ドライバー集合からランダムに選択された 30 人とする。用いる運転データ数は合計 6,248 件で、80%を訓練データ、20%をテストデータとして使用する。なお、識別に用いる特徴ベクトルは全ての特徴量で構成する。識別器として、SVM、ニューラルネットワーク、ランダムフォレストの 3 種を用いる。各識別器に対して 5-交差検定によるグリッドサーチによってハイパーパラメータチューニングを行う。SVM のハイパーパラメータの探索空間は、誤識別数に対するペナルティC の値域は $\{0.1, 1.0, 10, 30, 50\}$、カーネル関数は {RBF カーネル, 線形カーネル, 多項式カーネル, シグモイドカーネル } の 4 種の中から選択し、多クラス識別方式は {One-vs-One 方式, One-vs-Rest 方式 } の 2 種から選択する。ニューラルネットワークについては隠れ層の層数 {1、2、3} の中から、各層のユニット数 {50、100} の中から瀬選択する。ランダムフォレストについては決定木の最大の深さを $\{10, 15, 20, 25, 30\}$ から、決定木の本数を $\{100, 500, 1,000, 1,500\}$ から選択する。

チューニングの結果、SVM は誤識別数に対するペナルティ C は 10、カーネル関数は線形カーネルが、One-vs-One 方式のものが選択され、ニューラルネットワークは中間層が 3 層のもので、各中間層のユニット数はそれぞれ 1000 のものが選ばれた。ランダムフォレストの決定木の最大深さは 20、決定木の本数は 1,000 のものが選択された。

識別器ごとの予測精度を表 3 に示す。ランダムフォレストによる識別が、全ての指標において最も良い精度となったため、大規模ドライバー識別問題ではランダムフォレストを用いる。

表 3: 30 人ドライバー識別での各識別器の性能

識別器	SVM	NN	ランダムフォレスト
Recall@1	0.726	0.788	**0.885**
Recall@3	0.939	0.955	**0.981**
Recall@5	0.984	0.985	**0.997**
Recall@10	0.999	0.999	**0.999**
MRR	0.842	0.875	**0.934**

6.3 本実験の設定

本実験では、(1) 時間に関する特徴量と位置に関する特徴量のみを使って予測を行なう場合、(2) 速度と加速度に関する特徴量のみを使って予測を行なう場合、(3) 全ての特徴量を使って予測する場合の 3 つの予測モデルに (4) ベースライン手法を加えた 4 種類のモデルを比較する。識別器としては予備実験で決定されたランダムフォレストを使用する。ランダムフォレストのハイパーパラメータとして、森の中の決定木の本数、決定木の最大深さは改めて調整を行う。ベースライン手法としては、単純にデータ数の多いドライバー順にランキングして、これを予測結果とする方式を用いる。すなわち、識別対象とするドライバー数を M として、使用したデータ数で降順ソートしたドライバーを $y_1, \ldots, y_M$ とする時、ベースライン手法による予測結果は、上位から $y_1, \ldots, y_M$ となる。

ドライバーによって運転データの数は異なるため、100 個以上のデータが存在するドライバーはおよそ 100 個のデータを、100 個未満しかデータがないドライバーについてはそのユーザのデータ全てを使用する。このようにして得られた全データ中の 80%を分類器の学習に使い、20%を性能評価用のテストデータとして使う。なお、実験に用いたドライバーはデータ数に関して降順ソートした順に選択する。例えば 10 人識別であれば、ドライバー集合からデータ数の多い順に 10 人のドライバーを選択する。したがって、識別人数を増やすに従い 1 人当たりのデータ数が減っていくことに注意されたい。

識別対象とするドライバーの数は 10、100、1,000、5,000、10,000 人とした。評価指標には 6.1 節で述べた Recall@k ($k = 1, 3, 5, 10$)、平均逆順位を用いて評価を行なう。用いるデータ数は、訓練データとテストデータを合わせて、10 人識別では 1,189 件、100 人識別では 11,358 件、1,000 人識別では 107,135 件、5,000 人識別では 430,844 件、10,000 人識別では 609,279 件である。

6.4 本実験の結果

各特徴ベクトルの設計方法について、Recall@k ($k = 1, 3, 5, 10$)、平均逆順位を識別人数ごとにプロットしたグラフを示す。なお、横軸の識別人数は対数プロットしている。図 1 が Recall@1、図 2 が Recall@3、図 3 が Recall@5、図 4 が Recall@10、図 5 が MRR を示したものである。All が全ての特徴量で識別する場合、Loc and Time が位置と時間に関する特徴量で識別する場合、Speed and Acc が速度と加速度に関する特徴量で識別する場合、Baseline がベースライン手法でドライバーを識別をする場合の結果を示している。

識別人数が 10、100、1,000、5,000、10,000 人の全ての場合において、設計した特徴ベクトルを用いてドライバー識別を行なった方がベースラインよりも良い結果となり、本研究で構成した特徴量がドライバー識別に有効であることが分かった。10,000 人という大人数でのドライバー識別においても、全ての特徴量を用いて識別を行った場合平均逆順位が 0.33 という比較的高い精度を得ることができた。全ての特徴量を用いて識別を行なった場合と、位置と時間による特徴量のみを用いて識別を行なった場合には大きな精度の差はなかったが、ほぼすべての評価指標において、速度と加速度による特徴量を用いることでのわずかな精度向上が見られた。また、速度や加速度によ

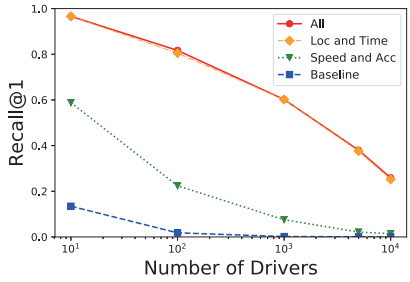

図 1: Recall@1 による各手法の比較

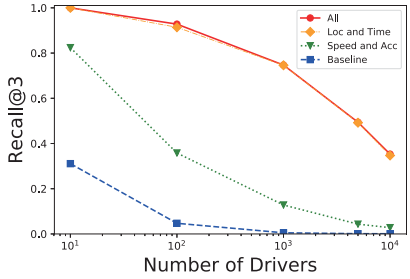

図 2: Recall@3 による各手法の比較

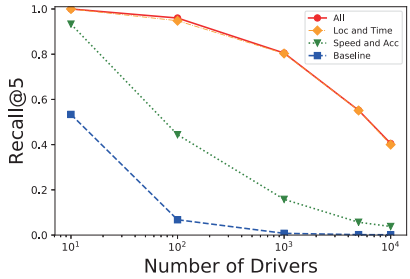

図 3: Recall@5 による各手法の比較

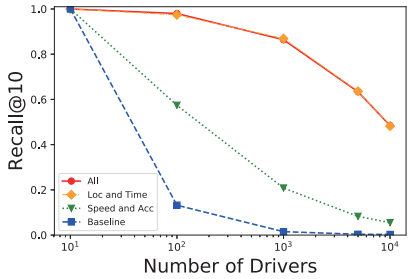

図 4: Recall@10 による各手法の比較

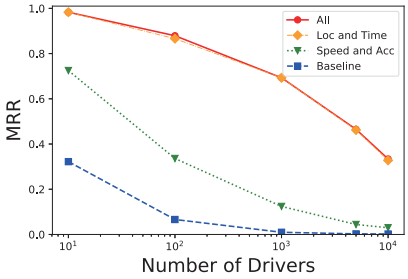

図 5: 平均逆順位による各手法の比較

る特徴量のみから識別を行なった場合に、ベースライン手法よりも良い精度を得たことからも、速度や加速度の情報は大規模ドライバー識別問題において一定の役割をもつと考えられる。とりわけ、様々な理由により位置情報の使用が制限される場合には、相対的にその重要度は高まると考えられる。また、識別対象の人数が増えるに従い、識別問題の難易度が上がるため、精度が落ちていくことも確認できる。

7. 結論

本研究では実運転データを用いた大規模ドライバー識別に取り組んだ。1 万人のドライバー識別実験では、ベースライン手法と比較して高い精度でドライバーを識別できることが示された。また、位置や時間に関する特徴量、特に位置による特徴量が大規模なドライバー識別において極めて有効であることを確認するとともに、速度や加速度もドライバー識別に一定の寄与があることを示した。

なお、本研究で使用したデータセットは、様々な車種の自動車に持ち込まれた携帯電話から収集されたデータをもとにしたものであり、速度と加速度といった特徴量は、加速性能やブレーキ性能、車両重量、トランスミッション方式などの、車種によるスペックの違いも大きく受けていると推測される。そのため、特に速度と加速度による特徴量のみによるドライバー識別問題は、車両識別問題とドライバー識別問題が混在した問題になっていると考えられる。今後はこれらを区別したより詳細な実験設定を行っていくことを予定している。

参考文献

[Fung 17] Fung, N. C., Wallace, B., Chan, A. D., Goubran, R., Porter, M. M., Marshall, S., and Knoefel, F.: Driver identification using vehicle acceleration and deceleration events from naturalistic driving of older drivers, in *Proceedings of the 12th Annual IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 33–38 (2017)

[Hallac 16] Hallac, D., Sharang, A., Stahlmann, R., Lamprecht, A., Huber, M., Roehder, M., Sosic, R., and Leskovec, J.: Driver identification using automobile sensor data from a single turn, in *Proceedings of the 19th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 953–958 (2016)

[Miyajima 07] Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., and Itakura, F.: Driver Modeling Based on Driving Behavior and Its Evaluation in Driver Identification, *Proceedings of the IEEE*, Vol. 95, No. 2, pp. 427–437 (2007)

[Van Ly 13] Van Ly, M., Martin, S., and Trivedi, M. M.: Driver classification and driving style recognition using inertial sensors, in *Proceedings of the 24th IEEE Intelligent Vehicles Symposium (IV)*, pp. 1040–1045 (2013)