2016 IEEE International Congress on Big Data

Do Drivers' Behaviors Reflect Their Past Driving Histories? – Large Scale Examination of Vehicle Recorder Data –

Daisaku Yokoyama
Institute of Industrial Science,
The University of Tokyo,
Tokyo, Japan
yokoyama@tkl.iis.u-tokyo.ac.jp

Masashi Toyoda
Institute of Industrial Science,
The University of Tokyo,
Tokyo, Japan
toyoda@tkl.iis.u-tokyo.ac.jp

***Abstract*—We present a method for analyzing the relationships between driver characteristics and driving behaviors on the basis of large-scale and long-term vehicle recorder data. Previous studies relied on precise data obtained under critical driving situations, which led to overlooking routine driving behaviors. In contrast, we used a dataset that was sparse but large-scale (over 100 fleet drivers) and long-term (one year's worth) and covering all driving operations. We focused on classifying drivers by their accident history and examined the correlation between having an accident and driving behavior. We were able to reliably predict whether a driver had recently experienced an accident (f-measure $>$ 86%). This level of performance cannot be achieved using only the drivers' demographic information. We also found that taking into account the driving circumstances improved classification performance and that driving operations at low velocity were more informative. This method can be used, for example, by fleet driver management to classify drivers by their skill level, safety, physical/mental fatigue, aggressiveness, and so on.**

***Keywords*-Vehicle recorder; Driving safety; Accident history;**

I. Introduction

Smarter transportation systems that are safe, efficient, and comfortable are essential to achieving smart cities. A key to achieving such systems is driver management, i.e., preventing accidents by predicting unsafe behaviors and educating drivers on how to improve their driving.

We have developed a method for analyzing the relationships between driver characteristics and driving behaviors on the basis of large-scale and long-term vehicle recorder data. It can be used, for example, by fleet driver management to classify drivers by their skill level, safety, physical/mental fatigue, aggressiveness, and so on. Transport companies typically manage their drivers on the basis of their demographic attributes, not on the basis of their driving behaviors.

Several studies [1][2][3] have analyzed driving behaviors. They relied, however, on detailed and precise data for a small number of drivers, so it is difficult to extrapolate the results to the general driver population. Many transportation companies have introduced dashboard cameras (dashcams) and/or vehicle data recorders (which collect GPS, velocity, and acceleration data) into their fleets. Although the amount of data collected tends to be sparse due to storage limitations, data can be collected for a large number of drivers.

Our method classifies drivers on the basis of long-term records of the kinematic variables (maximum velocity, acceleration, etc.) related to their driving operations (braking, steering, etc.). It is based on the assumption that the distributions of these variables differs from driver to driver. We focused on classifying drivers who had recently been involved in an accident and examined the correlation between having an accident and driving behavior. Our findings are useful both for educating drivers and preventing accidents.

Many studies [4][5] have analyzed driving behaviors as a means to estimate drivers' risks. They focused only on the occurrence of critical driving operations involving high acceleration. However, a driver's characteristics such as driving skills are reflected in all situations, not only in critical situations; for example, a skillful driver will brake smoothly even when driving in slow traffic. The previous studies thus overlooked the information to be obtained from operations performed in non-critical situations. We used *all* driving information derived from the data stored in a vehicle recorder to better estimate a driver's characteristics.

This work makes three main contributions:

- Intensive examination of large-scale vehicle recorder data covering all driving operations demonstrated the effectiveness of our method for analyzing the relationships between driver characteristics and driving behaviors. It was able to reliably predict whether a driver had recently experienced an accident (f-measure $>$ 86%). This level of performance cannot be achieved using only the drivers' demographic information, which is widely used to estimate drivers' safety.
- It showed that the kinematic variables during driving operations are affected by the driving circumstances, and combining these information can help to analyze drivers' characteristics. In particular, taking into account the vehicle's velocity and road width when an operation occurred was shown to improve driver classification.
- Operations performed at lower velocities were found

978-1-5090-2622-7/16 $31.00 © 2016 IEEE
DOI 10.1109/BigDataCongress.2016.58

to be more informative than those performed at higher velocity. This means that using data for all driving operations is useful in understanding driver characteristics.

In Section II, we overview related work. In Section III-A, we explain our analysis. We explain the dataset we used in Section III-B and examine the data in Section III-C. In Section III-D, we present our proposed method for analyzing the relationships between driver characteristics and driving behaviors and evaluate its effectiveness. This article ends in Section IV with a summary and a look at future work.

II. Related Work

There has been research on using vehicle recorded data, such as velocity and location, for various purposes [6][7]. They are grouped into two categories; researches that utilize large-scale vehicle location data [8] and that investigate small amount of driving operation data. We believe that this is the first research to investigate both driving operation and vehicle location data in large-scale (more than 1000 drivers).

The 100-Car Naturalistic Driving Study [9] is one of the largest studies on the use of vehicle recorded data. It used many types of precise driving information and driver demographic data (age, gender, personality, etc.), and driver information was intensively statistically analyzed. Several studies have used driving information in this archive to assess driver risk. For example, Guo et al. [4] reported an effective model for identifying high-risk drivers using driver demographic information and the occurrence of critical-incident events. Their model mainly uses demographic information and does not consider non-critical driving behavior. Zheng et al. [5] collected data on naturalistic driving and analyzed the relationship between the kinematic information and driver risk-taking behavior. Their analysis focused on kinematic information for critical driving operations involving high acceleration.

There are some research efforts tried to classify drivers on the basis of the aggressiveness of their driving behaviors with the aim of improving driving safety. Higgs et al. [1] analyzed the car-following behaviors of three drivers and identified the differences among them. Dang et al. [2] focused on the lane-changing behaviors of 12 drivers driving on a highway and found that some variables, such as the lane-changing frequency, differed among them. Miyajima et al. [10] used data for 276 drivers and tried to identify drivers on the basis of their car-following behaviors and pedal operations. However, their data collection required the use of pedals with specially designed sensors. Their study and the other previous research relied on precise information on driving behavior, which is not always available.

III. Classification of drivers' accident history

A. Approach

Our research purpose is to identify the characteristics of drivers through their driving behaviors. In this study, we focused on classifying drivers as either safe or unsafe on the basis of their driving records. Instead of using only critical operation records, we used a large amount of vehicle recorder data that included all driving operations and investigated how effective such data is for classifying drivers.

A driver performs various driving operations (braking, steering, etc.), each with several variables (maximum velocity, acceleration, etc.). A driver can be characterized by the distributions of these variables. We investigated ways to derive features from these variable distributions for use in classifying drivers as either safe or unsafe using Support Vector Machine (SVM).

Each driving operation is affected by the factors of the moment, such as velocity, road condition, degree of congestion, and time of day. We need to take into account the effects of these factors in order to derive good features from the operation records. Here we focus on two circumstances: velocity and road width. We derived several features from the distributions of operation variables, taking into account the factors of the moment, and evaluated the effectiveness of our method.

Table I
Summary of vehicle recorder dataset

All data	
Number of drivers	1469
Driving duration in total	77,450 hours

Driving days $\geq$ 20, driving hours $\geq$ 20	
Number of drivers	320
Driving duration in total	60,190 hours

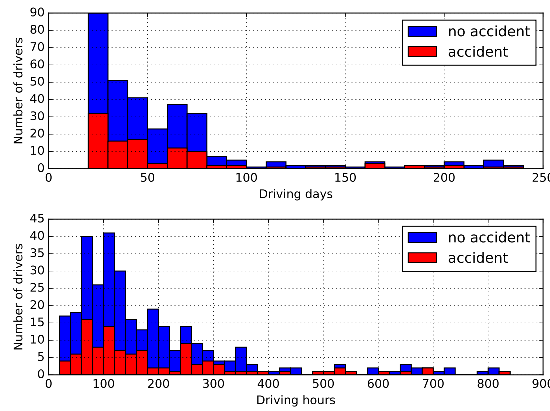

Figure 1. Driver statistics for whole dataset

B. Dataset

1) Vehicle recorder dataset: In our experiments, we used a large number of actual driving records[1] collected by a parcel delivery service company (transport company). The

[1]The vehicle recorder data was provided by Datatec Co., Ltd.

Table II
OPERATION RECORD VARIABLES

Operation	Variables
Braking	velocity (V), longitudinal acceleration (Gx), and jerk (derivative of acceleration with respect to time, Jx)
Steering	V, yaw velocity (Yr), yaw acceleration, and lateral acceleration (Gy)
Turning	{Gx, V} before turn, {V, centrifugal force (CG), yaw acceleration} during turn, and {V, CG} after turn
Stopping	V, Gx, and stopping duration

Table III
OPERATION RECORD STATISTICS

Operation	No. of records per driver (min)	(max)	No. of records (total)
Braking	114	45,861	1,993,341
Steering	239	46,452	2,783,723
Turning	121	21,027	1,218,957
Stopping	418	40,625	2,221,166

data were for about 1450 drivers working in the Tokyo area and covered 1 year (from 21 July 2014). A multifunctional data recorder in each delivery vehicle recorded longitudinal accelerometer, lateral accelerometer, gyro compass, and GPS data.

Since we focused on long-term driving behavior, we eliminated the data for drivers who had driven on fewer than 20 days or for less than 20 hours in total. A summary of the data is shown in Table I, and some driver statistics are plotted in Figure 1. The upper histogram shows the distribution of driving days, and the lower one shows the distribution of driving hours. The driving hour data does not include the time when the engine was turned off.

The vehicle data recorder automatically detected four basic driving operations: braking, steering, turning, and stopping. Several variables, including maximum velocity and acceleration, during each operation were recorded. The operation variables are listed in Table II. The numbers of recorded operations per driver are summarized in Table III. As mentioned, our dataset contained data on all driving operations while those used in previous studies contained data only on critical operations involving high acceleration.

2) Driver histories: With the cooperation of the transport company, we accessed the driver histories, including the traffic violations they had received and the accidents in which they had been involved. We used their histories to define their *accident experience* and *driving experience*.

Accident experience: Drivers who had had at least one accident during a certain time period were defined as an *accident* driver, and the others were defined as a *no accident* driver. Even though some accidents were only slight accidents without responsibility being assigned, we treated all accidents the same.

Driving experience: To estimate how long a driver had been driving, we used the oldest record in the driver's history to estimate the minimum number of driving years.

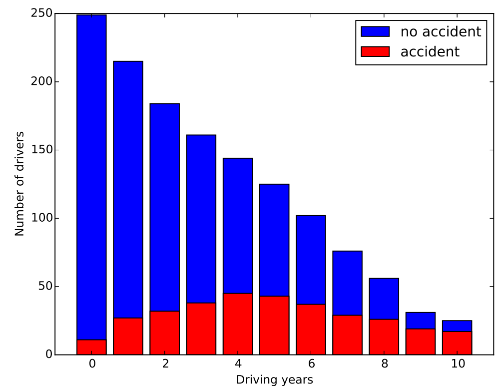

Figure 2. Number of drivers by number of driving years

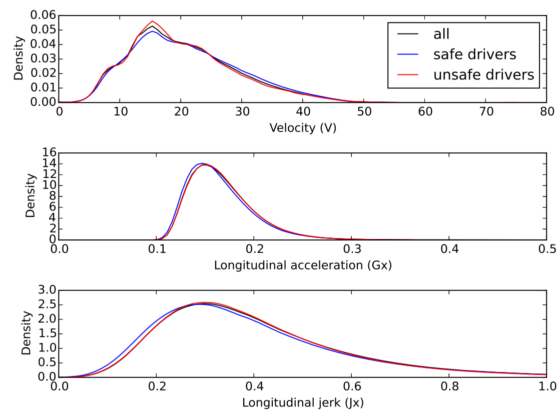

Figure 3. Distributions of braking operation variables

Using these definitions and estimate, we investigated the difference in driving operations between accident and no-accident drivers. Figure 2 shows the number of drivers by the number of driving years. The red bars represent the number of drivers who had had at least one accident during a certain years. The blue bars represent the number of the drivers who did not.

The no-accident drivers are not necessarily safe drivers. For example, a reckless driver may simply have been lucky enough to avoid an accident over the course of a year. We therefore focused on drivers who had had at least five years' worth of driving experience. We defined a driver who had had at least five years' worth of driving experience without any accidents in the previous five years as safe and otherwise as unsafe. There were 82 safe drivers and 43 unsafe drivers.

C. Observation of drivers' behavior in operation records

1) Distributions of operation variables: Figure 3 shows the distributions of the three variables for the braking operation for all drivers, safe drivers, and unsafe drivers. The y-axis indicates the density estimated by kernel density

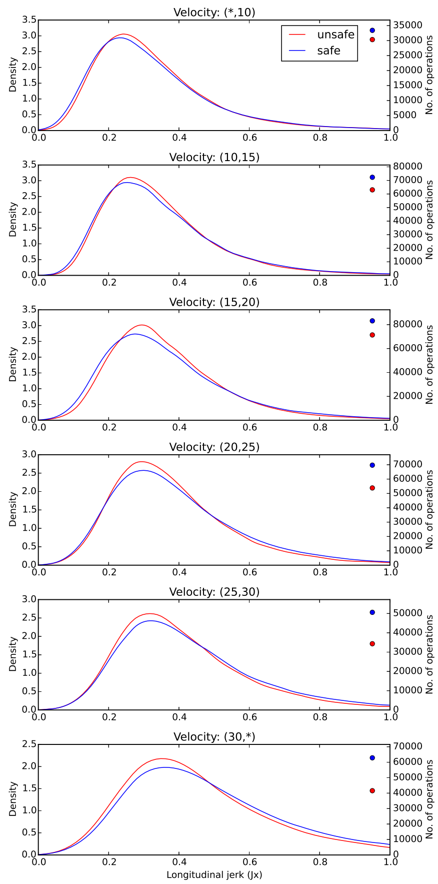

Figure 4. Density of longitudinal jerk during braking for various velocities

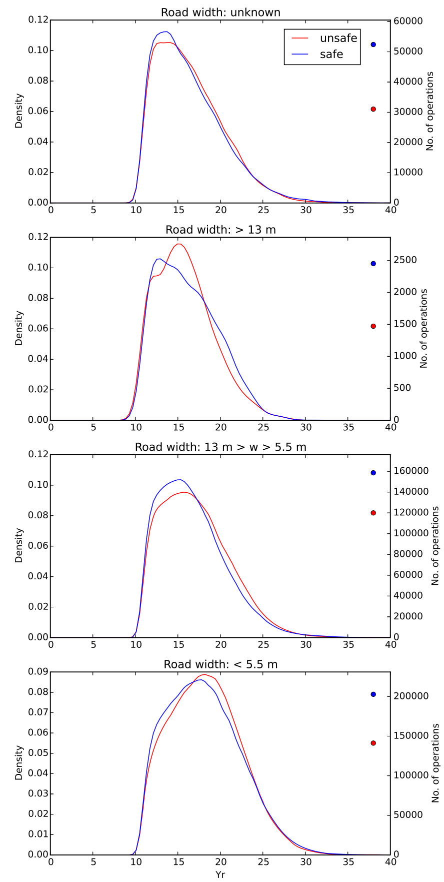

Figure 5. Density of yaw rate distribution during steering operation for various road widths

estimation. Although the distributions differ slightly, they are virtually the same.

2) Velocity: To overcome the problem of the distributions of variables differing only slightly, we consider the correlation between variables. For example, steering at a high velocity tended to cause a low yaw rate due to kinematic restrictions for both safe and unsafe drivers. We therefore treated velocity as the basic variable for each operation and split the operation records by velocity value. For example, we divided the braking operation records into six bins on the basis of velocity, and estimated the longitudinal jerk densities for each bin, as shown in Figure 4. The shapes of the distributions differ among the velocity bins, making the differences between the distributions for safe and unsafe drivers much clearer than shown in Figure 3. The two small circles at the rightmost of each panel indicate the number of operations by driver type, with the number shown on the right vertical axis. The unsafe drivers tended to operate the brake more frequently at lower velocity than the safe drivers.

3) Road width: Driving operations are also affected by the road width. For example, turning onto a narrower road tends to require more deceleration than turning onto a wider road. Since every operation record contained GPS data, we could match each operation location with a point on a digital road map[2]. We simply searched for the road segment nearest the operation location. If the nearest segment was more than 30 m away (due, for example, to being on a private site such as a factory or university), we considered that the location could not be matched to a point on the map and ignored that record. The road map contains information about the road width, represented in several ranks, and whether the road is bi-directional or not. If the road was bi-directional, we assumed that the width of the road segment was one rank narrower. We used four road width ranges: $>$ 13 m, $13 > w > 5.5$, $<$ 5.5 m, and unknown.

As shown in Figure 5, the distributions for yaw rate during the steering operation differed among road widths. The difference between the distributions for safe and unsafe drivers is clearer than in Figure 3, as in the case for velocity by bins.

[2]We used the "Advanced Digital Road Map Database" developed by Sumitomo Electric System Solutions Co., Ltd. The database was provided by the Center for Spatial Information Science at the University of Tokyo.

Table IV
COMBINATION PATTERNS OF OPERATION VARIABLES

Operation	Velocity-related variable (number of bins)	Other variables combined with velocity-related variable
Braking	velocity (6)	Gx, Jx
Steering	velocity (5)	Yr, yaw acceleration, Gy
Turning	velocity before turn (4)	Gx before turn
Turning	velocity during turn (4)	CG, yaw acceleration during turn
Turning	velocity after turn (5)	CG after turn
Stopping	velocity (5)	Gx

D. Features

1) Derivation: We used all 17 dataset variables listed in Table II to derive driver features. We also used driver demographic information known to be related to driving safety.

Demographic features: We used the driver's age, gender, and time since obtaining a driver's license as three demographic features. This informative is commonly used by insurance companies to set auto insurance rates.

License feature: In Japan, a driver who has not had any accidents and has not been cited for a driving violation during the preceding five years is categorized as a "gold license" driver and is generally considered to be a safe driver. We thus defined a binary feature for whether a driver had a gold license or not.

The license category is updated when one's license is renewed, and the renewal interval is 3 to 5 years. Therefore, a gold license does not always mean an accident-free driver; many drivers have had accidents in recent years and still hold a gold license. When we classify drivers as safe or unsafe by using their license category information alone, we achieved only 35% precision, which is virtually the same performance as with a random classifier.

Operation frequency features: We counted the number of instances for each of the four driving operations for each driver and normalized it by the driving duration.

Variable distribution features: We defined the shapes of the variable distributions as features. Each variable value was binned into one of ten intervals; the maximum and minimum bin breakpoints were chosen by hand, and the other bins were defined to have the same width. Therefore, each variable distribution was represented by ten values. There were thus 170 variable distribution features (17 variables $\times$ 10 values).

Road width distribution features: We defined each of the four road width ranges for each of the four driving operations as a feature.

Variable distribution by velocity features: Driving operations are strongly affected by the vehicle's velocity. We therefore selected six velocity-related variables for use in separating the operation records, and combined them with other variables, as shown in Table IV. The operation records were separated by the corresponding velocity-related variable, and the distributions of the other variables were calculated separately. The velocity-related variables were digitized into b values by intervals with a constant width (5 km/h). The other variable distributions were digitized with ten intervals, so the feature of a variable is represented by $b \times 10$ values.

Variable distribution by road width features: We separated the operation records by road width and calculated the variable distributions separately.

Variable distribution by velocity and road width features: We separated the operation records by their velocity-related variables and road width and calculated the variable distributions separately. This feature reflects the effects of both velocity and road width.

Increasing the number of combinations improved the accuracy of the depicted variable distribution for each driver. Although this helps to describe the difference between one's driving behaviors precisely, it may cause data sparsity because it reduces the number of operation occurrence in each bin, which means the features will be more strongly affected by noise.

2) Feature expression: We tested several methods for expressing the variable distributions as features.

Probability method: We accumulated each driver's frequency for each bin as N_i and computed each driver's occurrence probability P_i, which is N_i normalized by the number of operation instances for the driver. We used P_i itself as a feature.

Difference from average driver: We accumulate all the driver frequencies for each bin and computed the bin's probability as Q_i. This value represents the distribution for the average driver. The difference between a driver and the average driver may be related to the driver's characteristics. We tested two methods for representing this information.

- Use $P_i - Q_i$. Basically this representation does not differ from P_i itself because each Q_i is constant among all drivers. This representation may affect normalization during SVM classification and feature selection.
- Use $sign = \begin{cases} 1 & (P_i > Q_i) \\ 0 & (P_i == Q_i) \\ -1 & (P_i < Q_i) \end{cases}$.

Entropy-like method: We assumed that each occurrence of an operation has $-\log(Q)$ information and that the characteristic of each driver's distribution for that operation can be represented as $-P\log(Q)$. Such representation is often used in anomaly detection because it emphasize the occurrence of rare case. Basically, this representation does not differ from that of P_i due to the feature normalization; however, it may affect feature selection.

We represent whether P is greater than Q by using $-P\log(Q) \times sign$.

KL divergence method: We describe the difference between two distributions, P and Q. KL divergence [11]

Table V
FEATURE SETTINGS

Feature category (no.)	a	b	c	d	e	f	g	h	i	j	k	l
Demographic (3)	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
License (1)		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Operation frequency (4)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Variable distribution (170)					✓		✓	✓	✓	✓	✓	
Road width distribution (16)						✓	✓		✓	✓	✓	
Variable distribution by velocity (680)								✓	✓	✓	✓	
Variable distribution by road width (540)										✓	✓	
Variable dist. by velocity and road width (2160)											✓	✓
Number of available features	3	4	4	8	178	24	194	718	874	1414	3574	2168

Table VI
PARAMETERS FOR GRID SEARCH

Kernel	Hyperparameter
Linear	$C : [2^{-5}, ..., 2^{10}],\ w_{accident} : \{1, 2, 3, 5, 10\}$
Polynomial	$C : [2^{-5}, ..., 2^{10}],\ \gamma : [2^{-10}, ..., 2^{3}],$ $degree : \{2, 3\},\ w_{accident} : \{1, 2, 3, 5, 10\}$
Gaussian	$C : [2^{-5}, ..., 2^{10}],\ \gamma : [2^{-10}, ..., 2^{3}],$ $w_{accident} : \{1, 2, 3, 5, 10\}$

is a representative definition for the distance between two distributions: $KL(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$.

We use $P_i \log \frac{P_i}{Q_i}$ of each bin as the feature.

3) Performance evaluation: We tested 12 combinations of features, as shown in Table V, and evaluated performance by 10-fold cross validation. All features were normalized beforehand. Three types of kernel functions (linear, polynomial, Gaussian) with hyperparameters (Table VI) were evaluated in a grid-search manner to achieve the best f-measure. We also used feature selection based on the χ^2 value. The best number of features was determined by the grid search. Once we selected the best parameter setting, we shuffled the driver records and evaluated performance ten times used those parameters.

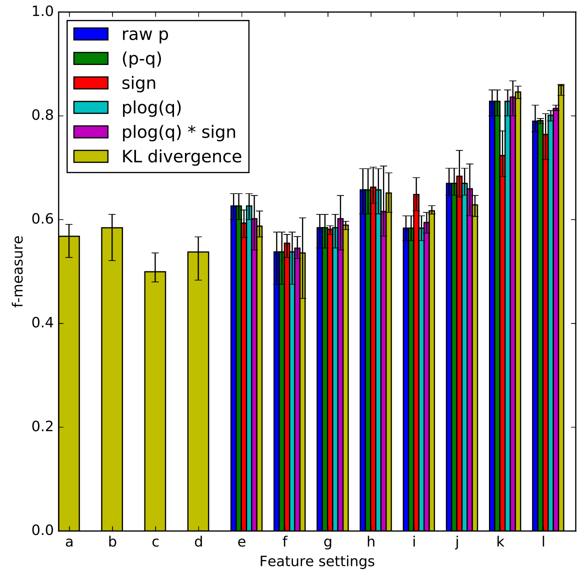

Figure 6. F-measure for different feature settings

Figure 6 shows the best f-measure for each setting. The average results are shown, with the maximum and minimum shown by the error bar. Representative results are shown in Table VII. The precision, recall, f-measure, AUC (area under the ROC curve) are the average results. The random classifier was used as a baseline; it had a precision of 37% (= 43 / 125).

The demographic information was not so helpful in classifying drivers although its use resulted in slightly better performance than the random classifier: the AUC values for settings (a) and (b) were greater than 0.5. Since all the drivers are well-trained professionals, the demographic information may not reflect their driving skills well. When driving operation information was available, the demographic features were generally not selected (settings (e) to (l)) in the best settings by the feature selection process.

The use of the kinematic information obtained from vehicle recorders was helpful in classifying the drivers, as we can see from the results for setting (e). When we took into account the velocity at which the operation was performed, the performance was slightly better (see results for (e) and (h)). Adding the road information to the kinematic information greatly improved performance (see (k) and (l)). However, the road width was not of much help when it was used alone (see (f) and (g)) or combined with simple variable distributions (see (j)).

We evaluated feature expression methods and found that their performance was not much different in general. The "sign" method eliminates much information about frequency, so it makes sense that this method does not achieve good performance for settings (k) and (l). The simplest method, "probability," achieved reasonably good performance for the simple feature settings ((e) to (j)). In contrast, the KL divergence method was effective when the setting was more complicated ((k) and (l)).

Figure 7 shows the ROC curves of representative results. Performance using demographic information was almost the same as that using random classification ((a) and (b)).

Table VII
CLASSIFICATION PERFORMANCE

Setting	Method	No. of selected features	Precision	Recall	F-measure	AUC
a	-	3	0.42	0.88	0.57	0.62
b	-	4	0.45	0.84	0.58	0.65
c	-	4	0.35	0.88	0.50	0.51
d	-	8	0.41	0.79	0.54	0.60
e	p	50	0.50	0.84	0.63	0.70
f	p	24	0.41	0.78	0.54	0.60
g	p	5	0.46	0.80	0.58	0.65
h	p	50	0.55	0.82	0.66	0.73
h	KL	150	0.53	0.84	0.65	0.73
i	p	50	0.42	0.93	0.58	0.63
j	p	50	0.57	0.81	0.67	0.75
j	KL	30	0.47	0.97	0.63	0.69
k	p	100	0.88	0.78	0.83	0.86
k	KL	100	0.86	0.83	0.85	0.88
l	p	50	0.87	0.72	0.79	0.83
l	KL	50	0.94	0.79	0.86	0.88
Random classifier			0.37			

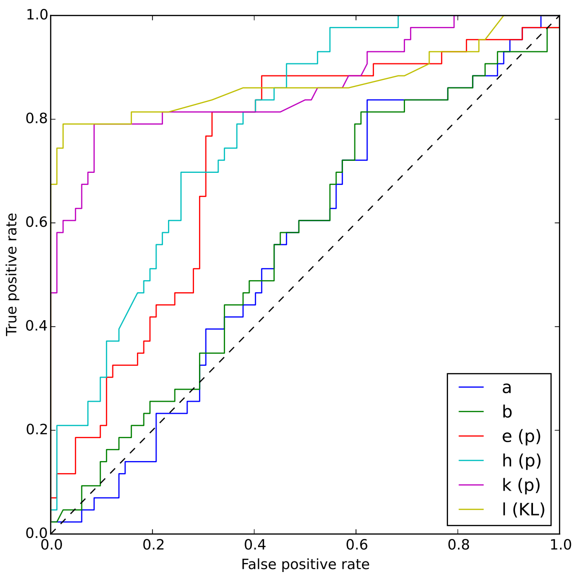

Figure 7. ROC curves of representative results

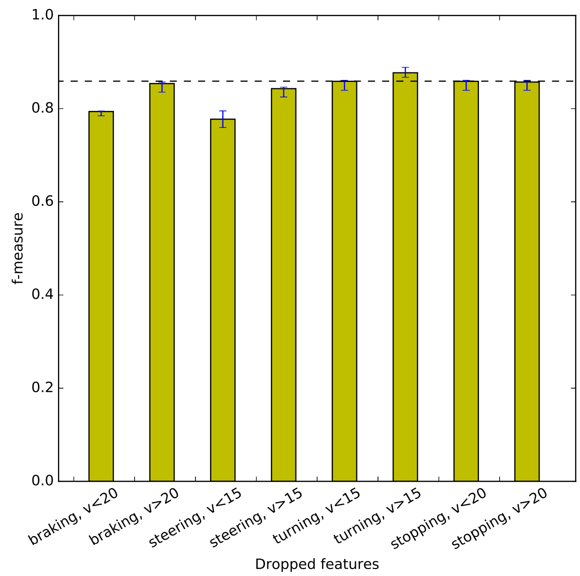

Figure 8. Performance when a feature group was dropped

Adding more information about driving operations improved performance ((e) and (h)). Adding road information achieved the best performance ((k) and (l)); however, doing so changed the shape of the ROC curve. The true positive rate increased quickly until 80% of the unsafe drivers were covered, and then it saturated. This indicates that the proposed method may not be suitable for some of the drivers. This could be because we did not consider the severity of or responsibility for the accidents. Some of the drivers categorized as unsafe may have been involved in only a minor accident. This will be investigated in more detail in future work.

4) Informative driving behaviors: We investigated what kinds of driving behaviors were strongly related to identifying safe drivers. We defined a certain threshold of velocity for each operation and placed each record into a high- or low-velocity group. This resulted in the "variable distribution by velocity and road width" features being categorized into eight groups (four driving operations × two vehicle velocities). We then created eight different feature sets where in each one, one of the feature groups was dropped, and evaluated the resulting performance. The baseline was setting (l) using the KL divergence method. The kernels, hyperparameters, and number of features were independently selected for each feature set by a grid search.

Figure 8 shows the f-measure for each feature set. For example, the bar labeled "braking, $v < 20$" indicates the best result for the feature set from which was dropped the features for braking operations at less than 20 km/h. A threshold value was defined to divide the number of operation records into roughly half (see Figure 3). The dashed line indicates the baseline f-measure of the original feature set. We can see that dropping the "braking, v $<$ 20" and "steering, v $<$ 15" features substantially degraded performance. Of particular interest is that the operations performed at lower velocity had much more information relevant to safe driving. This means that collecting information on all driving behaviors, not only the critical ones, is beneficial.

5) Subjective evaluation by driver managers: We computed a confidence value for each driver's classification result in the best setting and listed the ones most likely to be unsafe drivers. Some of them had not had any accidents, but they were included on the list because their driving operation variables were similar to those of drivers who had had an accident. We wanted to verify whether they really required extra attention. We also wanted to know the current driving characteristics of the listed drivers because our method only takes into account driver history. We thus interviewed the managers of the listed drivers. On the basis of their experience and intuition, they generally agreed with our assessments. They felt that most of the listed drivers needed extra attention.

IV. Conclusion

We intensively examined a large-scale archive of recorded vehicle data to clarify the relationship between safety and driver behavior. We used all driving operation information, which was mostly ignored in previous studies, and used the distributions of the operation values. Our proposed method successfully classified drivers as either safe or unsafe (f-measure $>$ 86%). This level of performance was not achieved with only driver demographic information, which is widely used to estimate drivers' risk. The most informative features were derived from the variables for operations at low velocity. This possibility was often overlooked in previous studies. We also found that taking into account the driving circumstances during operations is an effective way to improve classification performance.

This is the first step toward understanding the relationship between safe driving and driver behavior. Although this work considered only past accidents, the knowledge acquired will be helpful in investigating driver safety and preventing future accidents. We thus plan to apply our method to predicting accidents. Our findings on the characteristics of drivers through their driving behaviors will be helpful in educating drivers.

Our approach, i.e., focusing on the differences in the variables related to driving operations works well. Since the frequencies for the rare-occurrence bins are low, meaning that the related operations are unlikely to occur in the short term (e.g., several days), a daily review of vehicle recorder data is of little use in identifying unsafe behaviors. Regular review of long-term archived data is required.

We plan to continue archiving the vehicle recorder data to monitor long-term behavior changes in the drivers. We also plan to include other information such as geo-location and weather data.

References

[1] B. Higgs and M. Abbas, "A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles," in *16th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2013, pp. 857–862.

[2] R. Dang, F. Zhang, J. Wang, S. Yi, and K. Li, "Analysis of chinese driver's lane change characteristic based on real vehicle tests in highway," in *16th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2013, pp. 1917–1922.

[3] G. Castignani, R. Frank, and T. Engel, "Driver behavior profiling using smartphones," in *16th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2013, pp. 552–557.

[4] F. Guo and Y. Fang, "Individual driver risk assessment using naturalistic driving data," *Accident Analysis & Prevention*, vol. 61, pp. 3 – 9, 2013.

[5] Y. Zheng, J. Wang, X. Li, C. Yu, K. Kodaka, and K. Li, "Driving risk assessment using cluster analysis based on naturalistic driving data," in *17th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 2584–2589.

[6] D. Johnson and M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *14th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2011, pp. 1609–1615.

[7] W. Wu, W. S. Ng, S. Krishnaswamy, and A. Sinha, "To taxi or not to taxi? - enabling personalised and real-time transportation decisions for mobile users," in *IEEE 13th International Conference on Mobile Data Management (MDM)*, Jul. 2012, pp. 320–323.

[8] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi gps traces to social and community dynamics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, pp. 17:1–17:34, Dec. 2013.

[9] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," National Highway Traffic Safety Administration, Tech. Rep. DOT HS 810 594, 2006.

[10] C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, and F. Itakura, "Driver modeling based on driving behavior and its evaluation in driver identification," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 427–437, Feb. 2007.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.