Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment

Umberto Fugiglando, Emanuele Massaro, Paolo Santi, Sebastiano Milardo, Kacem Abida, Rainer Stahlmann, Florian Netter, and Carlo Ratti

Abstract—Cars can nowadays record several thousands of signals through the controller area network (CAN) bus technology and potentially provide real-time information on the car, the driver, and the surrounding environment. This paper proposes a new methodology for near-real-time analysis and classification of driver behavior using a selected subset of CAN bus signals, specifically gas pedal position, brake pedal pressure, steering wheel angle, steering wheel momentum, velocity, RPM, longitudinal and lateral acceleration. Data have been collected in a completely uncontrolled experiment involving 54 people, over 2000 trips have been recorded without any type of predetermined driving instruction on a wide variety of road scenarios. While only few works have analyzed the driving behavior of more than 50 drivers using CAN bus data, we propose an unsupervised learning technique that clusters drivers in different groups, and offers a validation method to test the robustness of clustering in a wide range of experimental settings. The minimal amount of data needed to preserve robust driver clustering is also computed, showing that by properly choosing a subsampling strategy it is possible to reduce the size of the database as much as 99% without impairing the clustering performance.

Index Terms—Driving behavior, CAN bus, feature extraction, unsupervised learning, drivers’ segmentation.

I. INTRODUCTION

MODERN cars are equipped with several hundreds of sensors and electronic control units (ECUs) [1] that, beyond guaranteeing an optimal functioning of the engine, provide the driver with more safety, control and entertainment. These almost real-time data provide information about the car, the driver and the surrounding environment and can be used to study, analyze, predict, and understand a large variety of problems, such as traffic congestion, vehicle energy consumption and emissions, urban mobility and drivers’ habits [2].

This huge amount of diverse data have been made available by the Controller Area Network (CAN), a serial broadcast bus developed by Robert Bosch in 1986 [3] that allows communication among the electronic control units devices mounted on the car. CAN technology has become de facto a standard in car embedded systems [4], providing access to data from an order of several thousands signals, recording at a sub-Hertz frequency information about the car and its surroundings, as evidenced by the dataset analyzed in this paper.

With this technology being implemented in modern cars, the amount of collected data increases with respect to the state of art of GPS-based technologies [5]. All the aforementioned applications can thus be extended and improved. Data availability is not a restrictive aspect anymore as insights from travels can be collected automatically, without the need to modify the car structure or to specifically design an experiment. Moreover, in the present research an order of few gigabytes per hour data stream is leveraged, which represents just a significative sub-sample of all the information traveling on the CAN bus: this amount of data will only increase with the advent of new autonomous driving cars [6].

A. Driving Behavior

Human driving behavior is a complex concept that, in general terms, describes how the driver operates the vehicle controls in the context of the driving scene and external conditions [7]. The characterization of driving behavior is not only crucial for accident prevention, as most of car accidents are due to human mishandling [8], but it is also important for designing driving models, which is one of the core algorithms that might make the future of self-driving cars possible [9]. In fact, autonomous vehicles have to interact with other vehicles (even non-autonomous ones) and understanding their driving style can provide valuable information in order to avoid traffic accidents [10].

Driving behavior characterization is useful also for car insurance companies to quantify accident risk and provide personalized rates [11]. State-of-art technology implements models mostly based on GPS location, traveled distance and coarse grained speed profile [12], [13]. A richer information like the one coming from CAN bus could better characterize human driving behavior and, consequently, accident risk.

In order to be able to use CAN data to characterize drivers in real application scenarios two very challenging problems [7] are to be solved: (1) providing a methodology for consistently identifying driving behavior in a completely uncontrolled environment, and with very limited knowledge of the surrounding conditions; (2) minimizing the communication and computational load needed to solve (1). This paper introduces and
discusses ideas to tackle these challenges and bring CAN bus based driver characterization closer to reality.

More specifically, the goal of the present research is to extract features from CAN bus signals and assess to what extent they are useful for finding similarities among drivers using a clustering algorithm. Given the enormous amount of data generated by the CAN bus – in the order of a few gigabytes of data per hour – it is not feasible to communicate and process the raw output of the CAN bus in real time to characterize drivers. As such, feasibility of the devised driver characterization methodology is bounded to the definition of a strategy to substantially reduce the amount of data to be processed to perform the driver identification task. Thus, in the second part of the paper different data subsampling methods are explored, that allow minimizing data communication between vehicle and infrastructure while guaranteeing robust driver behavior characterization.

The paper is organized as follows. Section II describes the details of the data collection process and the signals considered. Section III is devoted to the clustering of the drivers. Section IV addresses the sampling method question. Finally, section V concludes the paper providing a summary of the future research directions.

B. Related Work

In general, research on driving behavior in scientific literature can be classified according two perspectives: (1) the purpose of the research, e.g. driver recognition, maneuver recognition, aggressive or eco-friendly driving detection, etc. or (2) the data used for the analyses, i.e. GPS locations, CAN bus data, audio-video data, cellular phone data, car simulator data.

Early studies have been made with the aim of characterizing driving behavior by building a dynamic model to eventually implement a control system that would react like a human, to be used for example in self-driving cars. Models have been proposed to anticipate the driver actions by few seconds [14] or to predict the driver’s intended cruising speed up to 20 seconds in advance of reaching that speed [15]. All these works have been validated using data coming from car simulators. Data acquired by a simulator have also been used to quantify the drivers’ skills [16].

Some other work, on the other hand, have been conceived to recognize driving maneuvers (e.g. passing, changing lines, turning, starting and stopping) leveraging CAN data: for example, in [17] the drivers were asked by an instructor in the vehicle to perform given maneuvers.

Carmona et al. [18], through a novel hardware tool designed to integrate data from CAN bus, GPS and an Inertial Measurement Unit (IMU), attempt to classify real-time normal and aggressive driver behavior. The classification was performed in an experiment where 10 drivers have been asked to drive the same route twice, in a normal and aggressive way respectively. The model implemented calculates, for the recorded signal, some statistics on time-based data (mean, standard deviation, maximum and median values) and on frequency-based data, using them as references for real time data.

CAN sensors have also been coupled with external devices, designed and mounted specifically on the vehicle for the purpose of the experiment, like 3D cameras for eye monitoring or wearable devices used to collect biomedical signals. These experiments are more “human-centric” and are aimed at understanding how drivers’ bad habits or distractions are reflected in how they drive: Choi et al. [19] and, lately, Li et al. [20] detected and classified distraction tasks (e.g. tuning the radio, interacting with an automatic voice portal) using audio and video data coupled with CAN bus data.

On the other hand, some work focuses on the driver recognition problem, which attempts to distinguish different drivers only by looking at the CAN bus data. Wakita et al. [21], using data coming from a car simulator, made a comparison between parametric and nonparametric models, concluding that nonparametric approaches perform better in terms of percentages of drivers correctly recognized. Hallac et al. [1] leveraged the same database used in this work achieving a prediction accuracy of 76.9% for two-driver classification, and 50.1% for five drivers, using a random forest model. Miyajima et al. [22], [23] performed driving recognition modeling on pedal operation patterns acquired by CAN bus sensors by means of a cepstral method, both on a car simulator and on real cars involving 276 drivers. However, the exact setting of the experiment, the type of road the drivers navigated, and how they have been instructed to drive is not specified in the paper. Moreover, the vehicle used for data collection (a minivan, [24]), equipped with cameras, microphone, computer rack, power suppliers and amplifiers, suggests that the experimental conditions were far from an everyday context in personal driving.

More recent work uses data coming from mobile phone sensors (accelerometer, gyroscope, magnetometer, GPS, video): in [25], cell phone sensors data have been coupled with CAN bus data as a “ground truth” for isolating acceleration, braking and turning events: the problem of driver recognition is addressed, but the experiment involved only two drivers and reached only 60% of accuracy. Moreover, mobile phones sensors have been used to detect aggressive [26] or drunk [27] drivers.

In contrast to the present research, in which normal cars have been used, most of the previously cited works used cars developed in specific projects, like the UTDrive project [28], [29] or a specifically designed “vehicle corpora for research” [24], [30]. Finally, uncontrolled experimental settings have been used in the SHRP2 Naturalistic Driving Study [31], with the aim of learning more about driver decisions in order to make driving safer, and in another large experiment called EuroFOT (European large scale Field Operational Test on in-vehicle system) [32], where CAN bus data have been used with the only aim of evaluating the impact of 8 different driving assistance systems.

Comprehensive analyses of driving behavior models, tools and experiments can be found in [9], [10], and [20]. For what concerns the size of the data analyzed and the number of drivers involved only in [22], in [17], in the EuroFOT, in the UTDDrive, and in the SHRP2 studies, more than 50 drivers
where considered. Summarizing, none of the existing work analyzes the usage of CAN bus data for driver classification in a completely uncontrolled and open driving environment. Furthermore, the issue of how to reduce the communication and computational load related to driver classification has, to our best knowledge, not been addressed so far.

C. Motivations

As shown in the previous section, the main novelty of this paper in the field of human driving behavior analysis is the combination of (1) large number of drivers, (2) completely uncontrolled experimental settings and (3) quantity of data recorded.

This sets new limits and possibilities to the present research: limits in terms of the variety of the signals acquired, carrying useful information not supported by “ground truth”, i.e. information that can be considered as “true” to which compare the experimental data (for example the “aggressiveness” of the driver, his driving skills or his number of incidents). Furthermore, the subsampling and data reduction technique presented in this paper opens the way to new CAN-based technologies that could find application in real-life scenarios.

II. DATA COLLECTION

A. Experimental Settings

The dataset used in the present research was collected during an experiment carried out by AUDI AG and Audi Electronics Venture. The data collection experiment took place in the city of Ingolstadt (Germany) and involved 64 different drivers, who had not been instructed in any way on the route they had to drive, on the speed or on the behavior they had to follow during the driving. This gives to the present study its unique characteristic of an experiment under uncontrolled testing conditions. A test fleet of ten Audi A3 vehicles was retrofitted with data loggers: this prototype system enables data acquisition for research purposes.

The data collection phase took place in 2014 with a total of 55 days of experiment. Cars were picked up by the drivers in a central deposit and had to be returned within the same day. Each time a user switched on the car engine, the computer registered a new session. A total of 1987 sessions were recorded, and more than 2135 hours of driving data for each of the 2418 sensors were acquired. Each user drove an average of 31 sessions, whose average duration was 64 minutes.

CAN bus signals were recorded on a data logger, stored in a PostgresSQL database and analyzed using the Pandas and Scikit-learn Python modules. The sampling is not uniform due to the particular characteristics of the CAN bus and of the signals. Therefore, high frequency signals are constantly sampled at 20Hz, while low frequency sensors reports their data only when there is a change in their value (e.g. rain sensors, seatbelt sensors, etc.) but for the sake of simplicity all the signals considered in the analysis have been filtered with a low pass filter having a cutoff frequency of 4Hz and resampled at 4Hz through linear interpolation.

1No personal information on the drivers have been recorded.

B. Signals Selection

Among the 2418 signals transmitted on the CAN bus, in this work the analyses focus on eight signals:

- Brake pedal pressure (BRK)
- Gas pedal position (GAS)
- Revolutions per minute (R.P.M.)
- Speed (SPD)
- Steering wheel angle (S.W.A.)
- Steering wheel momentum (S.W.M.)
- Frontal acceleration (F. ACC.)
- Lateral acceleration (L. ACC.)

These signals are directly or, in some cases, indirectly related to the interaction between the driver and the vehicle. For instance, pedals and steering wheel signals directly reflect driver’s movements and actions [33]; some other (speed, RPM and accelerations) represent on a phenomenological point of view quantities that a person can “feel” during the driving and could reflect specific driving habits, e.g. attitude for speeding or harsh accelerations. An example of the collected signals is reported in Figure 1.

III. GROUPING DRIVERS’ BEHAVIOR

In this section a methodology is proposed that allows grouping drivers in a consistent way according to common characteristics, with the aim of finding what are the signals and features that allow to cluster in the most “robust” way the drivers. This methodology is composed of 4 different steps: A) feature extraction, B) feature normalization, C) dimensionality reduction and D) unsupervised clustering.

A. Feature Extraction

Any signal \(x_i\) in the database can be represented as a set of pairs of the type \((x_i, t_i)\), where \(i \in \mathbb{N}\) and \(t_i\) is the timestamp corresponding to the acquisition of the signal value \(x_i\) where \(x_i\) is a floating point number. Similarly to [18], from each considered signal the following 7 indicators are extracted:

1) values of the signal for each sample: \(x_i\).
TABLE I
FEATURES DEFINITION

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Values of the signal for each sample</td>
</tr>
<tr>
<td>2</td>
<td>Difference quotient (discrete first derivative)</td>
</tr>
<tr>
<td>3</td>
<td>Time interval between two singular points</td>
</tr>
<tr>
<td>4</td>
<td>Values of the local maxima</td>
</tr>
<tr>
<td>5</td>
<td>Moving mean</td>
</tr>
<tr>
<td>6</td>
<td>Moving median</td>
</tr>
<tr>
<td>7</td>
<td>Moving standard deviation</td>
</tr>
</tbody>
</table>

2) difference quotient (discrete first derivative) of the signal between two consecutive samples: \( \frac{x_{i-1} - x_i}{t_{i-1} - t_i} \). This measure quantifies the intensity of signal variation over time. Let us now define \( J \) as the set of indexes for which the values \( x_i \) are singular points (local maxima or minima), i.e., \( J = \{ i : (x_i - x_{i-1})(x_{i+1} - x_i) < 0 \} \), and by \( J_{\text{max}} \subset J \) the set of only local maxima. Moreover, let us define on those sets a relation \( \prec \), where \( j \prec k \) means that \( j \) is the largest element of the set that precedes \( k \), i.e. \( j = \max\{i \in J : i < k\} \).

3) time interval between two singular points: \( t_j - t_k \), \( j, k \in J \), \( j \prec k \). This feature represents the frequency of its peak points, or in other words the rapidity of variation of the signal when it reaches extreme values.

4) value of the local maxima: \( x_j \), \( j \in J_{\text{max}} \). This feature provides the intensity of the extreme values of the signal.

In a temporal window of one minute\(^ 2 \) and remembering the 4 Hz sampling we define the set of indexes \( I_i = \{ i - 120, \ldots, i + 120 \} \) and the following. resume

1) moving mean, averaging the values \( x_i \) over a temporal window of 1 minute: \( \bar{x}_i = \frac{1}{240} \sum_{j \in I_i} x_j \).
2) moving median, the median value of the set \( \bigcup_{j \in I_i} x_j \).
3) moving standard deviation, the variance of the values in the set \( \bigcup_{j \in I_i} x_j \).

Table I summarizes the features defined above for a quick reference, while Figure 2 shows a plot of a sample signal and some of the features.

B. Feature Normalization

For any given signal \( x \) of floating point type, \( w^{k,u} \) denotes the vector of the feature \( k \) for user \( u \), obtained by calculating the functions defined above on the vector \( x \), joining all the sessions of the same user. Outlier removal was done by keeping only the values between the 2nd and 98th percentile. The vector \( w^{k,u} \) can be considered as a set of statistical samples that are used to build frequency histograms.

Then, each vector \( w^{k,u} \) is normalized in the following way. In order to get for each user histograms with the same bins,

\[
W^k = \bigcup_{u \in U} \bigcup_{i} \{ w^{k,u}_i \}
\]

where \( U \) is the set of users, and the interval \([\min W^k, \max W^k] \) is partitioned into 10 equal intervals\(^ 3 \) (bins) \( b_1, \ldots, b_{10} \). Then, for each user and for each indicator, the histogram \( H^{k,u} \) for the vector \( w^{k,u} \) with bins \( b_1, \ldots, b_{10} \) can now be computed, i.e. each bar of the histogram has a value \( h_i^{k,u} \) which is the number of items of the vector \( w^{k,u} \) belonging to interval \( b_i \).

Finally, all the histograms are normalized, obtaining new values \( \tilde{h}_1^{k,u}, \ldots, \tilde{h}_{10}^{k,u} \) according to the formula

\[
\tilde{h}_i^{k,u} = \frac{h_i^{k,u}}{\sum_{j=1}^{10} h_j^{k,u}}
\]

so that \( \sum_{i=1}^{10} \tilde{h}_i^{k,u} = 1 \).

According to the previous definition, features in form of histograms can be interpreted as a discrete version of the sample distributions of the indicator vectors. This definition, along with its probabilistic interpretation, has two main advantages: it allows performing analyses on objects which have a probabilistic meaning, while on the other hand it keeps machine learning algorithms relatively simple due to the low dimensionality of the data.

It is important to underline that some of the histograms for a signal show similar values, therefore, by using the Pearson product-moment correlation coefficients, it is possible to measure how similar these values are. An example of such analysis is reported in Figure 3, where the average correlation between these vectors for the steering wheel angle signal is shown. Anyway, the decision of considering partially overlapping features aims at gaining a better understanding of what are the best signals to use, because in some cases small differences in a specific feature can be used to better clusterize the drivers.

\(^3\)The number 10 has been chosen after some preliminary analyses. The rationale for choosing the number of bins was to have a sufficient number of bins to well represent the shape of the probability density distribution, but small enough to keep the computation of the machine learning algorithms feasible.
In the following analyses, for data homogeneity users are considered who drove in total at least 10 hours, reducing the number of considered users to 54 from the initial 64 involved in the data collection process.

C. Dimensionality Reduction

In this section the $K$-means clustering algorithm [34] is used to leverage the features defined in the previous section with the aim of grouping drivers according to common similarities. This is a novel approach in this field and therefore it requires an assessment of the validity of the method in terms of robustness and scalability.

It is worth remarking that the vectors $H^k_u$ are 10-dimensional data-points, being represented by histograms with 10 bins. In order to plot them on bi-dimensional space, therefore, a dimensionality reduction technique has to be performed. In this work we use Principal Component Analysis (PCA), a well known statistical procedure that decreases the dimensionality of a space projecting it into another one, whose dimensions (principal components) are orthogonal to each other and such that the variance of the projected data-points on the principal components is maximized [34]. It is worth remarking that PCA is performed just for visualization purposes, while the machine learning algorithms use the original 10-dimensional data-points as inputs.

Table II shows that for most of the combinations of signals and features, the first two principal components can account for more than 80% of the total variance of the original high dimensional data. Figure 4, consequently, reports an example of a bidimensional representation of the features (Feature 1 for the gas pedal signal) where each dot corresponds to a driver. It can be noticed that there are no well separated clusters: this can be expected considering that human behavior typically varies in a range that forms a continuum. For this reason, the word “segmentation” more accurately describes this process than “clustering”: some common behavior can be identified, while some “outliers” slightly deviate from the average.

<table>
<thead>
<tr>
<th>Features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRK</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>0.96</td>
<td>1.00</td>
<td>0.66</td>
<td>0.89</td>
</tr>
<tr>
<td>GAS</td>
<td>0.90</td>
<td>0.98</td>
<td>0.93</td>
<td>0.85</td>
<td>0.79</td>
<td>0.96</td>
<td>0.78</td>
</tr>
<tr>
<td>R.P.M.</td>
<td>0.61</td>
<td>0.95</td>
<td>0.57</td>
<td>0.78</td>
<td>0.70</td>
<td>0.98</td>
<td>0.73</td>
</tr>
<tr>
<td>SPD</td>
<td>0.61</td>
<td>0.88</td>
<td>0.54</td>
<td>0.77</td>
<td>0.55</td>
<td>0.91</td>
<td>0.65</td>
</tr>
<tr>
<td>S.W.A.</td>
<td>0.92</td>
<td>0.99</td>
<td>0.92</td>
<td>0.97</td>
<td>0.95</td>
<td>0.97</td>
<td>0.80</td>
</tr>
<tr>
<td>S.W.M.</td>
<td>0.79</td>
<td>0.96</td>
<td>0.79</td>
<td>0.94</td>
<td>0.89</td>
<td>0.98</td>
<td>0.88</td>
</tr>
<tr>
<td>F. ACC.</td>
<td>0.82</td>
<td>0.94</td>
<td>0.76</td>
<td>0.81</td>
<td>0.87</td>
<td>0.97</td>
<td>0.75</td>
</tr>
<tr>
<td>L. ACC.</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Inspired by the widely used method of cross-validation used in supervised learning, a new approach is here proposed for establishing the optimal number of clusters, based on the concept of “robustness” of the clustering to the road sampling. In fact, remembering that the clusters are made...
up of distributions that come from sampled data, the clusters should be invariant to a subsampling of the original data. In other words, comparing the clusters generated by different subsampling of the original data, those clusters should be similar.

The method proposed is described in Algorithm 1 and can be synthesized as follows. For each feature \( k \) and for each feature \( u \), the vector \( w^{k,u} \) is divided into two different vectors: 70% of its components, taken randomly, form the vector \( w^{k,u}_v \) (training vectors), while the other 30% form the vector \( w^{k,u}_v \) (validation vectors). After having computed the histograms for the two sets of vectors, a \( K \)-means cluster algorithm is performed separately on both the training set and the validation set, producing two different clusterings of the same set of drivers. These two clusterings are then compared using a metric called “V-measure” [36], a score ranging from 0 to 1 and evaluating the similarity of the clusterings: if the clusterings are exactly the same (except for permutations on the labels of each cluster) the score is 1, while the score is closer to 0 as the clusterings are more dissimilar. These operations are repeated for a number of clusters \( K \) ranging from 2 to 10. Moreover, the subsampling being random, for each value of \( K \) the algorithm is repeated 40 times: averages and standard deviations of the scores for each value of \( K \) are calculated and lead to plots like the ones in Figure 5A.

Therefore, for each feature and signal, the number of clusters is chosen. Table III provides the chosen values together with the mean and variance of their corresponding V-measures. In case of ties of the V-measure, the lowest value of \( K \) has been considered as the optimal one.

Results clearly show that there are some numbers of clusters that separate users in a better way in terms of “robustness.” For example, feature 2 for the gas pedal position separates drivers in two different groups, which stays exactly the same in all the 40 repetitions of the cross-validation algorithm, whilst it is not the same for \( K = 4 \).

Overall, some features and some signals perform better than others: the brake pressure signal is the one with the most promising results, followed by the gas pedal position and the steering wheel. This is a first important result, as it confirms what has already been found in the literature with data from an unstructured experiment [22]. Perhaps not surprisingly, these are signals generated through direct interaction between the driver and the car, indicating a better robustness of those versus signals that are generated through the mediation of the car response (acceleration, RPM, etc.).

Finally, Figure 5B reports the results of the \( K \)-means clustering for the brake pedal signal, the gas pedal signal, and the steering wheel momentum signal (see Figure 7 for the remaining signals), with values of \( K \) as in Table III.

IV. DATASET REDUCTION

Once it is verified that a consistent, robust clustering of drivers is possible also in completely uncontrolled, open traffic conditions, the second fundamental aspect for real-life application is tackled: the best sampling method and the minimum amount of data required to provide consistent results. In fact, state-of-the-art technology in car communication uses mobile connectivity to stream data from the car to the server where they are processed, and given the massive volume of the sampled data it is crucial to investigate a lower-bound for this data communication. Furthermore, identifying the minimum amount of data necessary to consistently characterize driver behavior is useful to set boundaries for driver privacy when using car-collected data, and to assess robustness to security attacks. Two methods that involve different spatiotemporal sampling of the data are compared, and the quality of the clustering with different quantities of analyzed data is studied.

The subsampling of the vectors \( w^{k,u} \) presented in Section III-D is completely random and does not consider any spatial or temporal dimension: in other words, it is an independent subsampling. This is compared with a different subsampling strategy, which we call contiguous subsampling, a subsampling conditioned to spatial contiguity defined as follows. Given the vector \( w^{k,u} \) of dimension \( d \), a random number \( r \in \mathbb{N} \) is extracted uniformly in the interval \([1, d]\). Setting \( l = \lfloor pd \rfloor \), where \( p \in (0, 1) \) is the percentage of the elements to be subsampled, the vector \( w^{k,u}_s \) is constructed considering the elements of \( w^{k,u} \) with indexes from \( r \) to \((r + l) \mod d \). In other words, the vector is subsampled selecting a random element and its \( l \) consecutive elements, considering the vector with a circular structure.

For each of the two subsampling strategies defined, an analysis is proposed that compares the clusterizations generated in two different ways: in the first, drivers are clustered using all the data in the dataset, i.e. data coming from all the roads they have driven on; in the second, drivers are clustered based on a portion of the data acquired. In this way,
Fig. 5. Plot of the analyses for selected combinations of signals and features (Brake, feature 5; Gas, feature 2; S.W.M., feature 7): (A) Output of Algorithm 1, plotting the V-measure for different values of $K$; (B) Drivers clusterings. The $K$-means algorithm has been run on all data in the database and for the optimal values of $K$ as in Table III.

Fig. 6. Subsampling methods for selected combinations of signals and features (Brake, feature 5; Gas, feature 2; S.W.M., feature 7): the graphs show the V-measures of the comparisons of the $K$-means clusters generated using all the data in the database, with the clusters generated by a subset of the data (validation set), for different sizes of the validation set (100%, 50%, 20% 10%, 5%, 2%, 1% of the original data) and for the optimal values of $K$ as in Table III.

Fig. 6 reports the results of the V-measure comparisons of the clusterings generated using all the data in the database with the clusterings generated by a subset of the data, for different sizes of subsets and for the two aforementioned subsampling methods. Every subsampling has been repeated 40 times with different random numbers and the $K$-means clusterings have been performed for each feature with the optimal value of $K$ found earlier.

Results clearly show that the independent subsampling strategy performs better than the contiguous one, and for some features and signals it is possible to reduce the original

TABLE III

<table>
<thead>
<tr>
<th>Features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAKE</td>
<td>2 (0.95, 0.11)</td>
<td>4 (0.99, 0.01)</td>
<td>2 (1.00, 0.00)</td>
<td>5 (1.00, 0.01)</td>
<td>3 (1.00, 0.01)</td>
<td>3 (0.95, 0.05)</td>
<td>2 (0.92, 0.07)</td>
</tr>
<tr>
<td>GAS</td>
<td>2 (0.96, 0.06)</td>
<td>2 (1.00, 0.00)</td>
<td>2 (0.93, 0.06)</td>
<td>4 (0.98, 0.03)</td>
<td>2 (1.00, 0.00)</td>
<td>2 (0.99, 0.03)</td>
<td>2 (0.99, 0.03)</td>
</tr>
<tr>
<td>R.P.M.</td>
<td>3 (0.99, 0.02)</td>
<td>2 (0.98, 0.05)</td>
<td>2 (0.85, 0.06)</td>
<td>2 (1.00, 0.00)</td>
<td>2 (1.00, 0.00)</td>
<td>6 (0.71, 0.06)</td>
<td>2 (0.92, 0.08)</td>
</tr>
<tr>
<td>SPEED</td>
<td>2 (1.00, 0.00)</td>
<td>2 (1.00, 0.02)</td>
<td>3 (0.81, 0.12)</td>
<td>2 (0.98, 0.05)</td>
<td>2 (0.93, 0.06)</td>
<td>6 (0.72, 0.04)</td>
<td>2 (0.86, 0.09)</td>
</tr>
<tr>
<td>S.W.A.</td>
<td>2 (0.98, 0.05)</td>
<td>5 (0.99, 0.02)</td>
<td>4 (0.78, 0.08)</td>
<td>2 (0.99, 0.09)</td>
<td>4 (1.00, 0.00)</td>
<td>2 (0.92, 0.14)</td>
<td>3 (0.97, 0.05)</td>
</tr>
<tr>
<td>S.W.M.</td>
<td>3 (1.00, 0.00)</td>
<td>2 (0.96, 0.06)</td>
<td>4 (0.91, 0.05)</td>
<td>2 (1.00, 0.02)</td>
<td>2 (0.92, 0.09)</td>
<td>2 (0.96, 0.06)</td>
<td>2 (1.00, 0.00)</td>
</tr>
<tr>
<td>F.ACC.</td>
<td>4 (0.98, 0.05)</td>
<td>6 (0.93, 0.06)</td>
<td>2 (0.88, 0.09)</td>
<td>5 (0.87, 0.07)</td>
<td>2 (0.98, 0.05)</td>
<td>2 (0.82, 0.09)</td>
<td>2 (1.00, 0.00)</td>
</tr>
<tr>
<td>L.ACC.</td>
<td>3 (0.99, 0.04)</td>
<td>2 (0.83, 0.09)</td>
<td>2 (0.86, 0.10)</td>
<td>2 (0.92, 0.12)</td>
<td>2 (0.94, 0.08)</td>
<td>2 (0.80, 0.10)</td>
<td>2 (0.97, 0.08)</td>
</tr>
</tbody>
</table>

the first clustering can be considered somehow as a ground truth (being the result of all the data available to us), while the second is the result of a partial subsampling.
Fig. 7. $K$-means clusterings for each signal and each feature considered. Every point of a subgraph represents a driver and the number of clusters has been selected according to the optimal values of $K$ as in Table III.
Fig. 8. Comparison the independent subsampling (red line, diamonds) and the contiguous subsampling (black line, circles). The V-measures reported is calculated on the clusters identified using the complete signal and the subsampled version of it.
dataset by a factor as high as 99% without impairing clustering performance. A comprehensive chart for all the combinations of signals and features can be found in Figure 8.

V. CONCLUSIONS

A. Results

In this paper, the problem of driving behavior analysis has been studied from a new point of view, that bridges the gap between driving behavior studies through uncontrolled experiments – leveraging only the GPS signal – and studies exploiting CAN bus data through very controlled experiments. This work proposes a methodology for delineating similarities among drivers using data collected in a completely uncontrolled experiment, through a clustering algorithm performed on seven different features of eight signals recorded by CAN bus sensors, with a distributional approach.

Results show that an optimal number of cluster can be identified and specific combinations of signal-feature provide very high performances in terms of “robustness” (as defined in III-D). For each of the considered signal there is at least one feature which gives a mean of the V-measure values of at least 0.99 (see Table III); the same result holds vice versa (for each feature there is at least one signal). Moreover, it has been shown that, by properly choosing the subsampling strategy, it is possible to reduce the size of the dataset by as much as 99% without impairing clustering performance.

B. Discussion

Given the almost ontological question of what driver behavior is, this work attempts to define it through a data-driven approach. Without any external knowledge (ground truth), though, it is unclear how to define the boundary between the performance of the proposed method and the fuzziness and the unpredictability of human behavior. However, the promising results obtained in this study suggest that the present approach could be considered as a methodology for testing new signals, features and clustering methods which, coupled with additional field knowledge, may lead to pragmatic interpretations of the different clusters in terms of physical and behavioral characterization of driving styles.

It is important also to outline some limitations of this work: the number of users, 64 later reduced to 54 for data homogeneity reasons, likely does not offer a rich enough variety of driving behaviors to enable a comprehensive identification of common attitudes and outliers. Furthermore, the proposed location independent approach might neglect potential association between time/location and specific driving behaviors. Finally, an aspect that needs further investigation is the interaction of the different indicators and the signals directly in the clustering process.

C. Applications and Future Work

This paper projects the problem of driving behavior characterization using CAN bus technology from a research-oriented approach into an application-oriented technology that opens the way to wide-scale and real-time implementations. In fact, as mentioned, the presence of the CAN bus data in almost every car could scale-up any possible application in a very broad and cost-effective way.

Car insurance companies, for example, are interested in assessing the risk of accidents for each user based on real data coming from their driving sessions [37]. User segmentation in fact, to the best of our knowledge, today is only performed – besides the accident history – on general information like the geographical location, distance traveled, and velocity. More sophisticated concepts like “aggressiveness” or “nervousness” could be fully characterized. However, in order to do so, further studies have to be performed, comparing the insurance companies’ driver profiles with the clustering obtained in this work, allowing their characterization based on a ground truth.

Another application is driver recognition, aiming to recognize a driver only based on the CAN bus data. This driver “fingerprint,” already studied [33] but never tested in an uncontrolled experimental scenario, could let the car itself identify the driver for security reasons or adapt settings for comfort or efficiency optimization.

Finally, integration of this modeling technique with physical detection technologies including sonar devices, stereo cameras, lasers and radar would allow better understanding and modeling of driver behaviors, to improve the development of self-driving cars and to have safer road networks.

ACKNOWLEDGMENT

The authors would like to thank Volkswagen ERL, Cisco, SNCF Gares & Connexions, Brose, Allianz, UBER, Fondation OCP, Liberty Mutual, Ericsson, Saudi Telecom, Volkswagen Group America, Philips, Austrian Institute of Technology, Fraunhofer Institute, Kuwait-MIT Center for Natural Resources, SMART-Singapore MIT Alliance for Research and Technology, AMS Institute Amsterdam, Victoria State Government, and all the members of the MIT Senseable City Lab Consortium for supporting this research.

PRIVACY DISCLAIMER

The data reported herein was collected during experiments performed with drivers who were hired and were explicitly informed of the data collection process. In case the presented methodology should be used with consumer vehicles, it is fundamental to properly inform the customer about usage of data and the purpose of the collection. This needs to be done in order to comply with data privacy laws and regulations, but also to support customers’ awareness and self-determination – especially in cases where the realization of an application requires providing personal data to third parties. It is the decision of the customer based on a declaration of consent, if personal data may be collected and for which purpose it may be used.

REFERENCES

Paolo Santi received the Laurea and Ph.D. degrees in computer science from the University of Pisa, Italy. He is currently a Research Scientist with the MIT Senseable City Laboratory, where he leads the MIT/Fraunhofer Ambient Mobility initiative, and a Senior Research with the Istituto di Informatica e Telematica, CNR, Pisa. His research interest is in the modeling and analysis of complex systems ranging from wireless multi hop networks to sensor and vehicular networks and, more recently, smart mobility, and intelligent transportation systems. In these fields, he has contributed over 120 scientific papers and two books. He is a member of the IEEE Computer Society and has recently been recognized as a Distinguished Scientist by the Association for Computing Machinery. He has been involved in the technical and organizing committee of several conferences in the field, and he is/has been an Associate Editor of the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and Computer Networks. He was a Guest Editor of the Proceeding of the IEEE special issue on Vehicular Communications: Ubiquitous Networks for Sustainable Mobility in 2011, to which he also contributed a paper.

Sebastiano Milardo received the bachelor’s and master’s degrees in computer engineering from the University of Catania in 2011 and 2013, respectively, and the Ph.D. degree in information and communication technologies from the University of Palermo in 2018. From 2014 to 2015, he was with the Italian National Consortium of Telecommunications, as a Researcher within the NEWCOM# and SIGMA Projects. His research interests include software-defined networking, sensor networks, network protocols for the Internet of Things, and big data.

Kacem Abida received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada. He is currently a Senior Engineer with the Volkswagen Group of America, Electronics Research Laboratory, where he is also leading the big data projects. His areas of interest include speech and natural language technologies and machine learning-based big data analytics.

Rainer Stahlmann received the Diploma degree in electrical engineering and computer science from the Ingolstadt University of Applied Sciences, Germany, in 2009. He is currently pursuing the Ph.D. degree with the Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany. Since 2019, he has been with AUDI AG, Ingolstadt, Germany, where he is currently with the Department of Data Strategy and Analytic Services. In cooperation with the Chair for Computer Networks and Communication Systems with the Friedrich-Alexander-Universität Erlangen-Nürnberg. His research is focused on vehicular data processing and analytics and on technical evaluation of V2X communication systems.

Florian Netter received the Diploma degree in mechanical engineering from the Technical University of Munich, Germany, in 2010, and the Ph.D. degree in cooperation from the Karlsruhe Institute of Technology, Germany, in 2015. Since 2010, he has been with AUDI AG, Ingolstadt, Germany, where he is currently a Senior Engineer. He is currently with the Department of Data Analysis involved in vehicular data stream processing in cloud computing environments. During his research he focused on complexity adaptation of simulation models in entire system simulations to identify quantification attributes for a high goodness of fit and in spite of increasing computational power still maintaining a short simulation period.

Carlo Ratti graduated from the Politecnico di Torino and the École nationale des ponts et chaussées and received degrees in civil engineering and received the M.Phil. and Ph.D. degrees with the University of Cambridge, U.K. He is currently the Founder and the Director of the MIT Senseable City Laboratory. As an Architect and Engineer by training, he practices in Italy and teaches at the Massachusetts Institute of Technology. He has co-authored over 200 publications and holds several patents. His work has been exhibited worldwide at venues such as the Venice Biennale, the Design Museum of Barcelona, Barcelona, the Science Museum, London, GAFTA, San Francisco, and The Museum of Modern Art, New York. His Digital Water Pavilion at the 2008 World Expo was hailed by Time Magazine as one of the Best Inventions of the Year. He has been included in Esquire Magazine’s Best and Brightest list, in Blueprint Magazine’s 25 People Who Will Change the World of Design, and in Forbes Magazine’s Names You Need To Know in 2011. He was a Presenter at TED 2011 and is serving as a member of the World Economic Forum Global Agenda Council for Urban Management. He is a regular Contributor to the architecture magazine Domus and the Italian newspaper Il Sole 24 Ore. He has also written as an op-ed contributor for BBC, La Stampa, Scientific American, and The New York Times.