Characterizing the “Driver DNA” Through CAN Bus Data Analysis

Umberto Fugiglando
MIT Senseable City Laboratory
Cambridge, Massachusetts
umbertof@mit.edu

Paolo Santi
Istituto di Informatica e Telematica
del CNR
Pisa, Italy
psanti@mit.edu

Sebastiano Milardo
University of Palermo
Palermo, Italy
milardo@mit.edu

Kacem Abida
VW Group Electronics Research Laboratory
Belmont, California
kacem.abida@vw.com

Carlo Ratti
MIT Senseable City Laboratory
Cambridge, Massachusetts
ratti@mit.edu

ABSTRACT
People’s driving behavior is influenced by different human and environmental factors, and several attempts to characterize it have been proposed. Nowadays, the standardization of the CAN bus and the increase of the electronic components units in modern cars offer a large availability of sensors data that make possible a more reliable and direct characterization of driving styles. In this work, we propose the concept of “Driving DNA” as a way of describing the complexity of driving behavior through a set of individual and easy-to-measure quantities. These quantities are responsible for some aspects of the driver’s behavior, just as – in the metaphor – genes are responsible for the tracts of an individual.

The concept has been tested on a dataset collected on the driving of more than 2000 trips by 53 people, in a wide scenario of road types and traffic conditions. The Driving DNAs have been calculated for each person, and a graphical visualization of their comparison is provided.

CCS CONCEPTS
• Computer systems organization → Embedded systems; Redundancy; Robotics; • Networks → Network reliability;

KEYWORDS
Driving behavior, Driving DNA, CAN bus, Accident risk, Fuel efficiency

1 INTRODUCTION
The process of understanding and characterizing human driving behavior has recently gained a lot of importance due to its fundamental applications in connected autonomous vehicles, electric vehicles and artificial transportation systems [11, 17]. Moreover, as driving style influences accident risk [1] and fuel efficiency [10], technologies able to classify driving behavior can improve safety and car eco-friendliness.

Differently from earlier studies, where models were based only on GPS location [7], these can now rely on multiple layers on information coming from several hundreds of sensors and electronic control units (ECUs) embedded in the car, whose intercommunication is made feasible through the CAN bus technology [9]. This not only implies richer and higher quality data, but expands the possibilities of large-scale data collections and extensive sensing applications [12].

Driving behavior (or driving style) has no unique definition nor measure, and it is a combinations of mixed factors and components [11]. In this work, we propose the concept of driving DNA, looking at individual driving behavior as a global resultant of single easy-to-measure characteristics (genes) that describe some specific aspects of driving attitudes. This can be seen as a process in which driving behavior is reduced to some single dimensions that can be characterized and associated to measurable quantities.

More in general, the purpose of the present work is to show that (1) it is possible, thanks to the CAN technology, to compute and visualize individual driving attitudes, in form of driving DNA; (2) different people have different driving behaviors (DNAs) and it is possible to visualize and compare them; (3) ultimately, driving DNA is useful for educative purposes in order to achieve a better driving style.

The concept of driving DNA has been engineered and tested upon a large dataset of CAN bus signals recorded in an uncontrolled experiment, that involved 53 different people and took place on a wide scenario of road types and traffic conditions.

2 RELATED WORKS
Previous works attempting to study driving skills and driving behavior implemented models tested only on driving simulators [18].
Studies dealing with data acquired with real cars, however, involve very controlled experimental scenarios and different from the everyday driving conditions. For example, Oliver et al. [16] built a system recognizing driving maneuvers performed by the driver upon some directions by an instructor physically present in the car, while Carmona et al. [4] classified normal and aggressive drivers asking to 10 people to drive the same route twice, in a normal and aggressive way respectively.

One of the few studies that broke the barriers of data collections in a controlled environment is the one by Miyajima et al. [13], where models of driver recognition and pedal operation pattern were tested on a fleet of real cars involving 276 drivers. Moreover, Castignani et al. [5], similarly to [4], attempted to characterize driving aggressiveness using adaptive profiling mechanism instead of fixed thresholds: the fuzzy logic algorithm implemented, however, has the limitation that it is based on a 20 minutes calibration phase, which can be intentionally diverted by the driver.

Larger experiments, like the SHRP2 project [2], analyzed the driving of a big number of people using traditional technologies, though not using CAN bus data. Moreover, in a previous work [6], the authors of this paper leveraged the present database in order to cluster drivers according to their driving behaviors using unsupervised learning techniques. Clusters were not characterized as the database lacked of a ground truth.

The present work aims to overcome all the aforementioned limitations, using a large CAN bus signals database collected in a uncontrolled experiment and coupling together different aspects of driving behavior which so far have been treated separately, like fuel efficiency, aggressive driving and accident risk.

3 DATASET

The dataset used in this work is made up of more than 2000 CAN signals recorded from 10 different cars retrofit with a data-logger³. The data acquisition phase has been performed in 2014 by Audi AG and Audi Electronics Venture in the German city of Ingolstadt. A total number of 53 drivers have been involved in the data collection, providing a rich dataset of more than 2135 hours of driving over 55 days of experiments.

The datatype of the signals recorded is boolean, char, integer or float, and their sizes vary from few Mb to few Gb for each sensor. The original sampling frequency varies (up to 20 Hz), but for computation and visualization purposes all floating point signals have been downsampled at 4 Hz, significantly reducing the size of the database to few Mb of data per hour per car. In fact, considering that the data collected on the car have to be processed on-board or transmitted to a remote server via cellular network, it is important to perform a preprocessing in order to keep the size of the database as low as possible. This is particularly important as the amount of data collected will increase with the advent of autonomous driving cars [14].

In the following, we will denote any given signal as x, whose components x_i are the samples of the given quantity (e.g. speed, GPS position, RPM, acceleration) at a constant rate of 4 Hz.

4 DRIVING DNA

In molecular biology, genes are portions of the DNA sequence which affect an organism’s traits [3]. Looking at driving behavior as the expression of both exogenous and endogenous factors [11] (being the former driving skills, demographic, distractions, fatigue, etc.; the latter road type, weather and traffic conditions, etc.), we can interpret the Driving DNA as the mixed factors that determine, coupled with the external environment, a person’s driving attitudes.

In the present work we try to identify different dimensions or factors – which we call DNA dimensions (genes in the metaphor) – representing single characteristics of driving behavior, determined by measurable quantities upon specific rules. In the following, we will describe four dimensions: braking, turning, speeding, fuel efficiency. Each of them is associated with a score, ranging from 0 to 5, reflecting the “goodness” of the single factor: the higher the score is, the best behavior is achieved by the driver.

The scores are calculated individually for each driver, for each DNA dimension and for every signal sample, based upon some specific formulas defined in the following. We will call them instant scores, as they change at a high rate and can be implemented in a system that shows to the driver in “real time” the variation of the specific dimension.

The instant scores, separately for each of the DNA dimension, are then averaged in a suitable way over the sessions of the same driver, in order to come up with robust values that fully characterize each person.

In the following, the four proposed DNA dimensions are described.

4.1 Braking (cautious driving)

Signals used: frontal acceleration.

This measure represents the intensity of the braking actions, a measure that is related to aggressiveness, comfort and ultimately safety.

Given the frontal acceleration signal x, each value x_i is averaged considering other samples falling in a two-second moving temporal window, i.e. the set of samples \( U_i = \{x_{i-4}, \ldots, x_{i+4}\} \). For each sample \( x_i \), the averaged value follows the formula \( \bar{x}_i = \frac{1}{2}(\max U_i + \text{median} U_i) \), providing a value that represents both median and extreme values in the braking patterns over the two considered seconds.

In order to provide an instant score that ranges from 0 to 5 and able to assess the braking intensity, the values \( \bar{x}_i \) are then normalized over a percentile distribution considering all drivers in the database. In this way, a low score would represent a braking pattern worst than the average, a high score instead would refer a better one. More precisely, considering the distribution of the \( \bar{x}_i \) for this specific DNA dimension and all users, we consider the six quantiles \( q_1, \ldots, q_6 \) and assign to \( \bar{x}_i \) a score \( k \in \{0, 1, \ldots, 5\} \) such that \( q_k < \bar{x}_i \leq q_{k+1} \) (being \( q_0 \) the least number in the distribution) and the score of 0 in the case of \( \bar{x}_i = q_0 \).

The quantiles \( q_k \) could be evenly distributed in the interval \([0, 1]\), or could reflect a different distribution: in this way higher and lower scores would be assigned with different frequencies and therefore

³No personal information on the drivers have been recorded.
particular behaviors would be enhanced with positive/negative scores.

### 4.2 Turning (attentive driving)

**Signals used:** lateral acceleration.

A similar process to the one described above for the braking pattern is implemented for the lateral acceleration, a measure that describe the aggressiveness in turning and more in general in the use of the steering wheel, especially at high speeds.

### 4.3 Speeding (safe driving)

**Signals used:** speed, rain, GPS (road speed limit).

The speeding dimension represents the risk of accident induced by speeding, especially in case of adverse weather conditions. In fact, it has been calculated that the accident risk increases by 50% in case of rain [8, 15].

The measure is calculated as follows: for each timestamp \( t_i \), the function \( f(t_i) \) is computed as

\[
f(t_i) = \begin{cases} \frac{\frac{v(t_i)}{v_{\text{max}}(t_i)}}{1 + \frac{1}{2}r(t_i)}^2 & \text{if } v(t_i) > v_{\text{max}}(t_i) \\ 1 & \text{otherwise} \end{cases}
\]

where \( v(t_i) \) is the speed of the car at time \( t_i \), \( v_{\text{max}}(t_i) \) is the speed limit\(^2\) of the road where the car was travelling at time \( t_i \), and \( r(t_i) \) is a boolean function returning the weather condition\(^3\) in the location of the car at time \( t_i \), defined as

\[
r(t_i) = \begin{cases} 1 & \text{if it is raining in } t_i \\ 0 & \text{if it is not raining in } t_i \end{cases}
\]

The value of the function \( f \), also called CAR (Car Accident Risk), is equal to 1 in case the speed limit is not reached and increases linearly as the speed goes over the speed limit, with a factor of 1.5 in case of bad weather.

The values are then averaged over a window of two seconds, defined as (remembering the 4 Hz sampling rate)

\[
x_i = \frac{1}{9} \sum_{j=-4}^{4} f(t_j).
\]

Finally, instant scores (ranging from 1 to 5) are calculated comparing the single quantities \( x_i \) to the quantiles \( q_1, \ldots, q_6 \) of the distribution of \( x_i \) for all the drivers, similarly to what has been described in §4.1 for the frontal acceleration signal.

### 4.4 Energy efficiency (fuel)

**Signals used:** RPM.

Fuel consumption is a quantity that could not be directly measured from the information contained in the database, as the fuel level signal is too noisy for any type of analysis. Moreover, the fuel consumption – beyond reflecting an eco-friendly driver’s behavior – depends on many external factors, such as the steepness of the road, vehicle occupancy and weather conditions [11], which are hard to quantify.

In this work we assess the fuel consumption primarily basing on the engine’s revolution per minute signal (RPM). This quantity, in turn, is not necessarily tight to a bad driving behavior nor to a non-eco-friendly attitude, for the same reasons explained before. In order to overcome this difficulties, we implement a “swarm” approach, comparing each driver’s performance with the ones of different drivers in a similar situations. This establishes a benchmark that – although not extremely precise – gives a quick and direct measure without relying on individual thresholds. It is important to note that all the cars used in the data collection were of the same model.

More in particular, each road is divided in road segments, each of them delimited by the intersections of the two closest roads. In other words, being \( V \) the sets of all the road intersections, we partition the set of roads upon all points in \( V \). In this way it is ensured that each driver is compared to a driver that drove on the same road segment and for all its length (discarding the drivers who began or ended the driving session in the middle of a road segment), in similar conditions.

Given the RPM signal \( x \), the score in this case is assigned comparing the single quantities \( x_i \) recorded on road segment \( s \) to the quantiles \( q_1, \ldots, q_6 \) of the distribution of \( x_i \) for all the drivers on \( s \), similarly to what has been described in §4.1 for the brake pedal signal.

### 5 SYNTHETIC SCORE

Having computed the instant values for all the DNA dimensions and for each driver, averaged values are obtained by averaging the values of all the sessions, individually for each driver. In conclusion, thus, each driver \( j \) will be fully characterized by his driving DNA \((d_{j,1}, \ldots, d_{j,4})\), being \( d_{j,d} \) the four DNA dimensions described above.

The last step of this works is devoted to provide a synthetic score to each driver, a unique scalar from 1 to 100 that synthesizes the goodness of each person’s driving behavior. This is a function of the values \( d_{j,1}, \ldots, d_{j,4} \), and the type of function can be designed in different ways, for example:

(i) calculating the arithmetic average of the values \( d_{j,1}, \ldots, d_{j,4} \);
(ii) performing a weighted average on the values \( d_{j,1}, \ldots, d_{j,4} \), and set the weights to reflect an assigned importance that each DNA dimension has with respect to the others;
(iii) maximizing the variance of the synthetic scores for all the drivers, performing a Principal Component Analysis (PCA) on the values \( (d_{j,1}, \ldots, d_{j,4}) \) and projecting them on a mono-dimensional space that maximizes the variances of the projections.

In this way, different drivers can be easily compared by means of the synthetic score, in one of its formulations.

We now propose a simple but effective way of visualizing the Driving DNA for each person, through a radar graph showing the values of the four DNA dimensions described above.

Figure 1 shows for each of the 53 drivers their DNA radar graph, whose axis indicate each averaged DNA dimension over all their driving sessions, ranging from 0 (the center of the graph) to 5 (the external edges of the axis). The bigger the area of the radar graph is,

\(^2\)Road speed limits have been retrieved from OpenStreetMap.

\(^3\)Historical weather conditions have been retrieved from Weather Underground.
the better the driver is performing. Underneath each radar graph, three numbers indicate for each driver their synthetic scores (in a scale from 0 to 100), computed with three different formulations.

6 CONCLUSIONS AND FUTURE WORKS
In this paper, the concept of Driving DNA is presented, allowing the characterization of driving behavior through four different easy-to-measure factors, called DNA dimensions. To each driver is then assigned a synthetic score, allowing a unique representations of their driving skills.

The formulations of the four dimensions is the result of a process of analysis of a real database that led to the definition of four factors which are related to driving comfort, accident risk and fuel efficiency. This work, therefore, besides providing a first application and example of the concept of Driving DNA, can be considered as a methodology to reduce a big dataset of CAN bus signals into easy-to-process smaller datasets.

The main difference of the present approach from classic machine learning methods (such as neural networks or other unsupervised or supervised learning techniques) is a very low computational complexity and that the variables definitions and the results keep a physical or phenomenological characterization or interpretation.

This makes it easy to build applications devoted to show and easily compare performances between different drivers. Moreover, this opens the possibility of an almost-real-time visualization of the DNA dimensions to be implemented on-board on the car’s dashboard in order to influence the driver towards a better driving style.

ACKNOWLEDGMENTS
The authors would like to thank Audi ERL, Allianz and all members of MIT Senseable City Lab consortium for supporting this research. The authors would also like to thank Hyemi Song for her assistance with Figure 1.

Privacy disclaimer The data reported herein were 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
Characterizing the “Driver DNA” Through CAN Bus Data Analysis


