DRIVING BEHAVIOR ANALYSIS USING DEEP LEARNING ON GPS DATA

Saurabh Kumar Singh, Utkarsh Anand, Anurag Patel, and Debojit Boro

Department of Computer Science and Engineering, Tezpur University, Napaam, Tezpur, Assam, India-784028
admn@tezu.ernet.in
http://www.tezu.ernet.in/dcompsc/
{csb18055, csb18039, csb18058}@tezu.ac.in, deb0001@tezu.ernet.in

Abstract. Aggressive drivers are often considered to violate traffic rules and adopt dangerous driving behavior. This requires the development of effective and robust classifiers for unsafe drivers. Driving behavior analysis is the classification of driving behavior based on the driver's GPS trajectory. With ever-increasing GPS trajectory data, dangerous driving behavior can be thoroughly analyzed and better classified using a deep learning model. Behavioral analytics can help us analyze and identify dangerous drivers that contribute to traffic safety and promote safe driving behavior. In this paper, we propose a novel feature extraction model using a statistical approach to extract the important features from the GPS trajectory data and label the trajectory. To overcome the dataset dependency, we propose to use a deep learning model on our labeled data and finally classify the safe and unsafe drivers. The proposed method demonstrates high accuracy with reduced computational overhead.

Keywords: Driving Behavior Analysis · Deep Learning · Feature Extraction · GPS.

1 Introduction

Driving behavior analysis has been a prominent problem for the past decade with ever-increasing research and development in this field. Fueled by a vast amount of user GPS trajectory data publicly made available, driving behavior analysis has been getting more attention from researchers. The increased availability of real-time vehicle trajectory data is helping in the development of effective and robust techniques for analysis. The motivation for this study is:

- According to MORTH [1] roughly around 70 percent of all road accidents in India are caused mainly due to reckless and indiscipline driving. The identification of such drivers is very necessary to keep the road safe.
- Driving pattern of a driver affects fuel consumption and emission thus having an impact on the environment. Many automotive insurance companies are adopting pay-as-you-drive or pay-how-you-drive modes for price determination. Prices are decided based on the aggressiveness of the driver.

- Identifying unsafe drivers and taking suitable action on them will help prevent accidents rather than perform post-accident analysis.
- The analysis of human driving behavior is contributing a lot towards the development of autonomous driving as in certain conditions the autonomous algorithm relies solely on human decisions.

The study involves the classification of driving behavior based on GPS trajectories. Driving behavior involves identifying patterns and potentially dangerous behaviors in the GPS trajectory data through feature extraction and behavior evaluation. The main objectives of this study are:

- To extract features like speed, acceleration, and angle for approximating event features.
- To use event feature for labeling of GPS trajectories using statistical approaches.
- To classify and predict drivers associated with unsafe driving behavior by training a machine learning model to map GPS coordinates directly with their classes.

The contributions of our work can be summarized as follows:

- Extraction of basic features like speed, acceleration, angle, and sensitive area to estimate deeper features like SASV (sensitive area speed violation), harsh turning, harsh braking, and harsh acceleration.
- Labeled the raw dataset using weighted CDF (Cumulative Distributive Functions) values of our extracted features.
- Trained a deep learning classifier on our labeled data to classify safe and unsafe drivers. Model achieved high accuracy with reduced computational overhead and overcome dataset dependency.

2 Background and Related Work

The increased real-time trajectory data available today allows learning complex driving patterns from the huge samples, unlike small data samples. Huge data samples have richer representations and high dimensions which present us with statistical and computational challenges. The trajectory data can be collected using navigation applications, most of which have built-in GPS sensor that records the real-time location of the users. GPS sensors can produce data that has a large scale of samples generated over real-time continuous datasets. Though these samples have their advantages, the data still have high noise, and extracting features or identifying patterns can be very challenging in such datasets.

Driver behavior analysis using GPS data in statistical models is quite expensive as the computation of high dimension data is not easy. Authors in [2] suggest that Principal Component Analysis (PCA) will be a successful method for learning features based on GPS data assuming that the input is independent as a Gaussian distribution. Authors in [3] state that Kernel Principal Component

Analysis (KPCA) showed a higher accuracy as compared to PCA. Neural Network is the most widely used approach in solving such problems. Convolutional Neural Network is considered as one of the state-of-the-art methods to model big data in the space domain. In recent studies combining statistical and deep neural networks is a new trend in this field of study. Recently authors in [4] developed a hybrid unsupervised deep learning model to study driving behavior and risk pattern. They approached the problem using an Autoencoder and Self-organized Maps (AESOM), to extract latent features and classify driving behavior. One of the approaches to solve this problem as proposed in [4] is through using a combination of Autoencoder and Self-organized Maps (AESOM) for the extraction of concealed features and classification of driving behavior. Identification of harsh acceleration and deceleration or irregular maintenance of vehicle position, speeding were studied under two cases. They conducted experiments and concluded that through multi-layer autoencoders back propagation is efficient for non-linear and multi-model dimensionality reduction. Whereas for a large GPS dataset the reconstruction errors were minimal.

Previous studies have analyzed driving behavior with smartphone sensors, GPS devices, or video collection through camera/wearable devices. Geo-fencing service by IBM corporation [5] analyzes speed, harsh acceleration, harsh breaking, harsh hard cornering with the help of GPS data of users and provides alerts to the car's dashboard screen. Another leading US insurance company called AllState [6] introduced a solution called Drivewise that rewards the driver for safe driving, which eventually promotes safe driving behavior. AllState uses driver's telematics data to analyze the behavior of the drivers based on which they reward the drivers. They also calculate a score to decide the insurance premium cost. Yu et al. [7] proposes a system that detects abnormal driving behaviors using smartphone. The model uses SVM and Neural Network algorithms to detect the abnormality. Hard cornering and hard braking parameters are taken for behavior detection. TD Insurance [8] a leading company in Canada collects and analyzes driving data, and assigns a driving score for each trip. It uses speed, hard braking, acceleration, and cornering to detect the behavior of the drivers. Though it collects a lot of parameters but uses only speed as the parameter to alert its drivers.Table 1 shows the main features of the existing methods.

Table 1: Existing methods

Approach	Algorithm	Dataset	Accuracy %
Features extraction from redundant driving behaviour data [2]	PCA & Deep Sparse Autoencoder	2-D artificial data sets	NA
Driver's mental fatigue classification, based on EEG [3]	KPCA & SVM	VR-based dynamic simulator	81.64
AESOM for feature extraction [4]	AESOM	Smartphone's GPS sensors (Shenzhen, China)	NA
Using smartphone sensor's data [7]	SVM & NN	Smartphone sensors data of 20 drivers (6 months)	95.36 (SVM), 96.88 (NN)

3 Problem Statement

The driving behavior analysis problem is an identification and classification of safe and unsafe driving patterns of the drivers after observing their GPS trajectories. It is challenging to clean and process the raw dataset with the GPS trajectories into an appropriate set of features to maximize prediction accuracy. The dataset may contain noise, uneven time intervals, or inaccurate GPS locations. Therefore, to classify driving behavior it is necessary to clean the dataset for extraction of relevant features by the proposed model.

3.1 Assumptions

The proposed model has certain limitations as the dataset does not provide complete information about the state of the vehicle environment. To overcome certain limitations we consider the below assumptions:

- The road condition is consistent throughout the trajectory.
- A particular vehicle is driven by only one person.
- The traffic condition is consistent throughout the trajectory.
- The maximum speed and class of the vehicles are the same.
- Population segment used for analysis is homogeneous and majority of drivers are safe drivers.

4 Feature Extraction

Models and approaches as stated in section 2 do not consider many parameters which are important for the precise prediction of driver behavior. Parameters like sensitive area speed violation (SASV) and turning angle on road are some of the most important parameters required for predicting driving behavior. In our method, these parameters are extracted from the GPS data and included for better prediction. The parameters are discussed in detail below.

Acceleration Events like harsh acceleration and harsh braking directly correlate to the acceleration of the vehicle. The effect of acceleration can determine and detect anomalous behavior of the driver.Acceleration can be defined as rate of change of velocity with respect to time. The value of acceleration can be calculated using the formulae given below.

$$\bar{a} = \frac{v - v_0}{t - t_0} = \frac{\Delta v}{\Delta t} \tag{1}$$

where $\bar{a}$ is the average acceleration, v is final velocity, v_0 is initial velocity, Δv is the difference between the final and initial velocity and Δt is elapsed time which can be collected from the dataset.

If g_1, g_2 are two consecutive GPS points then v is the velocity of GPS point g_2 and v_0 is the velocity at GPS point g_1, similarly t is the time stamp of GPS point g_2 and t_0 is the time stamp of GPS point g_1.

Angle Events like harsh turning are dependent on the change in direction with respect to speed. These events can be estimated by evaluating the change in angle of the direction of a vehicle where it is headed through its GPS coordinates. We can approximate the earth as a plane and use two-dimensional vector calculations.

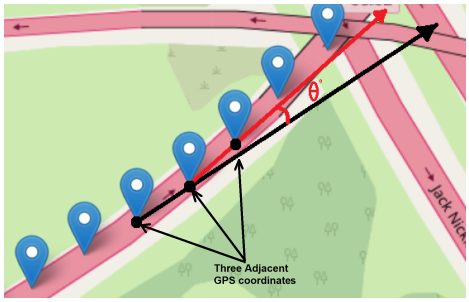


Fig. 1: Angle of direction change of the vehicle

Let g_1, g_2, g_3 and g_2, g_3, g_4 are the three consecutive GPS points and $\vec{u}$ and $\vec{v}$ represents them respectively.

Then the angle between GPS points g_1 and g_2 is calculated using equation no. 2.

$$\theta = \cos^{-1}\left(\frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\|\vec{v}\|}\right) \tag{2}$$

where $\vec{u}$ and $\vec{v}$ are two vectors derived from GPS co-ordinates, $\|\vec{u}\|$ and $\|\vec{v}\|$ are the magnitude of the velocity vectors, and θ is the angle between them as shown in Fig. 1. The equation represents finding angle between two vectors using dot product.

Sensitive Area Speed Violation (SASV) We propose a new way of marking the sensitive area around the vehicle trajectory. We collect and mark all the sensitive locations such as schools, hospitals, and college GPS points from different sources on the Internet. We then take a circular area of 300 meters around these marked locations as shown in Fig. 2. Then we calculate the distance between the vehicle trajectory GPS points and the sensitive marked GPS points. If the distance falls short of less than 300 meters, we mark that vehicle's GPS point as sensitive point and then compare the speed of the vehicle at that GPS point. If the speed is more than the maximum allowed, we give the driver more penalty in points of Speed violation.

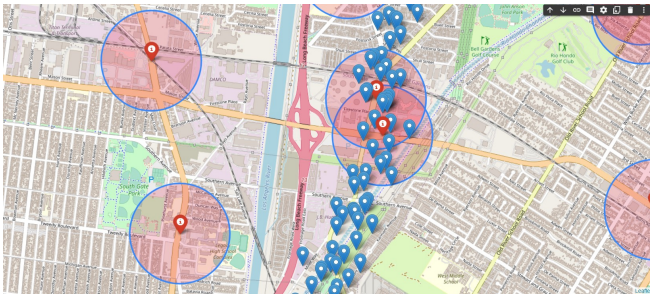


Fig. 2: Enlarged View of Trajectory Cascaded with Sensitive Zones: Red circles are sensitive areas, Blue marking are recorded GPS points of trajectory.

Calculating the distance between vehicle GPS trajectory points and GPS points of sensitive areas is computationally expensive as there is a large number of sensitive areas present in that region. Hence we further modify our approach. Based on the vehicle trajectory, we divide a region into many zones (as shown in Fig. 3). For each GPS point in different zones, the distance is calculated only from those sensitive marked areas which fall under that particular zone. All other sensitive areas falling in other zones to which the the GPS point does not belong are discarded. This approach reduces the computation time significantly.

We mark all the GPS points which are within the range of 300 meters as 'sensitive area' which indicate that the point is inside one or more sensitive areas. Else, we mark it as 'normal area' to indicate the GPS point is not in the sensitive area. After marking the GPS points we calculate the SASV feature based on speed limit of the specified area. The SASV feature is calculated by counting the occurrence of speed violation committed by the driver per kilometer.

Fig. 3: Zonal Division of Vehicle Trajectory.

5 Proposed Method

In this section, we propose a novel method to classify driving behavior by dividing the problem into two smaller problems. The first problem is to label the GPS trajectory data through feature extraction. To solve this problem we use statistical approaches on extracted features like Harsh Turn, Harsh Braking, and Harsh Acceleration to approximate a score based on the weighted sum of their CDF values. We then label the data with the help of these generated scores. The second problem is to use these labeled data to generate a model which can map the GPS trajectories directly to their classes. For this we a use publicly available state-of-the-art time-series classification model for the prediction of class labels based on only three features viz. latitude, longitude, and speed.

5.1 Feature Extraction Model

We process the raw unlabeled dataset to remove noise and garbage values, then we extract features like acceleration, angle, and sensitive area explained in section 4. The extracted features are then used to estimate the event type such as harsh acceleration, harsh braking, and harsh turning as shown in Fig. 4. Harsh acceleration and harsh braking is calculated using the acceleration feature, any sudden increase in acceleration is considered as harsh acceleration and high retardation is considered as harsh braking. Harsh Turning is calculated using the angle and speed feature, high speed and high angle change is marked as harsh turning. We calculate the score feature by estimating CDF[9] of event features.

Score Feature To calculate the score of the driving behavior for each driver, we find its weighted sum of the CDF(area under density function of normalized event per kilometer) values of all the event features using equation 3.

$$\text{Score}_{s \in S}(x) = \sum_{event} weight_{event} P_{event}(X \leq x) \tag{3}$$

where x is events per kilometer, $event$ is Event Type, $weight_{event}$ is weight of event, P_{event} is CDF value, and S is population segment. The Score feature specifies how safe a driver is compared to the general population. The higher the score value it is more likely that the driver has unsafe driving behavior. Weighted sum over CDF values on event features gave us a score to label the trajectories.

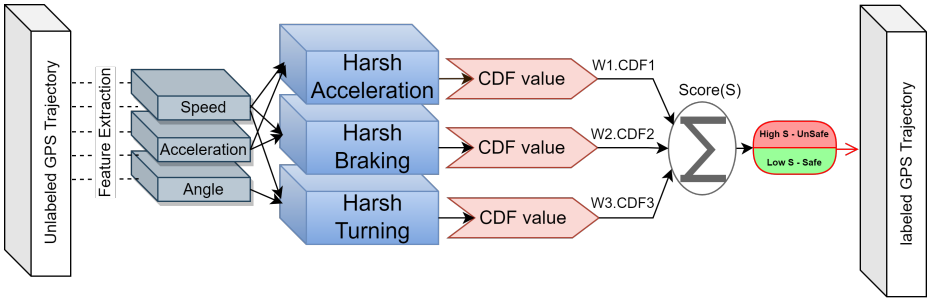


Fig. 4: Feature Extraction Model

5.2 Classification Model

Event feature is not present in every dataset and large scale calculation of these events are computationally expensive. To classify any raw GPS data which does not contains event features we trained a deep learning model MINIROCKET[10] on our labeled dataset to predict class labels. MINIROCKET classifier is trained with only three features latitude, longitude, and speed. The trained model eliminates the need for extraction of features and delivers similar performance when compared to statistical labeling on the test data.

6 Experiment Results

In our experiment, we use a web service Google Colab, the machine configuration comes with 12 GB of RAM and 12GB NVIDIA Tesla K80 GPU. We use DACT (Dataset of Annotated Car Trajectories)[11] that contains trajectory data of 50 driver trips and a total of around 48000 GPS points with a sampling rate of 1 second. Table 2 shows the sample description of the trajectories with their various features.

Table 2: Dataset Sample

TripID	TimeStamp	Speed	Acceleration	Heading	HeadingChange	Latitude	Longitude	Annotation	SegmentType
T-1	Wed Jul 24 15:58:22 EDT 2013	29	0	84	0	39.980572	-82.953895	NULL	NULL
T-1	Wed Jul 24 15:58:23 EDT 2013	29	-0.28	84	0	39.980575	-82.953735	NULL	NULL
T-1	Wed Jul 24 15:58:24 EDT 2013	24	-1.94	84	0	39.980582	-82.953587	NULL	NULL
T-1	Wed Jul 24 15:58:25 EDT 2013	21	-1.39	84	0	39.980602	-82.953477	NULL	NULL
T-1	Wed Jul 24 15:58:26 EDT 2013	20	-0.56	90	6	39.980603	-82.953347	NULL	NULL
T-1	Wed Jul 24 15:58:27 EDT 2013	19	-0.28	90	0	39.9806	-82.95325	NULL	NULL
T-1	Wed Jul 24 15:58:28 EDT 2013	19	-0.28	90	0	39.9806	-82.95313	Segment	Slow-Down

We analyse the driving behavior on two different parameters *Rule violation* and *Driving style*. To estimate the *Rule violation* we only use the normalized value of SASV feature as shown in Table 3.

Table 3: Rule Violation

Driver	Distance(Km)	Occurance of Violations	SASV	Rule Violation
T-1	15.8364	85	5.3674	0.1959842
T-10	24.0076	470	19.577	0.7148382
T-11	12.5815	302	24.003	0.8764621
T-12	22.1294	326	14.732	0.5379062
T-13	15.9831	400	25.026	0.9138135
T-14	22.1398	432	19.512	0.7124734
T-15	20.7581	533	25.677	0.9375582
T-16	14.8016	225	15.201	0.5550505
T-17	20.7034	567	27.387	1
T-18	21.6624	324	14.957	0.5461313

For labeling the *Driving style*, the dataset is processed with the help of the feature extraction model which combines CDF value of all event features for each driver and gives us a score relative to the general population which represents driving behavior. The CDF values and scores of few drivers are shown in Table 4. Higher scores are unsafe drivers while lower scores are safe drivers.

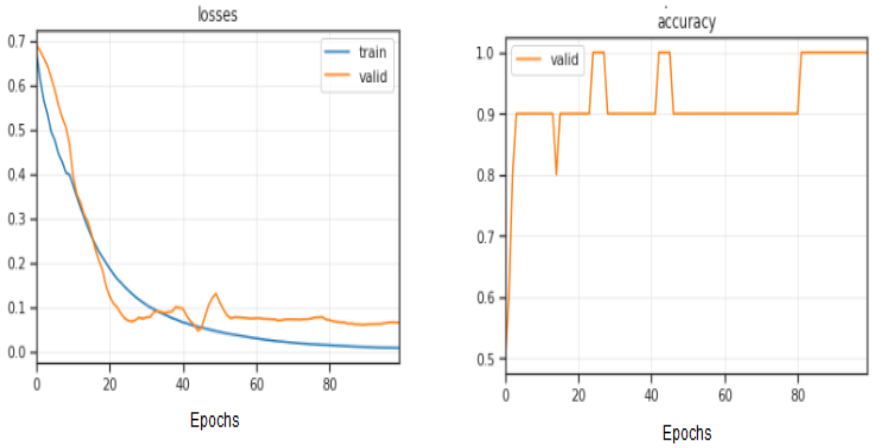
Table 4: Drivers with their Score and Label

Driver	**Calculated CDF values**			**Score(S)**	**LABEL**
	Harsh Acceleration	**Harsh Braking**	**Harsh Turning**		
T-18	0.9483	0.8541	0.3952	2.1977	UNSAFE
T-1	0.6436	0.7281	0.846	2.2177	UNSAFE
T-3	0.709	0.7607	0.78	2.2497	UNSAFE
T-11	0.7729	0.6691	0.8291	2.2711	UNSAFE
T-39	0.6474	0.7845	0.8663	2.2981	UNSAFE
T-44	0.8038	0.8079	0.8311	2.4428	UNSAFE
T-32	0.85	0.7974	0.9121	2.5595	UNSAFE
T-5	0.9471	0.8783	0.7869	2.6123	UNSAFE
T-16	0.932	0.8797	0.9313	2.743	UNSAFE
T-35	0.9333	0.8743	0.9372	2.7447	UNSAFE
T-43	0.0228	0.0974	0.3636	0.4837	SAFE
T-36	0.1922	0.111	0.181	0.4842	SAFE
T-37	0.1241	0.1742	0.3138	0.612	SAFE

Using the score feature value we assign a class label to the driver trajectory. We label higher scores as unsafe while lower scores as safe. We feed our labeled data with only three features latitude, longitude, and speed to MINIROCKET classifier and train for 100 epochs. We achieved 100 percent accuracy as shown in Fig. 5.

6.1 Discussion

The event feature of the dataset plays a very major role in the classification. We calculate the score feature for *Driving style* estimation using statistical methods on harsh events. We also produced a feature, "Sensitive Area" which represents whether or not a GPS point is present within a certain range of schools and



(a) Epoch vs Loss (b) Epoch vs Accuracy

Fig. 5: MINIROCKET training loss and test accuracy at 100 epochs

hospitals. We calculate a novel SASV feature based on the speed violation to estimate *Rule violation*. The *Driving style* and *Rule violation* play a very important role in determining driving behavior and contribute to road safety. We propose a model that classify *Driving style* using only the GPS coordinates and speed of vehicle. The model can be used on any raw GPS data without calculating event features reducing computational expenses and complexity.

7 Conclusion

Availability of labeled event feature datasets is very limited. The MINIROCKET classifier gave accurate results after training over our labeled dataset, Generalizing our model to any other dataset which may not have event feature, the proposed model can directly predict *Driving style* class labels, eliminating the problem of dataset dependency. We also calculate driving behavior in terms of *Rule violation*, which estimates how many times a driver violates road safety rules. In this work we only considered speed violations based on trajectory information, however there are other factors like lane change, vehicle overtaking etc. In future we look forward to work on these features and introduce new parameters for driving behavior analysis.

References

1. Ministry of Road Transport and Highways (MORTH) (2019). Road accidents in india.
https://morth.nic.in/road-accident-in-india (last accessed 28-01-2022).

2. HaiLong Liu, Tadahiro Taniguchi, Yusuke Tanaka, Kazuhito Takenaka, and Takashi Bando (2015). Essential feature extraction of driving behavior using a deep learning method. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 1054–1060.
3. Chunlin Zhao, Chongxun Zheng, Min Zhao, and Jianping Liu (2010). Classifying driving mental fatigue based on multivariate autoregressive models and kernel learning algorithms. In *2010 3rd International Conference on Biomedical Engineering and Informatics*, volume 6, pages 2330–2334.
4. Jingqiu Guo, Yangzexi Liu, Lanfang Zhang, and Yibing Wang (2018). Driving behaviour style study with a hybrid deep learning framework based on gps data. *Sustainability*, 10(7).
5. IBM (2019) Analytics :. real-world use of big data in insurance. https://www.ibm.com/services/us/gbs/thoughtleadership/big-data-insurance/.
6. Allstate. Stay smart on the road. https://www.allstate.com/drive-wise.aspx.
7. Jiadi Yu, Zhongyang Chen, Yanmin Zhu, Yingying Chen, Linghe Kong, and Minglu Li (2017). Fine-grained abnormal driving behaviors detection and identification with smartphones. *IEEE Transactions on Mobile Computing*, 16(8):2198–2212.
8. TDInsurance. TD MyAdvantage (2019). Myadvantage, https://www.tdinsurance.com/products-services/auto-car-insurance/my-advantage.
9. O Mendels, G Bertental, and T Kamara (2018). Unsupervised driver safety estimation at scale, a collaboration with pointer telocation.
10. Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb (2020). MINIROCKET: A very fast (almost) deterministic transform for time series classification. *CoRR*, abs/2012.08791.
11. Sobhan Moosavi (2017). DACT: Dataset of Annotated Car Trajectories.