Accident Analysis and Prevention 84 (2015) 54–64

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Driving risk assessment using near-crash database through data mining of tree-based model

Jianqiang Wang[a], Yang Zheng[a], Xiaofei Li[a], Chenfei Yu[a], Kenji Kodaka[b], Keqiang Li[a,*]

[a] State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 10084, China
[b] Honda R&D Co. Ltd., Automobile R&D Center, Tochigi 321-3393, Japan

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 11 May 2015
Accepted 3 July 2015
Available online 27 August 2015

Keywords:
Naturalistic driving study
Driving risk
Near-crash
Classification and regression tree (CART)
K-mean cluster

ABSTRACT

This paper considers a comprehensive naturalistic driving experiment to collect driving data under potential threats on actual Chinese roads. Using acquired real-world naturalistic driving data, a near-crash database is built, which contains vehicle status, potential crash objects, driving environment and road types, weather condition, and driver information and actions. The aims of this study are summarized into two aspects: (1) to cluster different driving-risk levels involved in near-crashes, and (2) to unveil the factors that greatly influence the driving-risk level. A novel method to quantify the driving-risk level of a near-crash scenario is proposed by clustering the braking process characteristics, namely maximum deceleration, average deceleration, and percentage reduction in vehicle kinetic energy. A classification and regression tree (CART) is employed to unveil the relationship among driving risk, driver/vehicle characteristics, and road environment. The results indicate that the velocity when braking, triggering factors, potential object type, and potential crash type exerted the greatest influence on the driving-risk levels in near-crashes.

1. Introduction

1.1. Background

In the past two decades, significant progress has been made in all aspects of vehicle safety systems, and experts from both academia and industry have conducted extensive research on vehicle safety (Young et al., 2014; Sepulcre et al., 2013; Takeda et al., 2011; Zheng et al., 2014). Efforts that aim to advance vehicle safety systems can mainly be divided into two areas (Jarašūniene and Jakubauskas, 2007): (1) active safety, which aims to avoid accidents and (2) passive safety, which helps reduce injuries in an accident. The active safety approach forecasts future driving states based on vehicle dynamics, infrastructure, and driver awareness (Wang et al., 2015), whereas the passive safety approach mainly focuses on enhancing vehicular safety systems such as seat belts, airbags and strong body structures (Jarašūniene and Jakubauskas, 2007). Although many encouraging achievements have been made, the number of road fatalities still remains unacceptably high, and traffic accidents are considered a major public health problem (DTM-China, 2010).

As the responsibility for traffic accidents involves the vehicles, drivers, and roadways, we must not only improve the safety performance of vehicles, but also better understand the factors that influence driving risk and identify the factors that result in accidents to make road transportation much safer. Many studies have attempted to better understand the factors that affect the probability and injury severity of crashes (Lord and Mannering, 2010). From a methodological standpoint, logit-based models are some of the most practical tools used for analyzing accident severity (Chen et al., 2012; Al-Ghamdi, 2002). Recently, non-parametric methods and data-mining techniques have been widely used to identify the factors associated with accident severity (Chang and Chen, 2005; Chang and Wang, 2006; Montella et al., 2011, 2012; Li et al., 2008; Harb et al., 2009). For example, Chang and Chen (2005) and Chang and Wang (2006) proposed a classification and regression tree (CART) model to establish the relationship among injury severity, driver/vehicle characteristics, and accident variables, indicating that vehicle type is a very important variable associated with cash severity. Li et al. (2008) evaluated the application of a support vector machine (SVM) model for predicting motor vehicle crashes, and showed that SVM models performed better than traditional negative binomial models. Montella et al. (2012) employed a decision tree and association rules to analyze accidents involving powered two-wheelers, and demonstrated that the curve alignment, rural areas, run-off-the-road crashes, night time, and rainy weather were

* Corresponding author.
E-mail address: likq@tsinghua.edu.cn (K. Li).

http://dx.doi.org/10.1016/j.aap.2015.07.007
0001-4575/

significantly associated with accident severity. These studies provided some insights into the factors that affect the likelihood of a vehicle accident. However, they were typically based on official traffic accident statistics, which have two major limitations: (1) lack of detailed driving data, and (2) difficult to collect and acquire (usually collected by traffic police agencies). Hence, the aforementioned studies usually do not consider the relationship between the accident severity and detailed driving data (*e.g.*, vehicle speed, acceleration, braking, and steering information).

Recent developments in vehicle instrumentation techniques have made monitoring naturalistic driving behavior and obtaining detailed driving data both technologically possible and economically feasible. For instance, NHTSA sponsored the project "100-Car Naturalistic Driving Study" which is a large-scale instrument-vehicle study to collect naturalistic driving data in the United States (Dingus et al., 2006). A series of technology tests of safety equipment was conducted in Michigan using the naturalistic driving technique (UMTRI and GMRDC, 2005). Takeda et al. (2011) reported a comprehensive project involving collecting large amounts of driving data on the actual road to study driver behavior and accident-causation-mechanism. With access to naturalistic driving data, traffic safety-related events could be observed and measured more precisely (Wu et al., 2014). Meanwhile, many researchers have proposed new methods and gained new insights into traffic safety (*e.g.*, Malta et al., 2009; Aoude et al., 2012; Guo et al., 2010; Jovanis et al., 2011; Jonasson and Rootzén, 2014). For instance, Malta et al. (2009) proposed a method to improve the understanding of driver behavior under potential threats using a large real-world driving database. Guo et al. (2010) assessed the factors associated with individual driver risk using naturalistic driving data. For naturalistic driving data, crash surrogates have received extensive research attention (see Guo et al., 2010; Wu and Jovanis, 2012, 2013; Moreno and García, 2013, for examples), because the number of crashes observed with naturalistic driving is typically small. Near-crash is frequently used as a surrogate measure for assessing the safety impact. For instance, Guo et al. (2010) employed two metrics, namely, precision and bias of risk estimation, to assess near-crashes, and indicated that using near-crashes as a crash surrogate could provide definite benefit when data about a sufficient number of crashes are not available. Recently, Wu and Jovanis (2013) proposed a multi-stage modeling framework to search through naturalistic driving data and extract near-crash events. All of these studies have demonstrated that naturalistic driving data could provide more controllable laboratory data as a useful supplement for traffic safety studies, and has the potential to further our understanding of crash causality, as well as improve road safety. Naturalistic driving data could not only provide more detailed driving exposure data, but also present the probability to identify more plausibly risky driving events and the associated factors.

1.2. Preview of the key results

This study focuses on the analysis of factors that influence driving risk using a naturalistic driving database. This database was obtained through designing a novel transcription protocol to code naturalistic driving data, which have two distinguishing features: (1) drivers drive in their normal states and (2) the instruments installed in vehicles can record drivers and road environments continuously during driving (Jovanis et al., 2011). The naturalistic database used herein contains only near-crash events because no actual crashes happened during the naturalistic experiments conducted on actual Chinese roads. Near-crashes refer to cases where drivers execute rapid evasive maneuvers (*i.e.*, emergency braking and/or steering operation) when facing a potential driving risk or a potential threat; in the absence of such an action, a real crash may occur. In the experiments, near-crash events in naturalistic driving were identified by detecting unusual vehicle kinematics using accelerometers and gyroscopic sensors installed in the experimental vehicle (Wu and Jovanis, 2013; Wu et al., 2014).

Recently, a few studies have focused on the assessment of risk in the driving environment, for example, individual driving risk (Guo and Fang, 2013) and momentary risk perception of a driving situation (Lu et al., 2012; Charlton et al., 2014). These studies employed indicators such as driver attributes and vehicle kinetic parameters to represent the risk level. Besides, critical braking and speed profiles were proposed to characterize the near-crashes in Moreno and García (2013) and Bagdadi (2013). In the present paper, we propose a novel method to quantify the driving-risk involved in a near-crash event. First, the driving-risk level is represented by the braking process characteristics, namely (1) maximum deceleration, (2) average deceleration, and (3) percentage reduction in vehicle kinetic energy. Then, the K-means cluster method is employed to classify near-crashes into different-risk levels based on the three aforementioned braking process features. Then, CART is employed for exploring the relationship among driving risk, driver/vehicle characteristics, and road environments. Identifying the factors associated with driving risk and further predicting high-risk driving scenarios will enable the adoption of proper safety countermeasures to reduce probable hazardous situations for high-risk groups, and thus improve overall driving comfort and safety. By analyzing driver characteristics, road conditions, and vehicle characteristics using the near-crash database, we obtained new insights into driving risk. The results indicate that the velocity when braking (V_BRA), triggering factors (T_FAC), potential object type (O_TYP), and potential crash type (P_CRA) had the greatest influence on the driving-risk level involved in near-crashes. These results can improve our understanding of the factors that affect driving risk, and help create polices and countermeasures to improve driving safety and comfort.

The remainder of this paper is organized as follows: Section 2 describes the near-crash database and presents some preparations, including experiment design, labeling protocol and driving-risk definition. The methodology employed in this study is presented in Section 3. Section 4 discusses the results, and some concluding remarks are given in Section 5.

2. Database and preparation

To build a firm foundation for the assessment of driving risk and enhancing driving safety, two components are essential: (1) real-driving data and (2) careful experimental design. Data collection is performed using naturalistic and low-intervention methods under actual traffic conditions. This section introduces the experimental equipment and experiment design, describes the near-crash database, and presents the definition and cluster analysis of driving risk.

2.1. Data-collection equipment and experiment design

2.1.1. Data-collection equipment

The naturalistic driving experiments were conducted using a Honda Crosstour, which was provided by Honda. The vehicle was equipped with instruments to collect driver, vehicular, and road data under real-world conditions. The data-collection system installed in the experimental vehicle included two driving recorders (DR) and four cameras (Fig. 1). The four cameras were used to record detailed video scenes including (1) forward view, (2) right-side forward view, (3) left-side forward view, and (4) driver's facial expression. One DR recorded data obtained by sensors, including GPS, brake signal, steering signal, three-axis