ABSTRACT

The design of a safe transportation system requires numerous design decisions that should be based on data acquired by rigorous scientific method. Naturalistic data collection and analysis methods are a relatively new addition to the engineer's toolbox. The naturalistic method is based on unobtrusively monitoring driver and vehicle performance under normal, everyday, driving conditions; generally for extended collection periods. The method generates a wealth of data that is particularly well-suited for identifying the underlying causes of safety deficiencies. Furthermore, the method also provides robust data for the design and evaluation of safety enhancement systems through field studies. Recently the instrumentation required to do this type of study has become much more cost effective allowing larger numbers of vehicles to be instrumented at a fraction of the cost.

This paper will first provide an overview of the naturalistic method including comparisons to other available methods. The focus of the paper then shifts to review the evolution of the data acquisition systems (DAS) and methods that have enabled naturalistic data collection. The goal is to provide readers with an understanding of how technology and unique partnerships has allowed the naturalistic data collection method to mature.

INTRODUCTION

Although driving becomes nearly second nature to the seasoned individual, in reality it is the most complex task most people perform on a daily basis. Drivers execute a continuous stream of decisions to navigate and control their vehicle based on sensory inputs provided by their surroundings. This interaction between the driver, their vehicle, roadway, adjacent vehicles, and the environment results in an incessant flow of potential hazards. Although drivers typically recognize and avoid these hazards, crashes are unacceptably prevalent; evidenced by the approximately 2.4 million injuries and 37 thousand fatalities reported each year [1]. The research community is focused on improving driver safety through investigation of the root causes of crashes and the development of strategies to reduce crash prevalence and mitigate the resulting injuries.

Naturalistic data collection has emerged over the last decade as a highly effective technique for gathering the information necessary to improve driver safety. The technique is based on unobtrusively monitoring drivers under normal, everyday, driving conditions. Typically, a driver's personal vehicle is instrumented with a data acquisition system (DAS) and no special instructions are provided. Drivers simply go about their daily routine while the DAS silently collects information about the driver's interaction with the vehicle, roadway, adjacent vehicles, and environment. The collected information is subsequently analyzed to identify the root causes of crashes and near crashes as well as a number of other uses such as virtual testing of countermeasures.

Although instrumented vehicle studies and field operational tests on live roadways have been conducted for many years, the first known application of naturalistic data collection methods in surface transportation was the “100 Car study” performed by Dingus et al., in 2002 [2]. This study collected 42,000 hours of continuous video and electronic data for 241 drivers over a 13 month period. This study was the first instrumented vehicle study to: 1) capture a meaningful number of crash, near crash and minor collision events, and 2) it was the first study that provided the opportunity to “mine data” to answer a large number of safety-related questions related to driver behavior and performance.

Naturalistic data collection relies on technology to replace the in-vehicle experimenter. Although this was initially difficult
and expensive, particularly for larger-scale studies, modern digital embedded technologies are revolutionizing the method by providing higher performance data collection at a fraction of the cost.

Today, naturalistic data collection is a mainstream technique with ongoing data collection efforts spanning the globe [3]. Within the United States, the Academy of Sciences in conjunction with the Transportation Research Board is embarking on the largest naturalistic data collection effort to date in the Strategic Highway Research Program 2 (SHARP2) Naturalistic Driving Field Study (SHARP2 NDFS, a.k.a “2000 car study”) [4]. This ambitious project relies on partnerships with research organizations such as the Virginia Tech Transportation Institute who is developing the data acquisition systems and coordinating the data collection effort. Other partners will manage individual data collection sites located from Seattle to North Carolina. Functioning as a team, this group is creating a national laboratory housing an unparalleled volume of driving performance data. Analysis of this data will allow researchers to perform a virtually limitless number of investigations to understand crash causation and driver behavior as it occurs on live roads with all the physical and cognitive factors present (i.e. Naturalistic).

In additional to pure naturalistic study, modern field studies often employ a pseudo-naturalistic strategy for collecting data. These field studies employ the naturalistic method to an extent; with the exception of a treatment which is being assessed and possibly additional controls such as asking a participant to drive in a particular area. For example, a pilot field study was recently performed for a cooperative intersection collision avoidance system for violations (CICAS-V) [5]. During this study participants were asked to drive a vehicle that was equipped with intersection collision avoidance technologies; however, there was no experimenter present in the vehicle. Although the driver's interaction with the technology was the subject of interest, the basic techniques employed during the data collection mirror that of a true naturalistic study.

As the prevalence of naturalistic data collection and the availability of naturalistic databases continue to grow, it is important for researchers and practitioners to be aware of naturalistic methods. The purpose of this paper is to provide an overview of naturalistic data collection emphasizing the technology advancements that have brought the collection method into mainstream science.

**DATA COLLECTION METHODS COMPARED**

Naturalistic data collection lies in the middle of the continuum of data collection methods (Figure 1). At one end of the spectrum is epidemiological data collection which is based on mining crash databases such as the Fatality Accident Reporting System (FARS) and the General Estimates System (GES). At the other end of the spectrum are empirical data collection methods such as simulator and test track studies. These three complementary data collection methods each have advantages; researchers should be careful to select the appropriate method(s).

Epidemiological data collection relies on actuarial data which provides very precise knowledge about crash risk. This precision is a direct result of measuring actual crashes after they have occurred rather than relying on a crash surrogates. Unfortunately, epidemiological data is reactive since detection of a safety deficiency is not possible until after a statistically significant number of crashes have occurred; potentially years after the deficiency manifests. Furthermore, crash databases typically have limited pre-crash information.

Most databases rely on data gathered from scene investigations (typically police reports) and/or self-report. The primary role of a policeman arriving at a scene is not filling out a police report. Rather, the police ensure public safety by providing traffic control and clearing the scene. The scene has likely changed and vehicles are frequently moved. It is difficult to obtain accurate assessments of pre-crash circumstances as drivers and passengers may be deceased, severely injured, or likely dazed. Furthermore, pre-crash events frequently occur so rapidly that key elements are forgotten or left out by witnesses. Finally, drivers and passengers may withhold information to avoid prosecution or embarrassment. These factors lead to errors in the database which can lead to false conclusions.

At the other side of the spectrum lies empirical data collection. Empirical data collection allows researchers to tightly control the research environment. They can hold a number of potentially confounding factors static while an independent measure is manipulated. This provides ordinal
crash risk information for the factors that were manipulated. It also allows proactive investigation of the treatment options such that risk can be assessed before the treatment is deployed into the real-world.

When performing an empirical study one must select a set of safety surrogate measures. Example measures may include lane keeping performance and average glance time. These surrogates are not a direct measure of crash risk which limits the ability to make conclusions regarding the safety impact in the population. Furthermore, the experimental scenario can alter the driver's behavior. For example, the driver in a simulator or test-track study will not have the same underlying motivation for driving and may purposefully modify their behavior to meet the perceived expectations of the experimenters.

Nestled in the middle, naturalistic data collection has the advantage of measuring ‘natural' behaviors in the full driving context. With synchronized video, the information density obtained far exceeds that of any other data collection method. The time-history data provides the ability to observe numerous pre-crash factors that cannot be captured in a crash database. The resolution of collected data also permits thorough examination of the seconds leading up to a crash such that detailed reconstructions are possible. This allows researchers to fully characterize factors leading to a crash, including those that were not identified a priori.

The collection of naturalistic data can be somewhat more complex than empirical data collection. The DAS must be able to operate for an extended duration with minimal maintenance and participant logistics must be considered throughout the study duration. When performing analysis, the researcher must be careful to account for potential confounds that could be been controlled in an empirical study. Finally, large naturalistic studies have traditionally been costly. This limitation is alleviated by the latest generation of low-cost embedded DAS coupled with improved collection methods and standardized procedures.

It should be noted that many efforts will not fit cleanly into one category. Field studies are the most common example in which most of the naturalistic methods are used with the exception of a treatment that is introduced. Field studies are not purely naturalistic, but rather blend the empirical and naturalistic methods. Such data collection efforts provide excellent data for determining the merits of a safety improvement strategy with all the complexities of the open-roadway prior to deployment.

**THE EVOLUTION OF NATURALISTIC DATA COLLECTION EQUIPMENT**

Naturalistic data collection relies on technology to enable unobtrusive monitoring of the driver and environment. Only basic naturalistic studies are possible without technology (e.g. counting seatbelt usage from outside the vehicle). In the 80s, it became possible to instrument vehicles using digital computers and video cameras (Figure 1). However, these vehicles were very expensive, time consuming to instrument, and they were far from unobtrusive. The figure is from the first instrumented vehicle evaluation of a moving map navigation system in 1984 [6]. The instrumented vehicle in this study employed two full-size video cameras mounted outside of the vehicle which captured the drivers face and the forward scene. A standard IBM PC and custom analog to digital conversion box was used to collect a variety of performance data including steering, accelerator and brake position.

In the late 90's technologies for collecting naturalistic data were refined. An example study is the naturalistic lane change study which provided participants with a specially equipped vehicle to drive as their personal vehicle for an two week duration [7]. The technologies put into these vehicles reduced the camera size significantly by using remote video cassette recorders (VCR) to capture the images. For the first time, video collection was choreographed by a computer that automatically started when the ignition was switched on and continuously collected data until the vehicle was shut off. Furthermore, this video was also loosely synchronized with a digital recording of sensor data including measures such as turn signal activation and lateral acceleration. This data collection system was considered very compact for the time despite filling a significant volume of the vehicle trunk (Figure 3).

Similar hardware configurations employing a combination of VCRs, laptops, and sensor arrays were used for a number of years. One of the most advanced systems of this early generation DAS was constructed for a study on drivers of long-haul trucks [8]. These trucks had to be capable of collecting data for approximately one week of data during cross-country trips without any maintenance. To enable a data collection of this duration, first data was only retained for an interval of time surrounding a triggered event (e.g.}

![Figure 1. Image of the first vehicle involved in an early data collection project](image-url)
longitudinal acceleration greater than a set threshold). To capture the video, five VCRs had to be choreographed to provide sufficient storage space.

A few years after the turn of the century, efficient digital video recording methods were surfacing and the cost of hard drive storage was plummeting. The hardened PC104 computer platform, originally developed by the space shuttle program, was also readily available. Motivated by the upcoming 100 Car Naturalistic study [2], engineers at VTTI took advantage of these technology advancements and designed a completely new DAS. This highly configurable DAS took advantage of controller area networks (CAN) allowing direct interfacing to the vehicle network as well as an expanded sensor array. The DAS had numerous additional enhancements such as cellular communications, a machine vision lane tracker, and a battery backup allowing for a complete data record; even in the event of power failure during a crash. This DAS was capable of collecting several months of continuous video and data without intervention yet took up about one quarter of the footprint of its predecessor (Figure 4).

Derivatives of the 100 car DAS were used for over six years for numerous projects. Enhancements such as developing a customized real-time Linux operating system, installing faster processors, expanding the sensor array, obtaining larger storage, and migrating to more efficient video compression techniques were incrementally performed. Although this DAS repeatedly demonstrated its utility, technology advancements were enabling the next leap in data acquisition.

Over the last couple of years, engineers at VTTI have been migrating to embedded technologies for our data acquisition needs. Moving to an embedded platform allows the engineers to develop a complete solution precisely targeted at the specific needs of vehicular data collection. At the core of this new DAS, known as the NextGen is a field programmable gate array (FPGA) which choreographs the real-time data collection. The FPGA lies on a custom board alongside two digital signal processors (DSP). The DSPs multiplex up to 10 video streams and house machine vision programs for face and lane tracking. The system gathers and records data streams asynchronously allowing each sensor to operate at its optimal collection rate. The data collection package is considerably more accurate than its predecessors yet is only the size of a large textbook (Figure 5).

The NextGen records data at millisecond accuracy and is attached to a high performance sensor suite. The base configuration provides the following measures; however, a virtually limitless array of sensors can be added or removed to meet the needs of a specific study.

- 5+ Digital video cameras/views
  - Machine Vision Eyes Forward
  - Machine Vision Lane Tracking
- Accelerometers (3 axis)
- Gyroscopes (3 axis)
- GPS
- Forward radar
- Luminance sensor
- Passive alcohol sensor
- Incident push button
- Audio
- Turn signals
- Vehicle network data from OBD
- Vehicle network data from private CAN
- Cell phone, Wi-Fi, Bluetooth, Ethernet, USB, CAN
- Expandable
One of the key motivators of the NextGen system is the economies of scale that can be achieved with an embedded platform. Studies including thousands of vehicles, such as those in the SHARP2 NDS, would not be feasible without the significant cost reduction of embedded platforms. To take economy a step further, VTTI has also invested a second embedded DAS platform. The miniDAS is currently in development and is intended for a slightly different application than its bigger brother.

The miniDAS is a single highly compact unit that includes the DAS and all sensors. The unit, which is not much larger than a deck of cards (Figure 6), is designed to be located on the windshield or dashboard of any modern passenger vehicle. Installation is as simple as mounting and plugging into the OBDII port. The miniDAS provides similar functionality to the NextGen, although sacrifices in available processing power, expandability, and camera configurations are required. The primary attraction of this unit is its compact size, ease of installation, and a dollar to performance ratio that pushes the technology envelope.

THE EVOLUTION OF DATA STORAGE

Alongside advancements in the DAS, data storage technology and strategies have also transformed. The storage model was relatively simple when the number of DASs in the field was small and video was stored on cassettes. Typical data storage requirements for an entire study were measured in few gigabytes of server space for ASCII files and a stack of cassettes for video. The magnitude of data produced by NHTSA and VTTI during the 100 car project required shifting to a dedicated data storage system.

At the time it was collected, the 100 car project was the largest naturalistic data collection in terms of the number of drivers, duration of collection, and digital space required per unit of time. The increase in storage requirements was primarily due to the inclusion of digital video which typically represents 80%-95% of overall storage requirements. A network attached storage system was installed to house the approximately 2 terabytes of data; which was indeed a significant amount of storage at the time.
After the 100 car study a number of additional naturalistic studies were performed through partnerships with organizations such as the Federal Motor Carrier Administration [9], the National Highway Safety Administration, the National Institutes of Health (NIH), and the Crash Avoidance Metrics Partnership [5]. This rapid database expansion increased storage demands to over 100 terabytes and highlighted the need to improve the compression and access speed performance. These studies fueled a transition to dedicated standard (video) and high-speed (parametric data) storage server racks attached to database management systems based on the structured query language (SQL). An independent local area network (LAN) protects from outside intruders while the array of independent storage disks (RAID) and routine offsite tape backup protect data from hardware failures. Although a 16 processor windows computing cluster is used for the analysis of large jobs, most analysis is performed on local machines.

The SHARP2 NDFS has motivated a complete restructuring of the data storage and computing framework. The VTTI and Virginia Tech partnered in the creation of a parallel high-performance computational (HPC) system. This system is currently being implemented and should be running within the next few months. The HPC includes three main components:

- Network file servers: total of 1PB file system storage, notably for hosting and serving video files
- Parallel computational cluster
- Parallel database cluster/data warehouse

The HPC is being designed for expansion such that it will scale effectively as future data collection efforts are undertaken. In addition, the HPC is also designed for efficient data analysis through parallel processing and rapid data access.

THE EVOLUTION OF DATA ANALYSIS

The early naturalistic data collection systems recorded in a standard ASCII format that could be read into any analysis software. A frame number, overlaid onto the video during recording, was used to manually synchronize video to the parametric data. To view the data, analyst simply looked at charts produced by a commercial analysis software packages (e.g. Excel and SAS) and used the frame-by-frame feature of a VCR to view the video. Data annotations were recorded in a spreadsheet and generally included broad observations such as incident type, weather conditions, and roadway type. This process required the analyst to search for the events of interest in the video and was relatively cumbersome.

With the migration to digital video, it became feasible to automatically synchronize the video and parametric data within a single user interface. In preparation for the 100 car study, software developers created a customized data viewer that allowed an analyst to quickly seek to events of interest and simultaneously play synchronized video and strip charts containing sensor data. This interface also provided integrated tools for adding annotation to the original data. Overtime the original viewing interface has undergone continuous improvements and is known in its current version as the Data Analysis and Reduction Tool (DART; Figure 7).

DART implements automated event identification algorithms to mark events that required manual, eyes-on, validation. Automated triggering algorithms take advantage of pattern detection methods that detect instances in which set thresholds are exceeded. Some examples of automated triggering algorithms identify instances of high accelerations, swerve events, and poor lane keeping performance. The DART classifies each of the triggered events it identifies and then queues the events for analyst validation.

The analyst opens the event and quickly assesses its validity. The validity check is an important step in the processes as sensor errors and imperfect triggering algorithms are not uncommon. Once validated, the analyst can annotate the data through the use of either epoch-level questionnaires or in-time data expansion. Epoch questionnaires allow the analyst to answer a number of questions about the event (such as even classification and adjacent vehicle presence). This provides an aggregated (sometimes called reduced) dataset describing the event of interest. In-time data expansion provides the ability for the analyst to add a new variable to the dataset. A common example of this is annotating eye-gaze in which the reductionist uses keystrokes to record the gaze region of the driver throughout the event of interest.

With the advent of the asynchronous DASs, a new data viewer is presently in development. The new viewer will provide a more fluid user experience. Since data is now loaded into large databases, the user will no longer open files but rather will locate drivers, trips, and events, of interest through novel navigation interfaces. There will also be new visualization tools that allow analysts to view aggregated data for entire collections as well as creating customized charts to display data of interest.

The new viewer is part of an entire re-tooling of the data analysis scheme. New database structures and permissions strategies are being implemented to prepare for the upcoming large data collection efforts. The new methods focus on development of standardized methods to provide users with a consistent experience regardless of which data collection they are working with. With software improvements, the typical time required for an analyst to validate and annotate an event of interest should be less than 4 minutes. This rate of analysis will allow higher production rates enabling eyes-on validation of a large number of events.
Most early analysis of naturalistic data focused on the aggregated data sets. This made analysis similar to the analyses performed on a crash database. Over time analysis methods have become more sophisticated. Case-control statistical methods are now being performed to develop accurate estimates of risk. Furthermore, functional data analysis is being used with increasing prevalence and will help researchers extract new insights from the naturalistic data. A number of new non-parametric methods are also being applied thanks to increasing processor performance, improvements in statistical analysis packages, and rapid data accesses.

**SUMMARY/CONCLUSIONS**

As evidenced throughout this paper, naturalistic data collection as it is widely employed today would not be feasible without modern data acquisition, storage, and analysis technologies. Research projects based on the method are providing valuable data this is leading to a safer drive for motorists. For example, some key discoveries found from naturalistic driving studies thus far include;

1. Most current database estimates, which rely primarily on self-report, place fatigue-related crashes at approximately 2% - 4% of the total crashes. Naturalistic data collection show moderate to severe fatigue is actually present in approximately 20% of both light and heavy vehicle crashes and near crashes and increases the probability of a crash or near crash event over six times that of driving while alert.

2. Teen drivers are four times more likely to be involved in a crash or near crash event while engaging in a secondary task (like cell phone dialing) than their adult counterparts.

3. Newly licensed teen drivers behave and perform largely like adults when an adult passenger is present but engage in much more risky behavior and exhibit poorer performance when an adult passenger is not present. This shows that teen are skilled drivers and understand the rules of the road and defensive driving techniques, but choose to engage in risky driving when unsupervised.

4. Engaging in visually demanding in-vehicle tasks increase the crash/near crash risk to a great degree. For example, in a heavy truck, texting while driving increases crash/near crash risk 23 times that of driving without engaging in a secondary tasks.

5. The largest single contributing factor in crashes and near-crashes is looking away from the roadway just prior to an unexpected event. This accounts for somewhere between 70% and 90% of unsafe events.

6. Roughly 10% of drivers account for almost 50% of the crash risk.

The new embedded technologies and data management methods will enable the next leap in naturalistic data collection. Research organizations will be collecting unparalleled data sets in terms of quantity and accuracy that will enable statistically robust conclusions that have not been possible through epidemiological or empirical methods. The new datasets will provide invaluable resource for identifying safety deficiencies and developing strategies to mitigate and eliminate them.

![Figure 7. Screen capture of the DART interface show data captured from a motorcycle](image-url)
REFERENCES


9. Hanowski, R.J., et al., The drowsy driver warning system field operational test, data collection final report. 2008: Washington, DC.

CONTACT INFORMATION

Zachary Doerzaph
Virginia Tech Transportation Institute
3500 Transportation Research Blvd
Blacksburg, VA 24162
Zac@vtti.vt.edu

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions made by the entire team at the Virginia Tech Transportation Institute with particular focus on the hardware and software engineers who enabled the naturalistic data collection method. Furthermore, we acknowledge the numerous partners who provided financial and technical support throughout the last several decades have brought naturalistic data collection into mainstream practice.