NATURALISTIC DRIVING DATA: MANAGING AND WORKING WITH LARGE DATABASES FOR ROAD AND TRAFFIC MANAGEMENT RESEARCH

K MURONGA and K VENTER

CSIR Built Environment, PO Box 395, Pretoria, 0001, Tel: 012 841 2337

e-mail: kmuronga@csir.co.za.

CSIR Built Environment, PO Box 395, Pretoria, 0001, Tel: 012 841 3856

e-mail: kventer@csir.co.za.

ABSTRACT

Naturalistic driving and field operational tests are used worldwide to collect data from drivers in order to better understand the human, vehicle and environment interactions. The fairly new methodology has already provided great insight into numerous driver behaviours that could previously not be observed directly. The data is collected with a data acquisition system which is installed in the vehicle. This system consists of cameras facing the driver (and passengers) as well as cameras facing outward. An on-board computer is installed in the vehicle and collects information about the vehicle. This information includes satellite positions, data and time as well as speed and acceleration and deceleration data. The system collects large volumes of data and the challenge is to manage this data efficiently as currently the datasets take-up much storage space, are in different formats necessitating that different software programs be used to download, transcribe and analyse the data. This paper provides an overview of the challenges experienced while working with these large data sets as well as some of the possible solutions identified. The findings and recommendations from this study should prove useful to other researchers and practitioners interested in working with naturalistic data.

1. INTRODUCTION

1.1. Background

1.1.1. Naturalistic Driving Data

Naturalistic Driving Studies (NDS) is/are a novel approach to the way that road safety research can be conducted in South Africa. The term "*naturalistic driving studies*" refers to an unobtrusive approach to studying driver behaviour. This methodology enables researchers to study driver behaviour in the context of the driving task and road environment as well as inform driver actions preceding crashes or near crash events. A vehicle or vehicles are instrumented with a data acquisition system (DAS). The DAS typically consists of an on-board computer which logs vehicle information obtained from various sensors. The data collected include date, time, satellite information, acceleration and deceleration data, speed and Global Positioning System (GPS) coordinates. Furthermore a number of cameras is connected to the vehicle and videos of the driver and driving environment are recorded and stored on the on-board computer. Data is logged every second that the vehicle is turned on. As can be expected a large number of quantitative and qualitative data are generated on a continuous basis.

1.1.2. CSIR ND study

In 2013, this methodology was applied in a South African context. Two DAS systems were obtained and installed in four participants' vehicles over a period of six months (2 systems in 4 vehicles –parent/child combinations for approximately 6 months each). Due to the volume of data that was/were being collected, it was necessary to download the data weekly or bi-weekly and to store the data in a secure manner. The data was downloaded and stored on an external hard drive.

1.1.3. Problem statement

The 1) volumes of data, the different types of data and the challenges that arose from working with this data, 2) warranted the investigation into strategies to manage this large database. There 3) was a need to integrate different datasets, to reduce the amount of time it took to standardise the data and to prepare it for eventual analysis.

1.1.4. Scope of this pape

This paper provides an overview of the challenges experienced in working with this large database. It explores potential strategies and methodologies which could in future be used to manage this data efficiently.

1.2. **Technology Background**

1.2.1. DAS Systems

The DAS system was obtained and installed in four participants' (volunteers) vehicles over a period of six months for each of the parent/child combinations. The research process is illustrated on Figure 1. Volunteers were asked to participate in the study. The technology was installed in the participants' own vehicles.

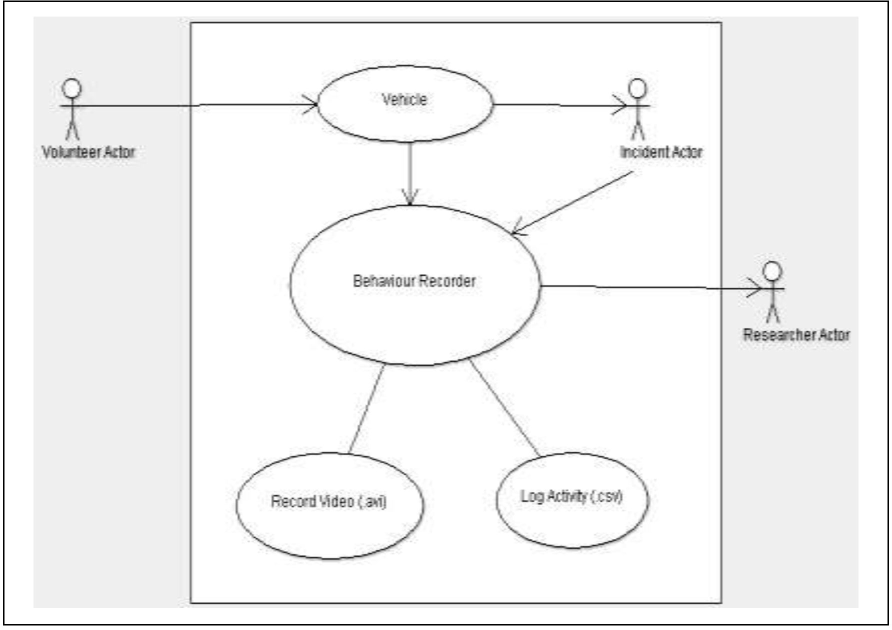

Figure 1: DAS System Use Case

The volunteer driver activated the system by switching the ignition on and driving the vehicle. All vehicle elements were recorded. The system also continuously recorded image material of the driving environment and driver. Every time an incident (i.e. speeding; drastic breaking; overtaking, etc.) occurs the incident actor will be activated. The incident is recorded in both video (.avi) and text (.csv) files simultaneously and this is recorded on a secure digital card (SD Card) for storage. The researcher then collects the card from the vehicle periodically to transfer the data to an external hard disk or server.

1.2.2. Data flows

There are two processes that are involved in the recording of the driver behaviour, when a DAS system is used. The first process records the video of the incident, storing it as an audio video interleaved file (.avi) and the second process writes the description of the incident on a text file (.csv), including GPS logs.

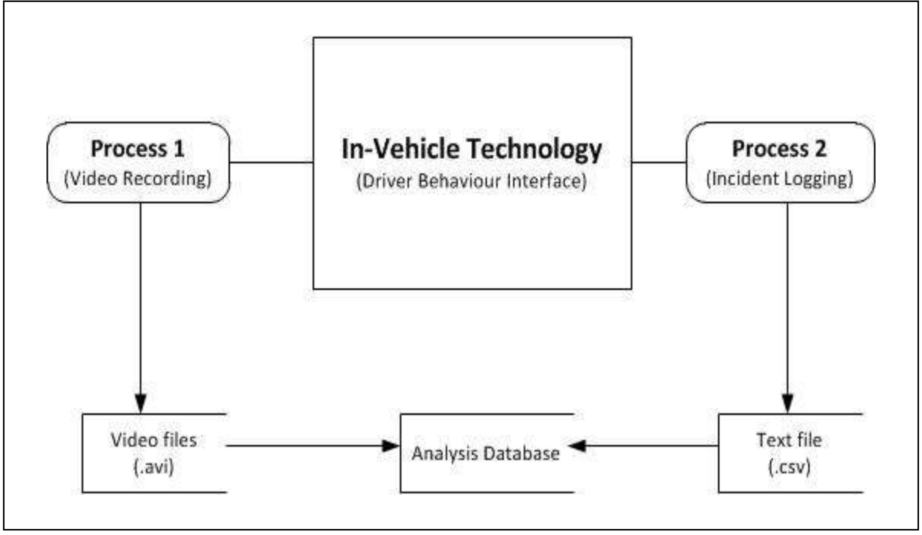

Figure 2: DAS system data flow

The data will then be transferred to an external storage facility and this needs to be done periodically to avoid data overwriting. The external storage will then be used as an analysis database.

2. CHALLENGES EXPERIENCED

2.1. <u>Data challenges</u>

Table 1 provides an overview of the type of data that was collected. After the data was/were downloaded from the DAS it was saved in two different Microsoft Excel files. The first file contained all the quantitative information and was adapted to include the headings as described above. The second file contained the qualitative information. It was essential that two types of data downloaded should have distinct identifying markers in order to later match the two types of data. This served the purpose of ensuring that the video and data and the recorded values from the data logger could be matched in the event that an incident or anomaly was detected in either of the files.

After the initial download, it was discovered that the first column of data (1-System time – local time) which should show the same as the GPS date and time, differed from the GPS time. This time stamp is also assigned to the videos. The GPS date and time (column 12) were correct and for the rest of this study use was made of the GPS time and date which provided the researcher with a continuous stream of data for every second, minute and hour that the system was turned on. However correcting this meant manually copying and pasting the *.csv* files into text files, transferring them into Microsoft Access, separating the date and time columns and then transferring them back to Excel. This seemed simple enough. However neither the Microsoft Access or Excel programmes were able to transfer the data in one step. It was again necessary to copy and paste the information manually from one programme to another.

Table 1: Type of data collected

Quantitative data	Qualitative data
1) System Time - Local Time	1) Driver Identification
2) G-Sensor X - Measured in mGal. (Unit of acceleration*)*	2)Date on which the data was downloaded
3) G-Sensor Y - Measured in mGal. (Unit of acceleration*)*	3) Date on which the video was generated
4) G-Sensor Z - Measured in mGal. (Unit of acceleration*)*	4) Trip identification number in log files (from data logger)
5) Satellite Fix Status - A value of “Y” indicates that a fix is currently obtained, whereas a value of “N” indicates that a fix is not obtained.	5) Video identification in video (normal file)
6) Latitude - Represents the current distance north or south of the equator. ("+" indicates north and "-" indicates south.)	6) Length of video in minutes
7) Longitude - Represents the current distance east or west of the Prime Meridian. (A value of "+" indicates east and "-" indicates west.)	7) Start time of video
8) Altitude - Antenna altitude above/below mean sea level, measured in meters.	8) End time of video
9) Speed - Indicates the current rate of travel over land, measured in km/h.	9) Length of trip per video
10) Heading - Indicates the current direction of travel over, measured as an "azimuth".	
11) Satellites - Number of satellites in use/Total number of satellites in view. (HDOP – indicators relative accuracy of horizontal position)	
12) GPS Time (UTC) - Date and Time calculated from GPS satellite signals.	

In order to code the videos, each video downloaded had to be transcribed into a “readable” .avi file. Each video generated 3 .avi files (driver, front and rear cameras).

Table 2: Steps in downloading and coding the videos/image material

Step 1	Transcribe videos (3) from BX4000 system into .avi files.
Step 2	Select videos per week for coding
Step 3	Activate and load the video in MAXQDA
Step 4	Look through video material in MAXQDA- (2 cameras driver and front)
Step 5	Assign time stamps to applicable behaviour
Step 6	Code each scenario allocated a time stamp
Step 7	Export codes and memos to Excel spreadsheet
Step 8	Manually correlate coded behaviour with the vehicle movements recorded in the data logger
Step 9	Analyse according to frequency, escalation of observable behaviour

The researchers experienced a huge challenge in terms of converting the visual data (driver’s behaviour) into a qualitative coding framework and integrating these with the incident logged text files (quantitative data).

Dealing with these large datasets manually was time consuming, and an inefficient way of dealing with the problem of downloading, transferring and integrating the different datasets.

2.2. Volume of data

Over the study period (238 days):

- More than a million vehicle movements were collected (1 284 153);
- 1755 videos were collected amounting to 255 hours of video material;
- A total distance of 14 119 kilometres was travelled.

These large volumes of data were stored on a 4 terabyte external hard drive. Despite this the large volumes of data slowed the computer down and it was realised that if this type of data is/are to be collected on a regular basis, a more powerful computer will be needed to not only store but also to analyse the data. In addition the automation of the process needs to be considered as the current manual approach is time consuming and resource intensive.

3. DESCRIPTION OF VERY LARGE DATABASES (VLDB)

3.1. Definition

A database is defined as "*an organised collection of data managed using a database management system*" (Wikramanayake and Goonetillake, 2012). Databases are used for storing and working with the data in order to ultimately manipulate data in such a manner that new meanings can be applied to the data.

3.2. Characteristics

In 1999, Jacobson et al. started to explore the use of video data for quantitative and qualitative research. The researchers emphasised that then the use of large scale video research studies was not common but that it might gain momentum as technology and software develops. Thirteen years later this has indeed become true and NDS is considered to be one of the most comprehensive research tools currently available to road safety researchers.

According to Wikramanayake et al. (2012) the term "*large*" is relative due to the fact that databases are growing along with technology and what is considered to be large today might not be considered large in future. However the researchers state that Very Large Databases (VLDB) are characterised by high numbers of rows or records or occupy large physical file system storage space due to wide tables with large numbers of columns or due to multimedia objects. The authors indicate that a VLDB typically contains more than 1 terabyte or billions of rows. The ND data is an example of both large numbers of entries as well as taking up more than a terabyte of space.

Tools that are used to analyse and manipulate scientific data are lacking behind in technology as compared to tools that are used to collect and store data. These present many challenges in managing the life cycle of scientific data (Grey et al. 2005). Researchers are challenged in deciding on what data to keep and for how long. Storing every data element presents a challenge as data grows too fast while creating large databases in the process.

4. POSSIBLE STRATEGIES TO MANAGE THE VLDB.

4.1 Hardware Storage

In this study the researchers tried to eliminate some of the challenges by upgrading the hardware used for storing and analysing the data. The desktop computer was upgraded with the following hardware:

- Intel Xeon CPU E5-2620@2.00GHZ
- RAM 16GB
- 1TB Internal hard disk space
- 4TB External hard disk space

Important considerations are to improve the processing speed of the computer as well as providing more storage space for the data. The DAS system only allows for saving of live incident data onto a SD Card and the researchers are also investigating strategies to improve the DAS in order to allow for other media to be used to prolong recording time with less periodical downloads. Initially a provision was made to make use of CSIR servers to store analysed data for safety purposes. Confidentiality issues are however an important consideration as video image material is used. This meant that the data could not be stored in traditional CSIR databases and provision had to be made to keep the data in a separate "off-line" secure location.

4.2 Proposed strategies for integration

Four integrative strategies for mixed methods analysis was identified and applied to this ND database (Bazeley, 2006):

- Data transformation, where one form of data is transformed into another for further analysis (Bazeley, 2006). Management and search strategies need to consider both types of data that is included in the ND databases. Johnson et al. (2010) and Morucutti et al (2010) proposed a semi-quantitative approach to the analyses of this type of data where segments of qualitative data is transcribed and coded that numbers are assigned to the qualitative segments in order to analyse the data. Hellerstein (2008) proposes that priority should be given to data profiling which is described as a process which gives an overview of the dataset in order to provide structure and values in the database. This would entail ensuring that quantitative data (integers, time series and so forth) are converted to uniform measurable units.
- Typology development where a classification scheme and categories are developed from one set of data and applied to the other (Bazeley, 2006). Qualitative data needs to be categorised with unique keys which can be used as identifiers (Hellerstein, 2008). Knoll and Stigler (1999) states that when working with video and the secondary data sources that has been generated (such as for example in the ND study) the data search procedures is extremely important.
- Extreme case analysis , in which the outliers or residuals revealed by one analysis are explored using alternative data or methods; and
- Data consolidation/merging to create new variables for use in further analysis (Bazeley, 2006).

Bazeley (2006) further states that other statistical techniques that might be useful include cluster analysis, correspondence analysis, and multidimensional scaling.

One particular challenge that needs to be addressed is the automation of “merging” and “matching” the corresponding quantitative and qualitative data. The research team is still investigating available theories and solutions that can be used for automating the merging of visual and text data into possibly a SQL database where the coding framework can be applied. From the literature a number of possibilities have been identified and are currently under investigation. Rajan et al (2006) explored algorithms for organizing and indexing compressed video data into hierarchical trees. The complex video data can be quickly summarised and discriminating characteristics highlighted.

Object identification and pattern recognition is possibly the most sensible strategies to be considered. For example when searching the data, 3D C-string pattern recognition in the videos entails using the projections of objects spatial and temporal relations between the objects in a video can be identified (Lee, Chiu and Yu, 2002). The 3D-C string takes into consideration changes in the size of objects which when working with the ND data could be useful to for example identify road furniture (stop streets, traffic lights etc.) which will change in size, depending on the speed of the vehicle. Subspace morphing theory differs from other object identification theories as it does need normalisation of images and scales but rather makes use of the differences in the images to recognise objects in video that differ in size and dimensions (Zhang and Srihari, 2002).

5. CONCLUSION

The existing methodological and theoretical approaches for integrating large ND databases are very few. Currently the research team has not yet identified a single strategy for managing, integrating or searching this large database. It is however envisioned that the solution will probably entail a combination of theories and methods which is the subject of the teams’ future research.

REFERENCES

Bazeley, P., 2006, ‘The Contribution of Computer Software to Integrating Qualitative and Quantitative Data and Analyses’, *Research in the Schools Mid-South Educational Research Association* 13(1), 64-74.

Gray, J., Liu, D., Nieto-Santisteban, M., Szalay, A., DeWitt, D., Heber, G., 2005, “Scientific Data Management in the Coming Decade’, *CTWatch Quarterly,* 1(1).

Hellerstein, J.M.,2008, Quantitative Data Cleaning for Large Databases, viewed 27 January 2014 from http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf.

Jacobs, J. K., Kawanaka, T. and Stigler, J. W., 1999, ‘Integrating qualitative and quantitative approaches to the analysis of video data on classroom teaching’, *International Journal of Educational Research* 31, 717-724.

Johnson, B. J., Dunlap, E. and Benoit, E., 2010, ‘Structured Qualitative Research: Organizing “Mountains of Words” for Data Analysis, both Qualitative and Quantitative’, *National Institute of Health*, 1-22.

Knoll, S. and Stigler, J. W., 1999, ‘Management and analysis of large-scale video surveys using the software vPrism^{TM}, *International Journal of Education Research* 31(8), 725-734.

Lee, A. J. T. Chiu, H. P. and Yu, P., 2002, ‘3D C-string: a new spatio-temporal knowledge representation for video database systems’, *Pattern Recognition*, 35, 2521 – 2537

Morocutti, J. and Zanardini, F., 2010, ‘Managing the Electronic Collection with Qualitative and Quantitative data- A case study: the Wiley-Blackwell collection at the University of Milan’ *76th IFLA Conference - Measuring usage and understanding users, e-resource statistics and what they teach us*, Stockholm, 8 August 2010.

Rajan, H., and Chia, L.T., 2006, ‘A motion-based scene tree for compressed video content management’, *Image and Vision Computing*, 24, 131–142.

Wikramanayake, G. N., and Goonetillake,J., S.,2012, ‘Managing Very Large Databases and Data Warehousing’. *Sri Lankan Journal of Librarianship and Information Management.* 2(1), 22-29.

Zhang, M., and Srihari, R.K.,2002, ‘Subspace morphing theory for appearance based object identification’, *Pattern Recognition*, 35, 2389 – 2396.