Accident Analysis and Prevention 72 (2014) 177–183

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Computer vision and driver distraction: Developing a behaviour-flagging protocol for naturalistic driving data

Jonny Kuo[a,*], Sjaan Koppel[a], Judith L. Charlton[a], Christina M. Rudin-Brown[b]

[a] Monash University Accident Research Centre (MUARC), Monash University, Australia
[b] Human Factors North, Inc., 174 Spadina Avenue, Suite 202, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 1 April 2014
Received in revised form 2 June 2014
Accepted 6 June 2014
Available online 23 July 2014

Keywords:
Naturalistic driving
Driver distraction
Observational study
Computer vision
Video processing
Machine learning

ABSTRACT

Naturalistic driving studies (NDS) allow researchers to discreetly observe everyday, real-world driving to better understand the risk factors that contribute to hazardous situations. In particular, NDS designs provide high ecological validity in the study of driver distraction. With increasing dataset sizes, current best practice of manually reviewing videos to classify the occurrence of driving behaviours, including those that are indicative of distraction, is becoming increasingly impractical. Current statistical solutions underutilise available data and create further epistemic problems. Similarly, technical solutions such as eye-tracking often require dedicated hardware that is not readily accessible or feasible to use. A computer vision solution based on open-source software was developed and tested to improve the accuracy and speed of processing NDS video data for the purpose of quantifying the occurrence of driver distraction. Using classifier cascades, manually-reviewed video data from a previously published NDS was reanalysed and used as a benchmark of current best practice for performance comparison. Two software coding systems were developed – one based on hierarchical clustering (HC), and one based on gender differences (MF). Compared to manual video coding, HC achieved 86 percent concordance, 55 percent reduction in processing time, and classified an additional 69 percent of target behaviour not previously identified through manual review. MF achieved 67 percent concordance, a 75 percent reduction in processing time, and classified an additional 35 percent of target behaviour not identified through manual review. The findings highlight the improvements in processing speed and correctly classifying target behaviours achievable through the use of custom developed computer vision solutions. Suggestions for improved system performance and wider implementation are discussed.

1. Introduction

Driving has been described as a 'satisficing' task, for which drivers will develop and invest only the minimum level of skill and attention required to complete it (Hancock et al., 2008). As such, driver engagement in secondary tasks has been found to be highly prevalent, with insufficient attention being directed at tasks necessary for safe driving. (Young and Lenné, 2010). Research has shown that up to 23 percent of crashes and near-crashes can be attributed to driver distraction, and that when drivers direct their gaze away from the forward traffic scene for more than 2 s, their crash risk is more than doubled (Klauer et al., 2006).

Extended periods of data collection through the use of discreet, in-car video cameras has allowed for naturalistic driving studies (NDS) to objectively capture aspects of everyday driving, including those that may be indicative of driver distraction, that were previously inaccessible to researchers (Klauer et al., 2006; Hanowski et al., 2005; Stutts et al., 2005). However, with the significantly increased volume of data generated comes the potentially challenging and inherently error-prone task of observation and interpretation by human analysts. This gives rise to both logistical and inferential limitations. Firstly, the manual processing of NDS data by human analysts becomes more time and labour-intensive with growing dataset sizes. Previous pilot research by the authors has yielded 150 h of video footage (Koppel et al., 2011), with current efforts aiming for 700 h (Sun et al., 2012). To limit the total amount of data that analysts need to view, one approach has been the use of various statistical sampling methods to select a subset of data from the complete dataset to analyse (Stutts et al., 2005; Koppel et al., 2011). Such protocols can be easily implemented without the

* Corresponding author at: Monash Injury Research Institute, Building 70, Monash University, Victoria 3800, Australia. Tel.: +61 3 9905 1808.
E-mail address: jonny.kuo@monash.edu (J. Kuo).

http://dx.doi.org/10.1016/j.aap.2014.06.007

need for specialised hardware or software. However, fundamental statistical assumptions are made regarding the representativeness of the selected subset, the veracity of which is difficult to assess.

Another approach to data reduction has been the use of video triggers such as vehicle performance data (i.e. only reviewing epochs of video data that are temporally correlated with sudden braking or swerving manoeuvres, as recorded by vehicle 'black box'-type devices) (Klauer et al., 2006). However, while these critical incident-triggered epochs offer valuable insight into the proportion of crashes/near-crashes attributable to driver distraction, the prevalence ratios of driver distraction may not be validly inferred without also considering instances where the occurrence of driver distraction did not result in a critical incident. In addition to these reasons for specifically measuring the occurrence of driver distraction, distraction-triggered data (as opposed to crash-triggered) may provide unique insight into the mechanisms that differentiate incidents of distraction which result in crashes and those that do not.

Eye-tracking technologies such as FACELAB are an example of a distraction-centred approach to data reduction, using driver glance location as a surrogate measure of where a driver is directing his or her attention (Taylor et al., 2013). The implementation of dedicated eye-tracking hardware for NDS data collection has allowed researchers to gain a high level of detail relating to visual distraction of drivers as it occurs in their natural environment (Liang et al., 2012; Ahlstrom et al., 2012). However, eye-tracking using currently available solutions remains an a priori venture, requiring forethought in research design. While these tools offer researchers a high degree of fidelity in what they measure, datasets collected without such applications in mind (or before the development of these tools) remain incompatible and must rely on conventional manual coding protocols. This represents a significant underutilisation of resources, both in the quantity of data left unexamined and in the need for manual coding when automated techniques exist.

The application of machine learning and computer vision solutions to NDS data offers a promising approach to resolving the issues described above, with many sophisticated applications based on driver face-tracking developed in the field of computer science (Bergasa et al., 2008; Rezaei and Klette, 2011). However, few of these applications have been tested extensively with large NDS datasets. There is a need to develop machine learning solutions that are resilient to the inherent 'noise' in naturalistic data (Young et al., 2008), and that not only accommodate the physical and technical limitations of existing NDS datasets, but that may also potentially be used to analyse future datasets collected without the use of such dedicated hardware.

To address the challenges posed by manual coding protocols, the aims of the present study were to develop a computer vision solution for classifying driver glance behaviour captured using video-recording during NDS. Additionally, using results derived through manual coding from the Children in Cars data set (Koppel et al., 2011) as a benchmark of current best practice, a second aim was to compare the accuracy and speed of processing achievable by a computer vision solution. It was hypothesised that manual coding and computer vision approaches would differ significantly, with computer vision processing correctly classifying a greater number of off-road glances whilst requiring less processing time.

2. Method

2.1. Computer vision algorithms

Custom software was developed using Python programming language (http://www.python.org) and an open-source computer vision library, SimpleCV (http://www.simplecv.org). These tools were selected for their high level of abstraction, allowing for rapid software development. Specifically the find HaarFeatures module of SimpleCV was used for face detection. This module is based on the Viola and Jones (2001) framework for face detection. The technique makes use of a series of adjacent dark and light rectangular regions (i.e. classifier cascades) to identify whether target features are present in an image. For a classifier cascade to be able to recognise a specified target feature, it must first be 'trained' by being presented with examples of what is a correct instance (a positive image) of the feature and what is an incorrect instance (a negative image). This training process is computationally intensive and is typically an application-specific process, requiring many thousand, manually selected and cropped examples of positive and negative images (Lienhart et al., 2002). Compared to the labour-intensive process of manual review, the selection of these training images need only be performed once and may subsequently be used to classify any number of images (given the same target feature and environmental conditions). For face tracking applications, researchers have proposed the training and use of multiple classifiers to account for different head positions and lighting effects (Jones and Viola, 2003). Head position specificity of classifier cascades was exploited in the present study as a robust method to identify instances when drivers turned their heads away from the forward traffic scene.

Perhaps due to the limited range of participant faces available in the present dataset, preliminary analyses showed more performance variability between participants than among different lighting conditions, suggesting the need for multiple participant-specific classifiers.

To this end, two approaches were implemented. In the first approach, separate classifiers were developed for male and female drivers. The decision to discriminate on driver gender was based on visual observation of a gender difference in hair styles which was hypothesised to manifest as highly salient differences in light and dark regions, as per the underlying mechanisms of the Viola and Jones (2001) technique. In the second approach, a statistical-based method was used. Hierarchical cluster analysis was performed on averaged images of each driver to determine the minimum number of classifiers that would need to be trained. Full results of this analysis are presented in Section 3.1. In brief, three clusters were identified: two clusters of three male drivers each, and one cluster consisting of six female drivers plus one male driver.

2.2. Datasets

The dataset from which the test and training sets were drawn consisted of 621 discrete journeys (i.e. 165 h of vehicle travel). A summary of the data management protocol is presented in Fig. 1.

Participant characteristics, recruitment, and procedure used in obtaining the data set have been previously documented (Koppel et al., 2011; Charlton et al., 2010). In brief, 12 families were recruited from an existing Monash University Accident Research Centre (MUARC) database on the basis of regularly driving at least one child between the ages of 1 and 8 years who were typically seated in a child restraint system (CRS) in the backseat. Families were provided with a luxury model family sedan for a period of three weeks, during which they were instructed to drive as per their usual routines. The study vehicle was fitted with four discreet colour cameras set to automatically record driver and passenger in-vehicle behaviours. The following perspectives were recorded through the video system: the forward traffic scene, a view of the driver and front seat passenger, the rear left passenger, and the rear right passenger.