Accident Analysis and Prevention xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Validity and reliability of naturalistic driving scene categorization Judgments from crowdsourcing

Christopher D.D. Cabrall[a,*], Zhenji Lu[a], Miltos Kyriakidis[a,b], Laura Manca[a], Chris Dijksterhuis[a,c], Riender Happee[a], Joost de Winter[a]

[a] Delft University of Technology, The Netherlands
[b] ETH Zurich, The Netherlands
[c] Hanze University of Applied Sciences, The Netherlands

ARTICLE INFO

Keywords:
Naturalistic driving study
Data annotation
Crowdsourcing
Validity
Reliability
Dash cam

ABSTRACT

A common challenge with processing naturalistic driving data is that humans may need to categorize great volumes of recorded visual information. By means of the online platform CrowdFlower, we investigated the potential of crowdsourcing to categorize driving scene features (i.e., presence of other road users, straight road segments, etc.) at greater scale than a single person or a small team of researchers would be capable of. In total, 200 workers from 46 different countries participated in 1.5 days. Validity and reliability were examined, both with and without embedding researcher generated control questions via the CrowdFlower mechanism known as Gold Test Questions (GTQs).

By employing GTQs, we found significantly more valid (accurate) and reliable (consistent) identification of driving scene items from external workers. Specifically, at a small scale CrowdFlower Job of 48 three-second video segments, an accuracy (i.e., relative to the ratings of a confederate researcher) of 91% on items was found with GTQs compared to 78% without. A difference in bias was found, where without GTQs, external workers returned more false positives than with GTQs. At a larger scale CrowdFlower Job making exclusive use of GTQs, 12,862 three-second video segments were released for annotation. Infeasible (and self-defeating) to check the accuracy of each at this scale, a random subset of 1012 categorizations was validated and returned similar levels of accuracy (95%).

In the small scale Job, where full video segments were repeated in triplicate, the percentage of unanimous agreement on the items was found significantly more consistent when using GTQs (90%) than without them (65%). Additionally, in the larger scale Job (where a single second of a video segment was overlapped by ratings of three sequentially neighboring segments), a mean unanimity of 94% was obtained with validated-as-correct ratings and 91% with non-validated ratings. Because the video segments overlapped in full for the small scale Job, and in part for the larger scale Job, it should be noted that such reliability reported here may not be directly comparable. Nonetheless, such results are both indicative of high levels of obtained rating reliability.

Overall, our results provide compelling evidence for CrowdFlower, via use of GTQs, being able to yield more accurate and consistent crowdsourced categorizations of naturalistic driving scene contents than when used without such a control mechanism. Such annotations in such short periods of time present a potentially powerful resource in driving research and driving automation development.

1. Introduction

Further knowledge specifically of (background) driving scene contexts could benefit transportation research and ultimately road safety. This study presents and evaluates a new method using crowdsourcing to provide content characterizations of natural driving video footage. Brief descriptions of both topics are provided in the following introductory sections.

1.1. Naturalistic driving and driving videos

Naturalistic driving studies (NDS) have been growing in popularity with much success over the last few decades. NDS offer advantages with respect to other traditional driving safety research methods such as eye

* Corresponding author at: Mekelweg 2, 2628 CD, Delft, The Netherlands.
E-mail address: c.d.d.cabrall@tudelft.nl (C.D.D. Cabrall).

http://dx.doi.org/10.1016/j.aap.2017.08.036
Received 17 August 2016; Received in revised form 30 August 2017; Accepted 31 August 2017

witness recall (often being inaccurate or unavailable) within crash data evidence approaches and driving simulators (often causing artificial participant behavior) (Regan et al., 2012). However, a lack of experimental control (where extraneous variables except that of manipulative interest are held constant), has been a commonly recognized detriment to NDS. Thus, the accurate annotation of the situational aspects and conditional characteristics that freely vary in NDS becomes all the more important for the identification and understanding of potential causal factors. Augmented by accelerating developments in audio-visual technology, computing, and networking resources, blended research designs are emerging wherein stimuli can be naturally sourced from the real world, reproduced, and mixed with more controlled laboratory conditions.

Due to reductions both in size and costs of cameras, real life driving video is an increasingly accessible data resource that may allow recordings at a large scale and could help enrich other sources of data with otherwise missed contextualized information. However, so much video data might be recorded in naturalistic driving research and field operational tests that research resources are often overwhelmed to process such data libraries through pre-requisite rounds of organization and labeling (e.g., data reduction) towards fuller potentials of use. For example, challenges can arise regarding the availability of confederate researchers for laborious manual annotation or transcription tasks. Unfortunately for driving safety research, the use of real-life driving video footage has remained a relatively low-tapped exception (e.g., Crundall et al., 1999; Chapman et al., 2007; Borowsky et al., 2010) rather than a common resource, despite inherent strengths in face validity and generalizability of results.

1.2. Crowdsourcing

Compared to less than 1% in 1995, about 48% of the world population has an Internet connection to date, placing the approximate number of Internet users in excess of 3.5 billion people (www.InternetLiveStats.com/internet-users/). Online crowdsourcing services make use of this extensive connectivity to create an on-call global workforce to complete large projects in small chunks (a.k.a., micro-task workers). Gosling and Mason (2015) review a broad and growing use of Internet resources in recent psychological research. They conclude that harnessing large, diverse, and real-world data sets presents new opportunities that can increase the societal impact of psychological research. In the automated driving domain, research has recently begun to emerge utilizing crowdsourcing resources through global survey initiatives to capture large scale international public opinion (Bazilinskyy and De Winter, 2015; Kyriakidis et al., 2015). In regards to crowdsourcing as a research method, investigation into the differences between laboratory participants versus crowdworkers has found faster responses but higher false alarms with crowdsourcing (Smucker and Jethani, 2011). Additional methodological research has revolved around the assurance of quality from the quick and inexpensive results typically returned by crowdsourcing and have recommended predetermined answer sets for use both in the screening of unethical workers as well as for the effective training of ethical workers (Le et al., 2010; Soleymani and Larson, 2010).

1.3. Present study

Real-world driving datasets come with large labor challenges in terms of data reduction like manual annotation and categorization. Pairing together expansive datasets of naturalistic driving video footage with crowdworkers may be a powerful method for progressing driving safety research. As a prototypical example of the power of crowdsourcing, the online platform known as CrowdFlower can accomplish routine categorization work at relatively low cost and at high speed by distributing the work around the world, taking advantage of both differences in time zones and hourly wages. However, such new methods require an investigation of validity and reliability to ensure trustworthy results might still be retained when scaling up beyond a single researcher or small research team. The present study investigated the use of CrowdFlower in the categorization of large amounts of videos with diverse driving scene contents (i.e., presence of another vehicle, straight road segments, etc.) through manipulation of one of its central quality control mechanisms to ascertain the quality and capability of such a method.

2. Methods

2.1. Quality control settings

Within its documentation, the CrowdFlower system promotes Gold Test Questions (GTQ) as its most important quality control mechanism. By configuring this setting, we enforced that a set of categorizations with known answers (i.e., given by the experimenters) were randomly intermixed with the experimental categorizations of interest. Thresholds of performance on these GTQs were set in an attempt to reduce the amount of indiscriminate responses that may occur within the results due to the remotely distributed nature of work under unsupervised conditions.

2.2. Participants/Workers

Participants in this research consisted of external micro-task workers from the online CrowdFlower contributor community. From this network, workers were prescreened by a number of criteria selectable within the CrowdFlower interface. Specifically, within CrowdFlower, performance levels are automatically awarded based on CrowdFlower's criteria of accuracy across a variety of different Job types. We selected a performance setting of Level 2 workers from a three-level scale, representing the midpoint between anchors of "highest speed" (Level 1) and "highest quality" (Level 3). Moreover, across all 51 of its current possible Channels for sourcing external workers (e.g. BitcoinGet, ClixSense, CoinWorker.com, etc.), CrowdFlower was set to include workers only from those retaining a ratio of Trusted to Untrusted Judgments greater or equal to 80% (39 Channels were left toggled on and 12 set to off). All countries were permitted within the Geography setting, and no additional Language Capability requirements were selected.

Table 1 lists the countries and source Channels of workers obtained across different sets of categorizations performed within the present study along with distributions of unique worker IP addresses and CrowdFlower worker IDs while Fig. 1 depicts the country distribution of the workers. For external crowdworkers, identification of country was determined by CrowdFlower based on IP address.

2.3. Apparatus and stimuli

To support projects oriented around the human factors of automated driving (i.e., exposing participants to various HMI/functional research concepts, measuring constructs of vigilance, situation awareness, mental models, reaction time, eye tracking behavior, etc.), a set of stimulus material was desired that had both qualities of high visual realism and controllable levels of uncertainty in repetition, freezeability, etc. Initial searches of YouTube with the keyword "dash cam" were conducted to compile a sample database of naturalistic driving video footage. Videos had to feature relatively high and consistent visual quality; a large and consistent field of view; and uninterrupted driving in order to be included. Candidate videos were selected from the search results in order to acquire nominal driving footage (i.e.; excluding violations and crashes). We collected a set of 10 freely available YouTube videos ranging between 1 min and 1 h duration (but of bimodal typicality of about 3 or 13 min length) for a total of 6934 s of driving footage. The countries in which the recordings were filmed