Accident Analysis and Prevention 114 (2018) 25–33

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, Future Resilient Systems, Singapore – ETH Centre, Singapore
[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
Available online 12 September 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

Table 1
Overview of the five different sets of categorizations. These sets included differences in the amount of video segments to be categorized (C1 = 48 segments, C2 = 12,862 segments), the use of Gold Test Questions (C1b had none) and the relation of the annotators to the research (external = CrowdFlower workers; internal = confederate research team).

Condition	Countries (ISO 3166-1 alpha-3)	Channels	Unique IP's	Unique ID's
C1a	15 = AUT, BEL, COL, DEU, ESP, GBR, GRC, IND, MKD, PHL, PRT, ROU, RUS, SRB, TUR	5 = clixsense, coinworker, elite, prodege, tremorgames	18	18
C1b	9 = DNK, GRC, IND, MDA, PAK, PHL, SRB, TUR, VNM	3 = clixsense, elite, tremorgames	13	13
C1c	1 = NLD	1 = n/a (internal)	1	1
C2a	46 = ARG, AUS, AUT, BEL, BGD, BGR, BIH, BRA, CAN, CHL, CZE, DEU, ESP, FIN, FRA, GBR, GRC, HRV, HUN, IDN, IND, ISR, ITA, JAM, LKA, MAR, MDA, MEX, MKD, MYS, PER, PHL, POL, PRT, ROU, RUS, SAU, SRB, SWE, TUR, TWN, UKR, URY, USA, VEN, VNM	16 = clixsense, coinworker, fusioncash, gifthulk, hiving, indivillagetest, instagc, personaly, pocketmoneygpt, points2shop, prodege, superrewards, surveymad, tremorgames, yute_jamaica, zoombucks	247	200
C2c	1 = NLD	n/a (internal)	12	7

Note. Country abbreviations are according to ISO 3166-1 alpha-3.

were not known; but driving was always on the right hand side. Audio was removed from the videos.

Subsequently, new self-recorded dash cam driving recordings (6026 s) were filmed in the United States and saved as 39 different files (typically less than 3 min in length, but ranging up to 15 min). This complemented the videos collected from YouTube in order to exhibit a broader range of real-life and experimentally interesting driving situations. These additional recordings included driving at night, on mostly empty desert roads, in a visually complex metropolis, and via multi-lane freeways, as well as at different driving speeds.

Driving videos from both sources were uploaded as 49 new private link-only access YouTube videos (M = 264 s duration) with an aggregate of 12,960 s of near driver point-of-view video footage. Through a combination of MATLAB script and an online tool from www.tech-tipsforall.com (ttfaloopandrepeat.appspot.com), auto-cueing URL links were generated to access each of the 12,862 possible 3-s segments from each of these 49 video. These URL links were embedded as text only in our CrowdFlower surveys with one URL per Judgment. The video segments overlapped in a manner such that a randomly selected worker categorized seconds one to three from video 1, another randomly selected worker categorized seconds two to four from video 1, a third randomly selected worker categorized seconds three to five from video 1, etc., for all videos 1 through 49. Example screenshots from the driving video segments are shown in Fig. 2a–c.

A coding scheme was created wherein each video segment categorization (i.e., Judgment) contained two groups of questions. The first group consisted of 21 checkbox items pertaining to the non-mutually exclusive presence of others, namely, (1) cars/trucks/vans/buses, (2) motorcycles/scooters/mopeds, (3) bicycles, and (4) pedestrians. Each of these four categories contained additional possible sub-specification of their position/direction of travel, namely, (5–8) leading, (9–12) oncoming, (13–16) passing or being passed, and (17–20) crossing; all relative to the present point-of-view vehicle. Additionally, there was a checkbox item which should be ticked for (21) no one else was present.

The second group consisted of 10 checkbox items pertaining to presence of miscellaneous infrastructural elements and aspects of vehicle

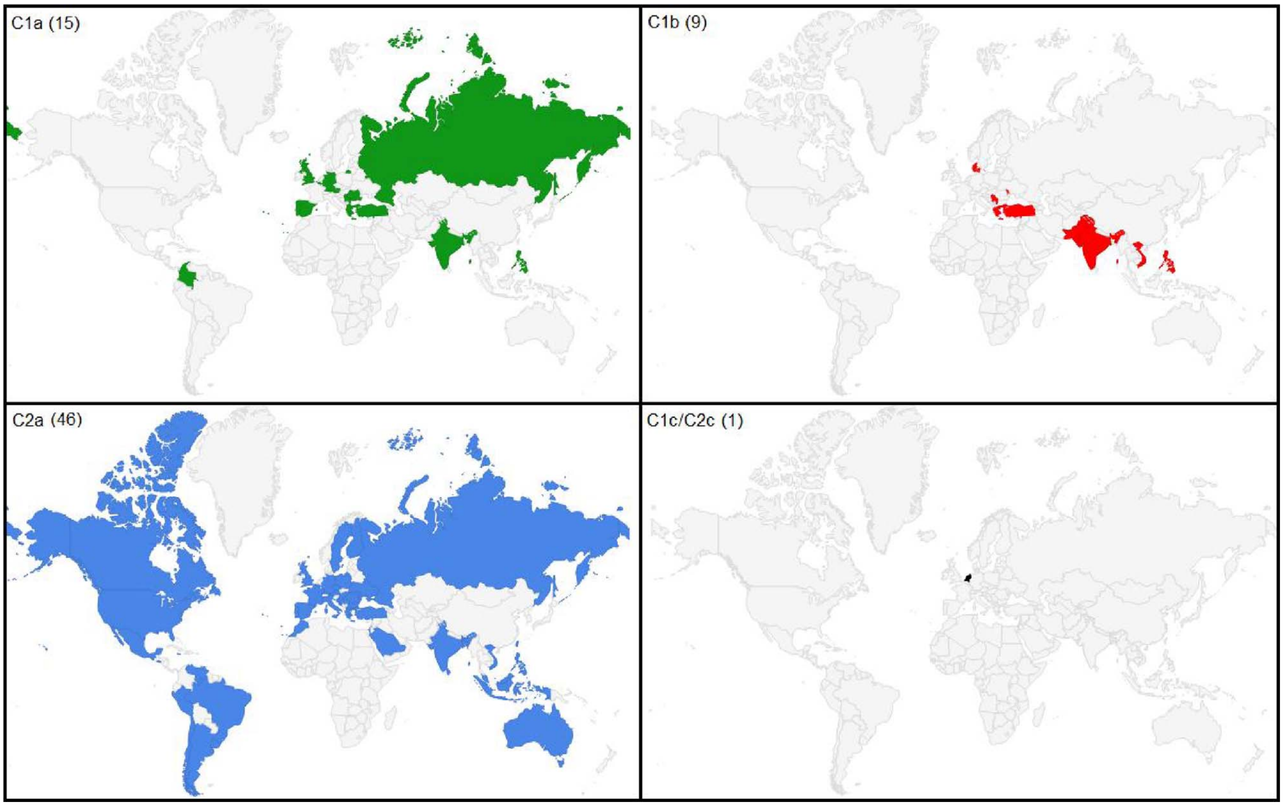

Fig. 1. Annotator country locations by condition.

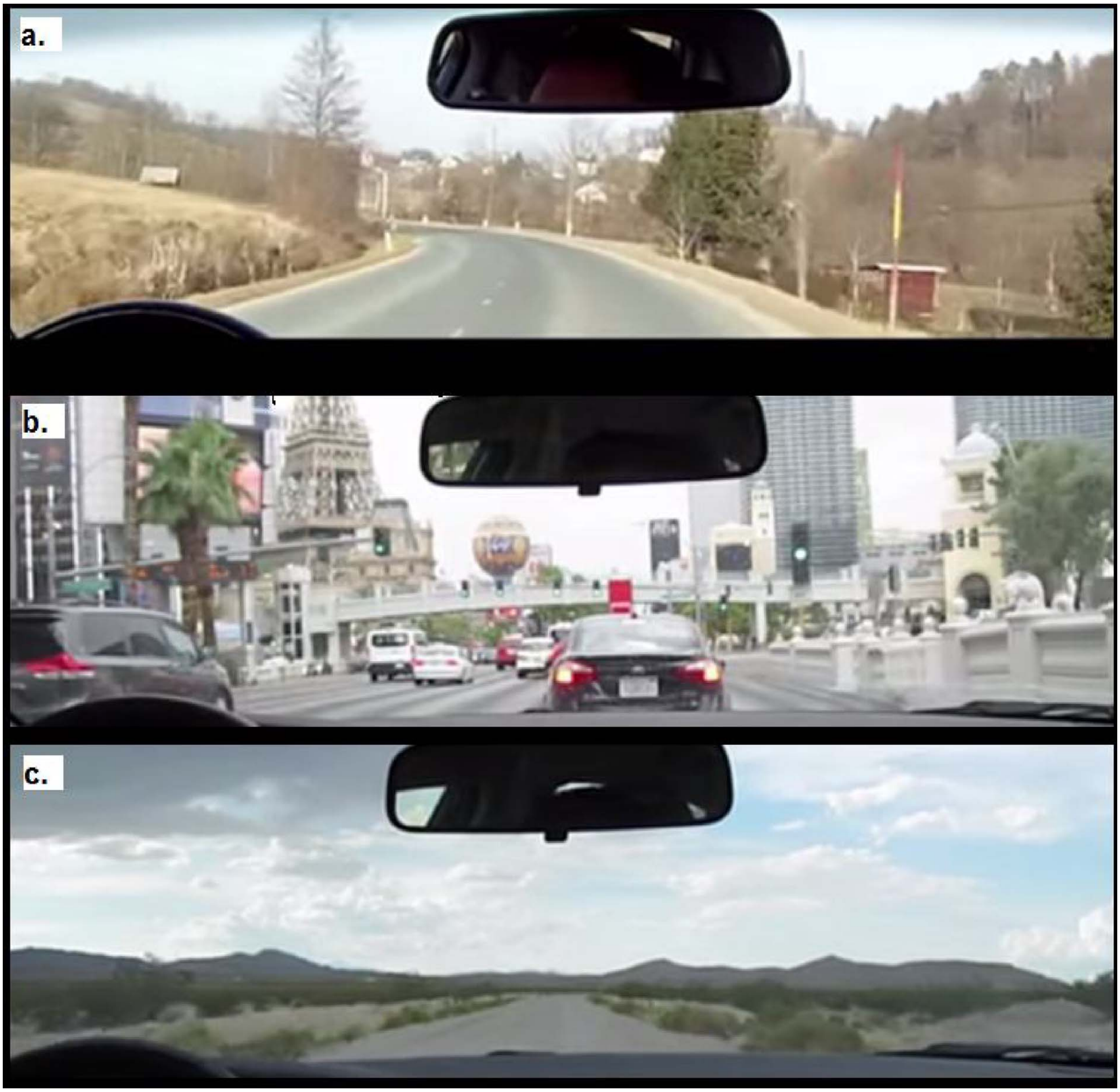

Fig. 2. Example screenshots from driving video segments a) recorded from within a publically posted dash cam YouTube video, b) recorded by the experimenters within a visually complex metropolis (i.e., Las Vegas strip), and c) recorded by the experimenters in a visually simple environment (i.e, Nevada desert backroad). Video resolution/quality here is only approximately representative as that initially made available to participants because differences in devices and browsers, full-screen viewing, etc. were not controlled for in the online survey.

behavior. These were: (1) straight road, (2) more than one lane per direction of travel, (3) signs/signals facing the driver, (4) road surface markings other than lane boundaries (e.g., crosswalks, arrows, writing, etc.), (5) lane change by this driver, (6) lane change by another vehicle, (7) turning by this driver, (8) turning by another vehicle, (9) this driver slowing to a stop, and (10) none of the above. In the second round of categorizations (C2, see Tables 1 & 2), the coding scheme was extended to include a position/direction item across all road user categories (i.e., of being parked/stationary), plus a miscellaneous item for overt video edits/alterations. Consequently, these extensions (for further data enrichment value) raised the total checkbox count per video segment to 36. The full coding scheme of annotation items (as well as the specific full training instructions given to annotators) is provided in Appendix A.

2.4. GTQ video segments: multiple purposes and representative examples

GTQ videos were selected from the full pool of video segments under the criteria to serve as effective screening and training devices. For the purpose of screening indiscriminate respondents, some of the easiest and most unambiguous scenes were selected, as for example a video segment where only an empty desert road is shown (see Example 1).

Example 1 https://www.youtube.com/embed/eS79DG08idY?start = 12;end = 15

For the purpose of explicating various annotation labels (e.g., surface paint markings, signage facing the driver), video segments were selected that contained certain items of interest, such as a segment where a railroad crossing sign appears on the side of the road as well as surface markings in the lane of travel (see Example 2)

Example 2 https://www.youtube.com/embed/vA5AiKbzIww?start = 82;end = 85

2.5. Conditions

Three different external CrowdFlower Jobs were conducted in two different rounds (C1 and C2), as shown in Table 2. In the first round, C1, a set of 48 unique three-second long video segments (randomly

Table 2
Categorization conditions.

Condition	Workers	Video segments categorized	Redundancy	Gold Test Questions	Video segments per Page	Worker payment per Page	Total CrowdFlower Cost
C1a	external	48	3	12	10	$0.50	$10.80
C1b	external	48	3	0	10	$0.50	$9.00
C1c	internal	48	1	12	n/a	n/a	n/a
C2a	external	12,862	1	53	11	$0.25	$349.32
C2c	internal	1012	1	42	11	n/a	n/a

Note. The total worker payment differs from the total CrowdFlower costs because CrowdFlower retained a margin of about 20%. Video segments per Page refers to the amount of videos the worker was assigned at a time (i.e., stacked vertically, with a scrollbar); total Pages completed varied between workers. A single Page consisted of 10 (C1) or 11 (C2) Judgments, that is, different driving video segments to be annotated.

selected from the larger full dataset of collected video footage) were categorized by external CrowdFlower workers with GTQs either turned on (C1a) or turned off (C1b). In C1a and C1b, the default triplicate redundancy setting in CrowdFlower was kept on and so the Job ran until three Judgments were collected for each video segment. Additionally, the same 48 segments were categorized offline by an individual internal worker (i.e., a confederate researcher) in C1c.

In the second round, C2, Judgments were performed on CrowdFlower across all 12,862 possible 3-s video segments of the full video dataset via external CrowdFlower workers (C2a) and over a subset of these video segments by an internal worker team comprised of multiple confederate researchers (C2c) using the same CrowdFlower structure as the external workers. Within the C2c round of internal team ratings, one team member accomplished a high volume of Judgments (n = 638) under two separate CrowdFlower accounts such that 38 different Judgments of the same driving scene segment from the same person were available to establish intra-rater reliability.

The required set of Judgments ordered for each CrowdFlower Job was specified at Job launch and included a redundancy option through a multiplier setting (x3 was used in C1, x1 was used in C2).

2.6. Analyses

In the investigation of the utility of CrowdFlower for annotating driving video content, multiple analyses from two different rounds of Jobs (Table 1) were undertaken to cover the separate but related psychometric aspects of validity (i.e., accuracy) as well as reliability (i.e., consistency).

In terms of validity, we ascertained to what extent categorizations returned from external CrowdFlower workers reflect what is actually visible in a given driving video segment. At an initial reduced Job scale, the same set of video segments was repeated with and without GTQs (Table 1, C1a vs. C1b) and compared to a reference set of categorizations of these same segments generated by a confederate researcher (C1c). For subsequent accuracy analyses at the greater Job scale (where GTQs were retained), ground truth was created by a team of internal confederates for a random subset due to the infeasibility (and self-defeating purpose) of checking the accuracy of each annotation at this scale.

In terms of reliability, we assessed how consistent categorizations of the driving video segments were when repeatedly administered. Supporting this aim, three analyses were conducted. First, from the second round of confederate categorizations (C2c) one internal team member was given a subset to categorize in duplicate to himself (i.e., randomly intermixed among his other categorizations, see 2.5 Conditions). Second, at the small scale Job (C1), each video segment was rated by three different external CrowdFlower workers (both in C1a and in C1b). Third, the full dataset categorizations of C2a provided an account of consistency due to the fact that the video segments overlapped such that any second of driving video footage was categorized three times. That is, for any second “x” bounded by start/end points [start, end] there existed a first segment: [x, x + 2], a second segment: [x − 1, x + 1], and a third segment: [x − 2, x].

2.7. Procedure

All workers were provided with a set of instructions and examples regarding the driving video segment categorization coding scheme that remained available for consultation throughout their work (Appendix A). A single Judgment consisted of a set of 31 (C1) or 36 (C2) checkboxes pertaining to features visible within a randomly selected 3-s long driving video segment (Section 2.3). A single Page consisted of 10 (C1) or 11 (C2) Judgments, that is, different driving video segments to be annotated.

In the conditions where GTQs were active (C1a, C2a, C2c), task workers were first given a single page of Quiz Mode GTQs Judgments to complete. Because of constraints of CrowdFlower, a GTQ Judgment had to be answered perfectly in order to be scored as correct, with no partial credit given (i.e., all 31 or 36 checkboxes had to be checked correctly against predetermined answers constructed by the experimenters). If workers achieved a threshold correctness Trust Score on these GTQs of 70% [i.e., 7 out of 10 Judgements] in C1, and 25% [i.e., 3 out of 11 Judgments] in C2, then workers were automatically allowed by CrowdFlower to continue through as many more Pages of Work Mode as they would like. Through trial and error, the set threshold was lowered from 70% in C1 to 25% in C2, because it turned out to be often highly difficult to obtain a perfect answer on each of the checkboxes of a Judgment. Additionally, in C2, participants were supported with further detailed feedback explaining the correct answers. For an incorrect answer to any checkbox item of a GTQ during Quiz Mode, workers were shown the correct answers of all checkboxes for that Judgment along with a brief justification. Each Page of Work Mode had one new not-yet-seen GTQ randomly presented within the other Judgments such that a worker was unable to identify which Judgments had a priori answers that their own answers would be scored against. As long as workers maintained a running average Trust Score above the set threshold (i.e., 70% in C1, 25% in C2), and there were still GTQs remaining that they had not yet seen, they were allowed to continue.

In the CrowdFlower condition without GTQs (C1b), workers were allowed to enter Work Mode straightaway without real-time screening criteria barring them from submitting Judgments. On a first-come-first-serve (optionally screened) basis, Jobs in CrowdFlower are run until a pre-determined amount of Judgments are completed by an indeterminate amount of workers.

In summary, the GTQ condition included further screening and training to enhance the responses of task workers than the condition without GTQs.

3. Results

The utility of the crowdsourcing platform CrowdFlower in the content categorization of naturalistic driving video footage was investigated through multiple analyses concerning both validity and reliability. Overall, the supposed utility of CrowdFlower in the present tasks was found to be supported (see Table 3). Results were indicative of significantly increased utility both in terms of validity and reliability in the presence of GTQs as compared to without GTQs. Results were

Table 3
Summary of analyses.

Section	Analysis aim	Relative Job size	Analysis outcome
3.1.1	Validity	Small	The GTQ condition yielded more accurate Judgments than the No GTQs condition. Accuracy was assessed by using the Judgments of a single internal confederate rater as ground truth.
3.1.2	Validity	Large	The GTQ condition yielded accurate Judgments. Accuracy was assessed by using the Judgments of a small team of internal confederate raters as ground truth.
3.2.1	Reliability	Small	A single confederate rater was found to be consistent to himself.
3.2.2	Reliability	Small	The GTQ condition yielded more consistent Judgments than the No GTQ condition, for full Judgments and at the item level.
3.2.3	Reliability	Large	The GTQ condition yielded Judgments of high inter-rater consistency for overlapping video segments. Consistency was assessed for known-to-be-accurate Judgments.
3.2.4	Reliability	Large	The GTQ condition yielded high inter-rater consistency for overlapping video segments. Consistency was assessed for unknown-to-be-accurate Judgments.

obtained both in the preliminary round of a reduced scale (C1: 48 video segments) and in the subsequent round conducted at a larger scale (C2: 12,862 video segments).

3.1. Validity

3.1.1. 48 Judgments, comparing GTQ with no GTQ

Results showed that there were 35 of 144 (24%) and 6 of 144 (4%) exact matches from C1a (with GTQs) and C1b (without GTQs) respectively, relative to C1c (taken as a measure of ground truth). Results thus indicated inaccuracies in the Judgments from both C1a and C1b (Fig. 3).

However, these inaccuracies occurred in different specificity/sensitivity biases. Phi correlation coefficients were computed between each full Judgment (i.e., an array of 31 binary checkboxes) from a condition (C1a or C1b) against the ground-truth Judgment returned by an internal confederate rater (C1c) matched for a specific video segment. The median across all 144 (48 × 3) correlation coefficients of the GTQ condition (C1a; $r = 0.78$) was significantly higher than for the No GTQ condition C1b ($r = 0.39$) (Mann-Whitney $U = 3756$, n1 = n2 = 144, $p < 0.001$ two tailed). Furthermore, greater total item accuracy across all 4464 (31 × 48 × 3) categorized items was found in C1a (4051 = 91%) than in C1b (3504 = 78%).

Among the 4464 categorized items in C1b (i.e., without GTQs), there were 396 false positives (i.e., items marked present but which were absent in the video segment according to the confederate researcher), yielding a false positive rate of 11% (396/3519). Furthermore, there were 564 misses (i.e., items marked absent that were present in the video segment according to the confederate researcher), yielding a miss rate of 60% (564/945). In C1a (with GTQs), the false positive rate was 1.6% (57/3519) and the miss rate was 38% (356/945). In other words, GTQs contributed to a reduction of both false positives and false negatives.

3.1.2. 1012 Judgments, comparing external versus internal workers

The confederate research team (C2c) performed 995 Judgments of video segments (17 video segments were removed due to video playback errors) which were randomly selected from C2a. Results showed that there were 257 (26%) exact matches between the Judgments from C2a and C2c. Phi correlations with the ground truth for both the smaller scale Job (correlation between C1a and C1c: median $r = 0.78$, see also Section 3.2.1) and the larger scale Job (correlation between C2a and C2c: median $r = 0.80$) were not found to significantly differ (Mann-Whitney $U = 65298.5$, n1 = 144, n2 = 995, $p = 0.083$).

From the 35,820 C2a items re-rated within C2c (995 Judgments x 36 items per Judgment) the false positive rate was 2.1% (682/31,564) and the miss rate was 27.6% (1176/4256).

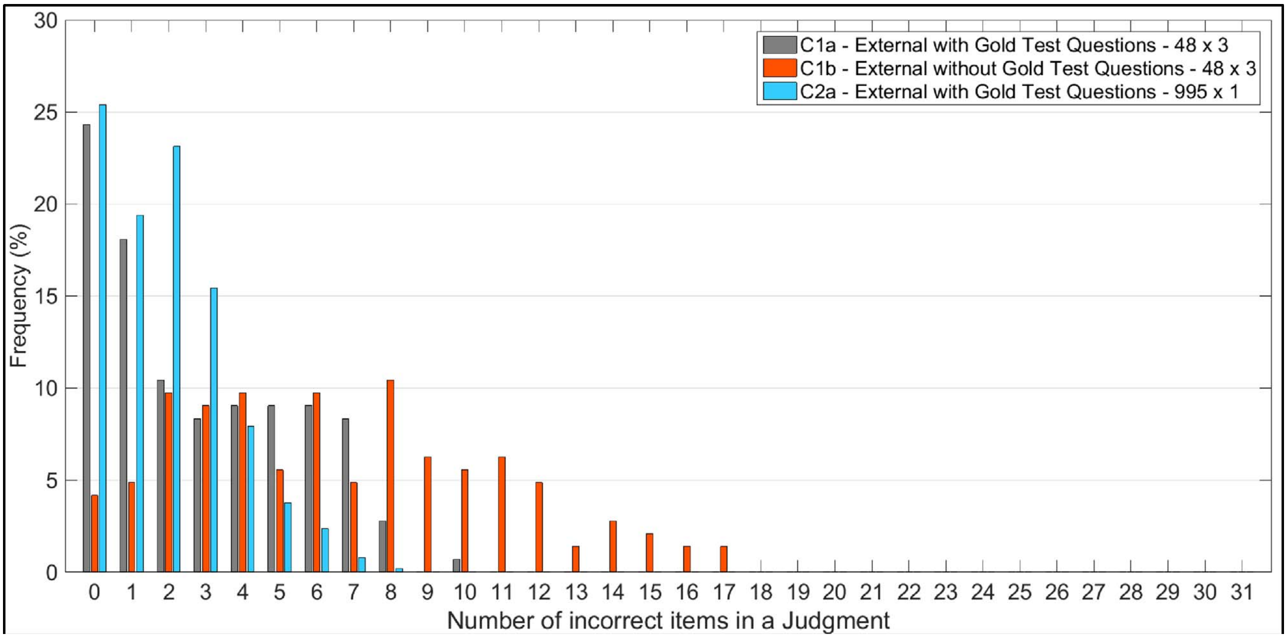

Fig. 3. Distribution of the number of errors per Judgment at the smaller C1 Job scale of 144 Judgments (with and without GTQs) and for a subset of 995 Judgments from the larger C2 Job scale (with GTQs). Errors were determined against known answers (C1c or C2c). A score of 0 signifies a perfectly correct Judgment.

3.2. Reliability

3.2.1. 38 Judgments, comparing confederate to himself

In condition C2c, one confederate performed 638 Judgments about evenly split under two different CrowdFlower accounts, with an approximate 10% subset of his Judgments from each account coded in duplicate ($n = 38$). Intra-individual test-retest reliability results for this same rater using the same software settings but across different sessions were: 34 (89%) exact matches, an average phi correlation of 0.98 across the 38 Judgments, and an overall item accuracy of 99.5% (i.e., 1361 out of 1368).

3.2.2. 48 Judgments, comparing GTQ versus no GTQ

During C1a and C1b, each video segment collected three external worker Judgments and so allowed for a consistency measure of how many categorization ratings (both for full Judgments and/or across items within Judgments) were returned identically between external CrowdFlower task workers. Unanimous agreement on all 31 items of a Judgement was found in 7 of 48 Judgments in C1a (with GTQs) and in 1 of 48 Judgments in C1b (without GTQs). Per item, the unanimous agreement percentage across the 48 Judgments was computed, and was found to be significantly higher for C1a ($M = 90\%$, $SD = 13$) than for C1b ($M = 65\%$, $SD = 19$, $n1 = n2 = 31$, $t(60) = 5.85$, $p < 0.001$).

3.2.3. 257 Judgments, comparing ratings by unanimous voting

For the correct 257 Judgments in C2 (see Section 3.1.2), a reliability analysis was conducted by comparing overlapping categorizations across sequential seconds of video footage. For example, the correct true/false answer provided for an item in a video segment that began at time x, was compared with the answer received for that same item by another external worker whose video segment began at time $x-1$ and additionally by another external worker whose video segment began at time $x-2$. It should be noted that some variation between overlapping video segments would be expected to exist (e.g., a car seen only in the last second of a segment that starts at $x = 0$ might not be visible in the previous videos $x-1$ and $x-2$). Due to such uncertainty, somewhat less than perfect reliability may be expected even from perfectly reliable raters. This necessitates consideration of proportional consistency analysis across the entire array of 36 items contained within a Judgment. In other words, it is assumed that while one or a few aspects might vary between overlapping videos, the majority of aspects should remain the same.

Results showed that 74 of 257 correct Judgments (29%) received the same true/false rating across all 36 items by three different external workers who rated overlapping video segments. Fig. 4 shows a distribution of the 257 Judgments according to the number of items yielding unanimous agreement. Judgments always had more than two-thirds (i.e., at least 25 out of 36 items) unanimous agreement, and the mean number of items yielding unanimous agreement was 33.9 out of a possible 36.

3.2.4. 12,862 Judgments, comparing ratings by unanimous voting

For all 12,862 Judgments, a reliability analysis of unanimous answers was conducted with overlapping sequential seconds again as in Section 3.2.3, but now for the full dataset. The first and last two Judgments of each video required removal due to a logical lack of full overlap, resulting in a total of 12,670 Judgments (12,862 − 4 × 48).

Regarding unanimity of full Judgments, 1129 of 12,670 answers (9%) received the same true/false value across all 36 items by the three different external workers. The mean number of items with unanimous agreement per Judgment was 32.6 out of 36 possible.

The distributions of Judgments in Fig. 4 shows that disagreement existed in the categorizations of overlapping sequential seconds of video footage; this occurred most frequently for two items.

4. Discussion, conclusions and recommendations

The CrowdFlower crowdsourcing platform may present great potential for driving research by bringing task workers from across the world to categorize a rapidly growing resource of naturalistic driving video data. Due to its inherently distributed structure, CrowdFlower and online tools of similar kind may be more susceptible to fraudulent or non-discriminating responses as compared to locally administered and more tightly controlled traditional methods. Specifically, the utility of CrowdFlower with (and without) its self-purported most important quality control mechanism of GTQs was investigated in the objective categorization of driving video contents via binary presence/absence flagging of pre-specified driving items of interest both at a preliminary reduced and a subsequently increased Job scale.

Exhibiting credible signs of validity and reliability (Table 3), the potential for the method of crowdsourcing the categorization of driving video contents can be considered in a meaningful and valuable way. For example, as a result of our settings in the present study, 12,862 CrowdFlower annotation categorizations were completed in about one

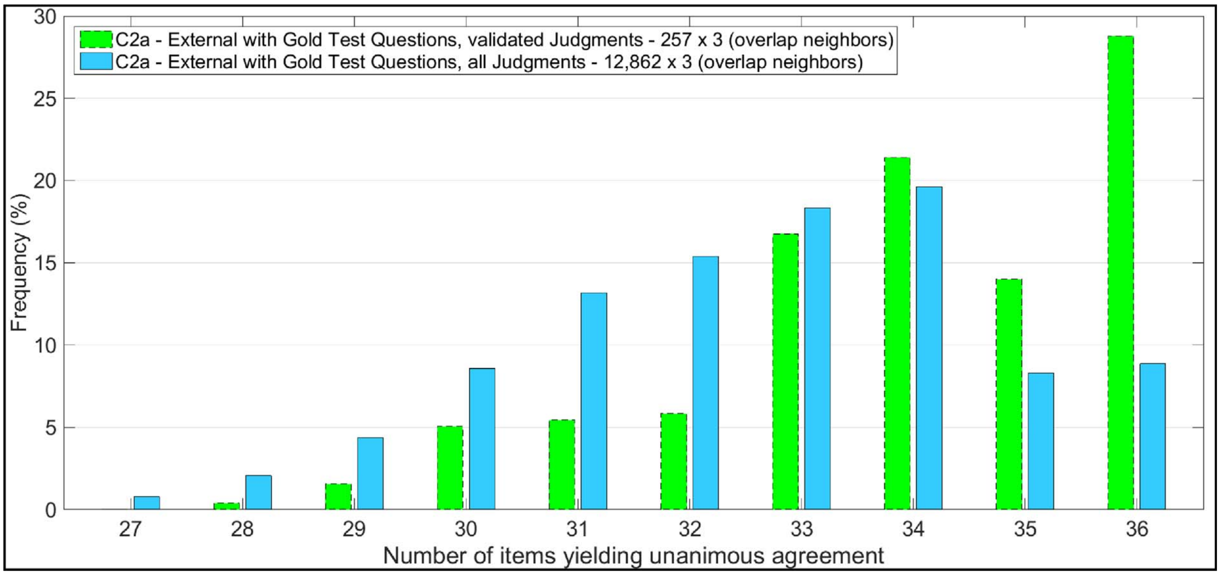

Fig. 4. Frequency of validated (i.e., 257 fully correct) and all returned Judgments (originally 12,862) from C2a according to number of items yielding unanimous agreement from three independent raters.

and a half days by 200 external workers from 46 different countries working at an hourly rate of 1.09 USD each (total cost of about 349.32 USD inclusive of a 20% transaction fee) with an average of 75 s per Judgment. Through volunteer confederate collaboration, 1002 annotation categorizations were completed in about two weeks by six internal confederate workers from the Netherlands working between/around their other work duties at a conservative estimated hourly rate around $20.25 USD each (total cost estimate of about $394.54 with an average of 70 s per Judgment). Thus, for the same approximate costs, the external workers returned categorizations about ten times faster.

Several limitations exist within the present study and are worth mentioning. The first and foremost, is that the GTQ mechanism is explicitly designed to work with objective tasks where there are clear and definable right and wrong answers and so it may not be suitable for many otherwise desirable subjective Judgments from a distributed task worker network. A GTQ is constructed in CrowdFlower to require predefined correct answers with as minimal ambiguity as possible as well as detailed and documentable justification/motivation of that answer (similar to how both annotator screening and training is used in more controlled laboratory experiments). It should be noted that the design of the present study does not lend itself towards some other research questions that might be addressed from pairing crowdsourcing to naturalistic driving data for example for purposes of investigating the general human ability in perception/annotation of various aspects of driving scenes (inter-item research questions) and/or the bearing of universal/local driving cultures on driving scene interpretation (intercultural research questions). Instead, the present study aimed to eliminate ambiguities on an equal par between conditions to test the principle manipulation of interest: the use or not of GTQs.

Nonetheless, some of our requested annotation items appear to have contributed to some confusion between some raters. The worst three annotation items, both in terms of accuracy and reliability, pertained to identification of fully straight roads, signage/signals facing the driver, and number of lanes per direction of travel. Overall, performance with these items averaged around 63% (reliability) and 79% (accuracy) compared to averages taken across all the remaining items of 93% (reliability) and 96% (accuracy). Without proper hypotheses/controls in place, we cannot propose these as particularly systematic nor meaningful results in human perception or suitability to crowdsourcing beyond our own inabilities to more thoroughly formulate such desired details for our driving video data library into more fully objective definitions/terms (see Appendix A). For example, while relative decreases in miss rates were obtained through use of GTQs, the absolute levels of miss rates (38% and 28%, in C1a and C2a respectively) might be indicative of annotation items requiring further scrutiny and/or ease in task criteria definition. Our annotation task contained a combination of both demanding visual search and items with low ground truth base rates. Thus, it would be logical or even possibly more natural for a rater to adopt a conservative strategy when faced with annotation uncertainty (i.e., not checking a box unless they have explicitly seen something). Relatedly, the high miss rates may reflect a bias due to the fact that all items were by default unchecked (absent) requiring checking as needed, rather than being checked (present) requiring unchecking as needed. Indeed, complexities in universal instructions, clear coding rule descriptions, and controlled balancing of default absence/presence question valences could be a relevant concern in crowdsourcing annotations from large, diverse, and remote participant populations without local remediation of a real-time physically present experimenter. However, it should be noted that we did not use any CrowdFlower geography/language settings and thus kept this aspect equally random across our external worker conditions so as not to confound our relative evaluations regarding potential benefits of GTQs.

Secondly, the specific items of the coding scheme created and used in the present study may be challenged further than issues of clarity towards aspects of organization and inter-item independence. The item checkboxes within a Judgment were pre-tested and arranged by probable frequencies of occurrence such that categorization speeds might benefit from predictable and likely emergent patterns of responses. Thus, the repetitive and non-random ordering of items may be a source of bias towards consistency (although, again it should be noted that the same structure was presented to both GTQ and non-GTQ condition groups).

Lastly, several dependency relations existed between items which may degrade the power of some of the analyses of the present study. For example, several items pertained to the identification of object classes (cars, motorcycles, bicycles, and pedestrians, respectively) that upon selection, each expanded with sub-item location information (i.e., leading, oncoming, passing, crossing, parking). For cases where only one object from the class was present, the sub-item location information thus became mutually exclusive rather than independent. As another example, items pertaining to actions of other vehicles such as “Lane change by another vehicle” and “Turning on/off between this and any other road by another vehicle” logically depend on presence of another vehicle and thus retain relations to ratings of item vehicle class identification.

More traditional and established methods for interrater reliability (e.g. Cohen’s/Fleiss’ kappa) were not pursued. The reason for that is the difficulty of determining a chance agreement for our Judgments that contained a composite of yes/no decisions with inter-item dependencies as described above. Instead, simpler measures of consistency, such as the phi coefficient and the proportion of unanimous Judgments, were used. Further studies with CrowdFlower more specific to questions of validity and reliability might limit such complexities in advance, sacrificing some annotation meaning in favor of stricter control, standard analyses, and afforded reflection regarding the broader annotation literature. Additionally, further assessments of the ground truth reliability of our internal rating team (beyond the single rater repetitions of the analysis in 3.2.1) would be desirable in future work. For now, the reliability agreements observed in our approach (Fig. 4) appear qualitatively consistent with levels from previous image annotation work (Nowak and Ruger, 2010; containing 53 annotations per image across a set of 99 without presuming the existence of two persons that annotated the whole set of images). Specifically, in comparison to the average identical accuracy they obtained of 0.906, following their Equation 2, we computed our own average unanimous annotation accuracies respectively as 0.941 (Section 3.2.3, Fig. 4) and 0.906 (Section 3.2.4).

Multiple ethical and privacy concerns can be raised in consideration of methods that employ crowdworkers with human annotation of naturalistic driving video data. Some of these may not be new and include attempting to anonymize video data in the sense that specific combinations of sensitive information are not presented in combination to result in personably identifiable information from both aspects of the drive (time, date, location, etc.) along with aspects of driver identity (name, face, home/work address, etc.). A major difference between the present method and the classical way of annotating naturalistic driving data is that in the present method the task is outsourced to crowdworkers who are themselves anonymous and residing in different countries, while in the classical way the annotation is done by trained team members who are typically local and known/approved by the principal investigator(s). Aside from the annotation integrity (accuracy/consistency) concerns specifically addressed in the experimental design and results of the present study, other new challenges are worth discussing such as legal requirements of the handling of data. In the present study, the video data were obtained from public sources, which is uncommon within traditional NDS approaches. Thus, any terms and conditions regarding data sharing, ownership, and viewership restrictions put in place a priori by the responsible parties would need to be considered and respected so as not to be violated. Additionally, the regulations and policies pertaining to the online reproduction/distribution of (video) data specific to each country or online hosting community should be adhered to, and this includes the presentation of potentially disturbing images such as might be the case with automobile crashes/accidents or illegal driving behavior.

A few positive privacy points regarding the present method are interesting to consider as well. Because the annotating work is distributed across many crowdworkers in distal locations, a relatively small amount of the total data is restrictively released to single/isolated persons at a time. For example, in the present study, only random 3-s clips from randomly different drives and randomly different drivers were distributed. Accordingly, it becomes much less likely that a crowdworker can come to recognize a driver's travel patterns or other aspects that may pose risks to privacy. This compares favorably in contrast to a classical annotation perspective where a single or smaller group of annotators may more likely become familiar with the travel patterns contained within the data. Additionally, the present study does not propose to share all data (e.g., geospecific, CANBUS, etc.) as may be accessible to classical annotators in naturalistic research but to selectively distribute only pieces of the full dataset (i.e., herein only video annotation was outsourced and only that of forward facing cameras from public roads where filming is allowed). Lastly, crowdworkers themselves are employed under certain terms of service to which they must accept and abide (e.g., https://www.crowdflower.com/legal/). If crowdworkers were to violate such terms (e.g., share proprietary data) they would be subject to consequences not limited to but including the likes of losing their worker privileges such as payment, membership, etc.

An increasing amount of real-life driving videos are being recorded both within naturalistic driving studies as well as from public channels of user generated content. For example, at the start of conducting the current research, there were approximately 795,000 returns for the term "dashcam" on YouTube (November 19, 2015). Upon presenting this work at the international conference for Road Safety on Five Continents (May 19, 2016), there were 1.13 million returns for the same search (i.e., +42% increase in about half a year), and by the time of manuscript revisions (August 8, 2017), a total of 4.26 million were available (i.e., +436% increase in less than 2 years' time). Categorizing such expansive data sets can be a costly and time-consuming manual process. One solution is to train automated algorithms to conduct coding tasks such as in machine learning and classification. However, such algorithms themselves often require some diligently pre-labeled examples for their own accuracy and only through diverse training sets may overcome common challenges of overfitting. Under the correct circumstances (e.g., open-access data) and quality control settings (i.e., the construction and use of GTQs), Crowdsourcing tools like CrowdFlower appear to have the potential for delivering equivalent accuracy and reliability utility as locally trained humans. It is therefore recommended that future driving research and ultimately driving safety itself might benefit from exploiting increasingly large scale and publically available data sets through embracing and channeling a growing global pool of human resources.

Acknowledgments

The research presented in this paper was conducted within the project HFAuto – Human Factors of Automated Driving (PITN-GA-2013-605817).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aap.2017.08.036.

References

Bazilinskyy, P., De Winter, J.C.F., 2015. Auditory interfaces in automated driving: an international survey. Peer J. Comput. Sci. 1, e13.

Borowsky, A., Shinar, D., Oron-Gilad, T., 2010. Age, skill: and hazard perception in driving. Accid. Anal. Prev. 42, 1240–1249.

Chapman, P., VanLoon, E., Trawley, S., Crundall, D., 2007. A comparison of drivers' eye movements in filmed and simulated dangerous driving situations. In: Behavioral Research in Road Safety, Seventeenth Seminar. London : Department for Transport.

Crundall, D., Underwood, G., Chapman, P., 1999. Driving experience and the functional field of view. Perception 28, 1075–1087.

Gosling, S., Mason, W., 2015. Internet research in psychology. Annu. Rev. Psychol. 66, 877–902.

Kyriakidis, M., Happee, R., De Winter, J., 2015. Public opinion on automated driving: results of an international questionnaire among 5, 000 respondents. Transp. Res. F: Traffic Psychol. Behav. 32, 127–140.

Le, J., Edmonds, A., Hester, V., Biewald, L., 2010. Ensuring quality in crowdsourced search relevance evaluation: the effects of training question distribution. Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010).

Nowak, S., Ruger, S., 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. Proceedings of the International Conference on Multimedia Information Retrieval, ACM 557–566.

Regan, M., Williamson, A., Grzebieta, R., Tao, L., 2012. Naturalistic driving studies: literature review and planning for the Australian naturalistic driving study. In: Australasian College of Road Safety Conference. Sydney, New South Wales, Australia.

Smucker, M., Jethani, C., 2011. The Crowd vs. the Lab: a comparison of crowd-sourced and university laboratory participant behavior. Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR 2011).

Soleymani, M., Larson, M., 2010. Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. Carvalho, V., Lease, M., Yilmaz, E. (Eds.), Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010).