CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Report No. UMTRI-2019-21 February 2019

Project Start Date: 9/1/18

Project End Date: 12/31/19

Machine Learning, Human Factors and Security Analysis for the Remote Command of Driving: An MCity Pilot

by
Walter S. Lasecki
Assistant Professor
University of Michigan

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3551747105 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Lasecki, W.S., Chung, J.J.Y., O'Keefe, S.D., Hampshire, R.C., Bao, S., & Song, J.Y. (2020). Machine Learning, Human Factors and Security Analysis for the Remote Command of Driving: An MCity Pilot. Final Report. USDOT CCAT Project No. 9
Identifier: http://hdl.handle.net/2027.42/156392

Contacts

For more information:

Walter S. Lasecki
Computer Science and Engineering
Bob and Betty Beyster Building
2260 Hayward Street,
Ann Arbor, MI 48109-2121
Phone: 734-764-4259
Email: wlasecki@umich.edu

CCAT
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48152
umtri-ccat@umich.edu
(734) 763-2498

Robert C. Hampshire
Gerald R. Ford School of Public Policy
Joan and Sanford Weill Hall
735 South State Street,
Ann Arbor, MI 48108-3091
Phone: 734-615-6975
Email: hamp@umich.edu

UMTRI · The University of Akron · Illinois Center for Transportation · Purdue University · Washtenaw Community College · Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Technical Report Documentation Page

1. Report No. UMTRI-2019-21	**2. Government Accession No.** Leave blank – not used	**3. Recipient's Catalog No.** Leave blank - not used
4. Title and Subtitle Machine Learning, Human Factors and Security Analysis for the Remote Command of Driving: An MCity Pilot Identifier:http://hdl.handle.net/2027.42/156392		**5. Report Date** December 2019 **6. Performing Organization Code** Enter any/all unique numbers assigned to the performing organization, if applicable.
7. Author(s) Walter S. Lasecki, PhD: 0000-0002-8204-1708 John J.Y. Chung, Stephanie D. O'Keefe, PhD, Robert C. Hampshire: 0000-0002-5269-3377, PhD, Shan Bao, PhD: 0000-0002-0768-5538, Jean Y. Song, PhD		**8. Performing Organization Report No.** Enter any/all unique alphanumeric report numbers assigned by the performing organization, if applicable.
9. Performing Organization Name and Address Center for Connected and Automated Transportation University of Michigan Transportation Research Institute 2901 Baxter Road Ann Arbor, MI 48109 Computer Science and Engineering Bob and Betty Beyster Building 2260 Hayward Street, Ann Arbor, MI 48109-2121		**10. Work Unit No.** **11. Contract or Grant No.** Contract No. 69A3551747105
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Report (Sept 2018 – Dec 2019) **14. Sponsoring Agency Code** OST-R
15. Supplementary Notes Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.		

UMTRI · The University of Akron · Illinois Center for Transportation · Purdue University · Washtenaw Community College · Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

16. Abstract

Both human drivers and autonomous vehicles are able to drive relatively well in frequently encountered settings, but fail in exceptional cases. These exceptional cases often arise suddenly, leaving human drivers with a few seconds at best to react—exactly the setting that people perform worst in. Autonomous systems also fail in exceptional cases, because ambiguous situations preceding crashes are not effectively captured in training datasets. This work introduces new methods for leveraging groups of people to provide on-demand assistance by coordinating responses and using collective answer distributions to generate responses to ambiguous scenarios using minimal time and effort. Unlike prior approaches, we introduce collective workflows that enable groups of people to significantly outperform any of the constituent individuals in terms of time and accuracy. First, we examine the latency and accuracy of crowd workers in a future state prediction task in visual driving scenes, and find that more than 50% of workers could provide accurate answers within one second. We found that using crowd predictions is a viable approach for determining critical future states to inform rapid decision making. Additionally, we characterize different estimation techniques that can be used to efficiently create collective answer distributions from crowd workers for visual tasks containing ambiguity. Surprisingly, we discovered that the most fine-grained and time-consuming methods were not the most accurate. Instead, having annotators choose all relevant responses they thought other annotators would select led to more accurate aggregate outcomes. This approach reduced human time required by 21.4% while maintaining the same level of accuracy as the baseline approach. These research results can inform the development of hybrid intelligence systems that accurately and rapidly address sudden and rare critical events, even when they are ambiguous or subjective.

17. Key Words	**18. Distribution Statement**		
Automated vehicles, Cross-cutting technology, Enabling Technology, Human Factors, Control and Operations, Transportation operations, Control, Detection and identification, Intelligent transportation systems, Multimodal transportation, Crowdsourcing, Instantaneous Crowdsourcing Workflows	No restrictions.		
19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	19	Leave blank – not used

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Table of contents

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

Abstract

Both human drivers and autonomous vehicles are able to drive relatively well in frequently encountered settings, but fail in exceptional cases. These exceptional cases often arise suddenly, leaving human drivers with a few seconds at best to react—exactly the setting that people perform worst in. Autonomous systems also fail in exceptional cases, because ambiguous situations preceding crashes are not effectively captured in training datasets. This work introduces new methods for leveraging groups of people to provide on-demand assistance by coordinating responses and using collective answer distributions to generate responses to ambiguous scenarios using minimal time and effort. Unlike prior approaches, we introduce collective workflows that enable groups of people to significantly outperform any of the constituent individuals in terms of time and accuracy. First, we examine the latency and accuracy of crowd workers in a future state prediction task in visual driving scenes, and find that more than 50% of workers could provide accurate answers within one second. We found that using crowd predictions is a viable approach for determining critical future states to inform rapid decision making.
Additionally, we characterize different estimation techniques that can be used to efficiently create collective answer distributions from crowd workers for visual tasks containing ambiguity. Surprisingly, we discovered that the most fine-grained and time-consuming methods were not the most accurate. Instead, having annotators choose all relevant responses they thought other annotators would select led to more accurate aggregate outcomes. This approach reduced human time required by 21.4% while maintaining the same level of accuracy as the baseline approach. These research results can inform the development of hybrid intelligence systems that accurately and rapidly address sudden and rare critical events, even when they are ambiguous or subjective.

Introduction

Both human drivers and autonomous vehicles are able to drive relatively well in frequently encountered settings, but fail in exceptional cases. These exceptional cases often arise suddenly, leaving human drivers with a few seconds at best to react—exactly the setting that people perform worst in. Autonomous systems also fail in exceptional cases, because ambiguous situations preceding crashes are not effectively captured in training datasets.

While remote driving itself is not a new concept, what we need is a super-human driver for sudden, critical events. We can achieve this by using input from multiple individuals, coordinated around a task at a moment's notice. Our approach builds on real-time crowdsourcing workflows—"human algorithms"—that coordinate groups to outperform individuals in terms of both reliability and task performance (e.g., latency) (de la Cruz, Peng, Lasecki, & Taylor, 2015). With the rise of artificial intelligence as a service, human-backed algorithms at scale have become the norm rather than the exception for intelligent systems. Google, Facebook, Apple (Siri), Samsung, Bloomberg, and countless others use large groups of human annotators and checkers behind their intelligent services.

However, while reliability and accuracy is important in all of these settings, none of these prior methods have used low-latency, real-time systems. Lasecki's previous work establishes the feasibility of this approach in low latency environments (de la Cruz, Peng, Lasecki, & Taylor, 2015; Bigham, Lasecki, & Wobbrock, 2015; Lundgard, Yang, Foster, & Lasecki, 2018). Work on combining Reinforcement Learning with crowd feedback found that we can get collective responses in <0.3s (roughly human reaction time, but more stable and less subject to distraction). We also have worked on more formally modeling collective input mediation strategies that can optimize for either input reliability or low latency (Bigham, Lasecki, & Wobbrock, 2015). Our approach asks groups of remote commanders with minimal direct interaction (to reduce interaction cost and allow for on-demand routing to tasks) to concurrently help with a given monitoring or control task within as little as ~350ms of a need arising. With video-based remote control latencies as low as 100ms, total latency for control can be under 0.5 seconds. In contrast to our prior work that introduced these approaches (Bigham, Lasecki, & Wobbrock, 2015), which learned how to effectively interleave and combine input from groups over short time spans, we modified these approaches so they work effectively for short, sudden bursts.

To improve the speed of responses, we introduce methods that directly leverage the AV's ability to understand possible futures that may arise in real settings (even when the system does not know how to respond to a possible setting) to pre-fetch possible configurations of the world. Using these future states, crowd workers can (in parallel) provide feedback before a system needs to know what action to take. What makes this possible is the speed of existing real-time crowdsourcing approaches. While 0.5s may be a relatively slow response time for an engaged driver to respond to an event (usually 200-300ms), the ability to respond this quickly means that we only need to pre-fetch future states of the world that are <1second into the future, which is a tractable task.

Prior work has shown that just-in-time (JIT) training can result in an average response time of below 3.5ms, a three-order-of-magnitude reduction in latency (Lundgard, Yang, Foster, & Lasecki, 2018). Further, the collective response is more likely to be correct than a single person's (i.e., driver's) response

(Bigham, Lasecki, & Wobbrock, 2015; Lasecki, 2011). The remaining challenge is scaling solutions that work in "laboratory" settings (simple, fully-controlled problems) to real-world settings with massive state spaces.

In order to accomplish JIT feedback in real-world scenarios, which often contain uncertain events or ambiguous contexts, we explored methods for human-guided selection when rapid intervention is required. We also investigated methods for eliciting and aggregating reliable annotations and human feedback on ambiguous times with less human time required than current state-of-the-art approaches.

Findings

Our findings are grouped by two themes. First, we introduce a scalable workflow that enables instantaneous crowdsourcing in dangerous driving scenarios. This work demonstrates that it is viable to invoke instantaneous crowdsourcing methods in a scalable and efficient way. Second, we investigate how to improve data collection methods for training ML models on ambiguous and subjective data. We demonstrate that it is possible to more efficiently generate data distributions for ambiguous visual data. We present the methods and results of these two lines of research below.

Developing an Instantaneous Predictive Crowdsourcing Workflow to Prevent Car Crashes

In this section, we present three studies on the expected benefits and challenges of implementing an instantaneous crowdsourcing approach in autonomous driving settings. In Study 1, we conduct an experiment to understand how people perform when completing predictive annotation tasks. In Study 2, we ran a study to understand how input method influences people's answer latency and accuracy in predictive annotation tasks. Finally, we present formative research on combining computer vision algorithms and human prediction to effectively leverage the complementary abilities of systems and humans in a Hybrid Intelligence approach. Below we present the methods and results of each study.

Study 1 - False positives are prevalent when using predictive crowdsourcing to anticipate collisions in driving scenarios.

In order to increase the speed of human operator decision making, we introduce a new crowdsourcing workflow. When an autonomous system encounters an unfamiliar dangerous situation where it cannot determine which action to take, control is handed off to an instantaneous crowdsourcing workflow that uses collective human ability to rapidly predict which objects in a driving scenario will be dangerous in the near future. However, there is not much known about how people make predictions in instantaneous crowdsourcing settings. We conducted this study to increase our understanding of how well people perform in predictive annotation tasks. In this study, crowd workers watched simulated driving scenarios and were asked to predict whether or not there was going to be a collision. The long-term goal of this workflow is to use pre-fetched information about objects in driving scenes to rapidly and accurately detect when a collision is likely to occur.

Participants were crowd workers recruited from Amazon's Mechanical Turk who were shown simulated driving videos either with or without a car crash. In order to prevent bias, in the tutorial crowd workers were asked to complete before beginning the annotation task they were shown two videos: one with and one without a collision.

Videos with collisions were varied by the following characteristics:

1. **Pre-appearance time**: This is the time interval between the start of the video and the first frame in which a vehicle collision appears. The interval was either 0 seconds, 3.8 seconds, or 7.6 seconds.
2. **Pre-indication time**: This is the time interval between the first frame in which the vehicle collision appears and any clear visual indicator of danger in the scene. This interval was either 0 seconds, 3.8 seconds, or 7.6 seconds.
3. **Indication time**: This is the time interval between the vehicle that is about to get in a crash maneuvering in a way that clearly indicates danger and the collision actually occurring. In other words, this represents how fast the crash happens or how quickly the vehicle is moving. This interval was either 0.6, 1.2, 2.0, 2.7, 3.3, or 4.0 seconds.
4. **Number of Vehicles in the Scene**: The number of vehicles shown in the scene varied from 1 vehicle to 3 vehicles.

Since we are most interested in understanding latency among human evaluators who are asked to predict when a car crash will occur, we varied three different time intervals. In the pre-indication time and indication time intervals, we defined physical indication of danger as the moment when the vehicle with the camera and another vehicle in the scene are expected to collide if they maintained their current physical dynamics. We also manipulated scene complexity by varying the number of vehicles shown on the road. In the simulated driving videos, the indication of danger was when another vehicle cuts into the lane of the target vehicle with the camera. This remained constant across all study conditions.

To measure how well people were able to predict collisions, three metrics were used to assess the collected video annotations. First, we measured **Prediction Time**, the time between the participant making their first correct prediction and the crash occurring. Next, we calculated the **Prediction Rate**, how many participants were able to correctly predict the crash after the beginning of the indication time interval. In our Prediction Rate analyses, we only included accurate predictions that were made after there was a clear visual indication that a crash was about to occur. We excluded these annotations because predictions made without a corresponding visual cue could be random or unrelated to what is happening in the scene, meaning future system predictions that are based on this prefetched data will be less precise. This also led us to measure the **False Positive Rate**, how many participants predicted there would be a crash before any visual indication of danger occurred.

We used linear regression to assess which of the four manipulated variables were associated with Prediction Time, Prediction Rate, and the False Positive Rate. For videos with a collision, 78.7% of participants accurately predicted the crash would occur. We found a negative correlation between the Pre-Appearance Time Interval and the Prediction Rate (coeff = -.018, $p < 0.001$), such that the longer the

Pre-Appearance Time Interval was, the less accurate people were at identifying the crash. Similarly, there was a negative correlation between the Pre-Indication Time Interval and the Prediction Rate (coeff = -0.029, $p < 0.001$), such that the longer the Pre-Indication Time Interval was, the less accurate people were at identifying the crash. Though these findings seem counterintuitive, the correlations may be explained by the prevalence of false positives. In other words, with a longer time interval before there is a clear visual indicator of a collision, people may make overly cautious predictions that are not grounded in visual cues. When including false positive predictions in the prediction rate metric, we did not find a significant correlation between the indication time interval and the prediction rate. In videos where a crash did not occur, the false positive rate was 30%. The pre-appearance time interval was positively correlated with the false positive rate (coeff = 0.025, $p < 0.05$; $R^2 = 0.548$), with longer pre-appearance time intervals being associated with a higher likelihood of a false positive prediction.

When examining Prediction Time data, we found that 20.7% of participants indicated that a crash would occur before the Indication Time Interval began. Since premature predictions are considered false positives in this work, these annotations were removed in the next set of analyses. We found when removing false positives from the data, Prediction Time was positively correlated with the Indication Time Interval (coeff = 0.47, $p < 0.001$), with participants taking more time to make a correct annotation when the Indication Time Interval was longer. We also found a negative correlation between Prediction Time and the Number of Vehicles in the Scene (coeff = -0.078, $p < 0.001$; $R^2 = 0.531$), meaning when more vehicles were shown in the scene, participants made accurate predictions faster. Across all videos, we found that on average participants made their first prediction 1.49 seconds ($\sigma = 0.76$) after the Indication Time Interval began. The best performing participants in our sample, who annotated the data quickly and accurately, made their first prediction within 0.90 seconds ($\sigma = 0.42$) of the Indication Time Interval starting time.

We conducted an additional analysis to see when participants make their first premature prediction, for videos with and without collisions. We find the majority of participants who made premature predictions made their annotation as soon as another vehicle appeared in the scene for the first time.

Study 2 - Annotation input method influences latency of crowd workers predicting if a dangerous object is present in a driving scene.

In our next study, we explored the influence of different input methods on the latency of crowd worker responses when they are asked to predict whether or not there is a dangerous object in a driving scenario. Dangerous objects were defined as any object that could cause a crash by moving and colliding into our target vehicle. In other words, dynamic objects, like people or other vehicles on the road, were considered dangerous while static objects, like trees or fire hydrants, were not.

We assessed human reaction time and accuracy in three input conditions:

1) **1H2K** - one hand and two keys (left and right arrow keys)

2) **2H2K** - two hands and two keys (left and right shift keys)

3) **1H1K** - one hand and one key on the keyboard (spacebar)

When two keys were used for input, participants selected whether or not the object was “movable” or “won’t move” by pressing the key that was assigned to represent each category. When one key was used for input, pressing the key indicated that the object was “movable” or dangerous and not pressing the key indicated that the object was not a threat and “won’t move.”

We recruited 60 crowd workers, with 20 crowd workers randomly assigned to each of the three input conditions. The task began with a two-part training period. In part one, participants were oriented to the input method and completed 20 trials in which they had 0.8 seconds to indicate whether or not an object was “movable” or “won’t move.” In part two of the training, participants completed two practice trials in which they watched a video from the driver’s perspective and were asked to assess whether or not there was a potentially dangerous object in the scene. In the main study task, participants made these ratings for 10 different videos. Half of the videos contained a dangerous dynamic object and the other half did not. The order in which videos were presented was randomized to reduce response bias.

We found that across all conditions participants were able to quickly indicate whether or not there was a dangerous dynamic object in the scene, with over half of them providing a response within 1.06 seconds. There was a significant difference in response time between conditions (Kruskal-Wallis, Chi−Square = 8.80, $p < 0.05$), with quicker responses in the one hand and one key condition compared to the two hands and two keys condition (Mann-Whitney $U = 7042$, $p < 0.05/3$ with Bonferroni correction). The three input conditions all had high similarly high accuracy, precision, and recall scores equal or greater than 0.87 (Cochran’s Q test, $N = 200$, $Q = 3.59$, $p > 0.1$).

Additionally, we examined the accuracy, precision, and recall throughout the duration of the task for individual workers and teams of four workers. For these analyses, the agreement threshold was 0.5, meaning that at least half of the workers needed to provide the same answer. With teams, each worker’s answers were accounted for as soon as they were provided, which meant a 50% agreement rate could be reached before all four of the workers provided an answer. For both of the two keys conditions, reliable performance could be achieved in just one second. The one hand one key condition has the most robust recall scores, but precision and accuracy with this input method dropped when examining team performance as opposed to individual performance.

Study 3 - Superpixel algorithms can accurately pre-segment images with many superpixels in a hybrid intelligence instantaneous image segmentation workflow, but perform poorly on images with few superpixels.

Next, our research addressed how to handle cases when machines fail to segment, or recognize the boundary of an object. Precisely segmenting the object is a crucial aspect in understanding the environment as it allows vehicles to decide how it should avoid objects in the environment. Also, as we are looking at the autonomous vehicle scenario, we should recover this failure within a short amount of time. To enable this solution, we proposed the approach of combining human computation and vision algorithms that do not require any data to accelerate reliable segmentation. With vision algorithms such as

superpixel generation algorithms, we can pre-segment object images with pixel information and hire a large group of crowd workers to decide whether each pre-segmentation belongs to a target object or not. With the initial idea, we conducted a small-scale experiment that investigates whether this direction is plausible. We found that superpixel algorithms can pre-segment objects accurately when the number of superpixels is high, but when the number is low, pre-segment tends to include an irrelevant part of the image. Hence, there is a trade-off between the number of pre-segments and the accuracy of pre-segment. Devising a workflow that can overcome this trade-off will be our future work.

Developing Efficient Elicitation Approaches and Estimation Techniques to Create Collective Data Distributions for Ambiguous Scenarios

Study 4 - Selecting all relevant labels from the perspective of others is the most effective annotation approach for efficiently estimating data distributions.

We published a paper that explores the best way to elicit annotations from crowd workers regarding ambiguous or uncertain situations (Chung, 2019). For instance, due to limited contextual information and differing perspectives, crowd workers can disagree on the best label for describing ambiguous data. In such scenarios, we propose that receiving annotation labels in the form of an answer distribution (instead of a single best answer) will be more informative for autonomous vehicle systems when encountering unfamiliar ambiguous environments and scenarios.

The goal of this study was to improve how ambiguous information and scenarios are understood by using crowdsourcing to generate answer distributions at scale. More specifically, we explored methods for efficiently eliciting rich responses from annotators that enable accurate estimation of an answer distribution. In ambiguous or subjective domains, multiple valid interpretations may exist, especially if there is insufficient contextual information available to annotators. In these cases, answer distributions can be a better representation of the data than a single answer, and capture a more nuanced representation of an ambiguous scenario. Existing approaches for generating answer distributions ask individuals to provide a single rating for queries about a target object or scene. While this is a standard annotation method, using this approach to estimate an answer distribution rather than a single best answer requires more data and more annotator time and effort, to yield a valid and robust data distribution. Our research examines how different querying strategies that vary in annotation granularity and perspective taking prompting can potentially reduce the time and effort needed to collect answer distribution data while maintaining or exceeding the quality achieved using traditional methods.

We conducted a 4 x 2 experiment in which we asked crowd workers to annotate the emotional valence of facial expression images using varying degrees of annotation granularity and from different perspectives. We varied the annotation granularity by asking them to provide either the 1) single best label, 2) all

relevant labels, 3) rankings of all relevant labels, or 4) real-valued probabilities for all relevant labels. We also varied the task with two perspective taking prompts, in which crowd workers were asked to provide their answer(s) from their own perspective or they were asked to provide the answer(s) they thought a group of crowd workers who completed this same task provided.

To test the effectiveness of the eight proposed elicitation methods, we conducted an experiment on Mechanical Turk. In this experiment, we asked crowd workers to annotate images of faces by providing ratings of how positive or negative the person's emotional expression appeared. We chose to use an emotion annotation task in this study because facial expressions can be ambiguous, especially with limited situational context, and raters may have differing interpretations. In this study, we used images from the FACES dataset which contains 2,052 facial expressions from 179 people. The facial expression images convey one of six emotions: happiness, neutrality, anger, fear, sadness, or disgust. Since these images are already labeled, this allowed us to validate our findings and to create baseline data distributions by using images that vary in ambiguity. An emotion annotation task is also ideal to establish which elicitation methods are most effective for generating answer distributions at scale before exploring complex domains with multiple dynamic target objects, like autonomous driving scenarios.

To create the gold standard data distributions, we collected 50 single label annotations for four different images from the FACES dataset and used bootstrapping to measure stability. The four images selected for creating the gold standard answer distributions had perfect agreement on the selected emotion label in the FACES annotation dataset. Using the data labels and agreement scores from the FACES dataset, we also selected the 40 most ambiguous images for our study. We collected baseline data for the ambiguous images from 500 crowd workers. We tested the stability of the gold standard distributions by randomly resampling the gold standard distributions with replacement 10,000 times and computed the Wasserstein distance between the first distribution and the resampled ones. The average distance between the gold standard and resampled distributions was 0.11 ($\sigma = 0.07$), which indicates the gold standard distribution is stable.

After creating the gold standard distributions, we collected 30 annotations for the 40 ambiguous images in each of the 8 elicitation conditions -- a total of 9,600 annotations. We used a between-subjects design, in which crowd workers annotated five images in one of the eight elicitation conditions. While some annotations were lost due to technical issues, we continued recruiting participants until we reached 30 annotations per each condition. At the start of the task, crowd workers were given instructions specific to the condition they were assigned. As an attention check, after being shown the instructions, crowd workers were asked to complete a short survey to ensure that they understood the task details. After this stage, crowd workers began annotating the emotional valence of five images of facial expressions. In the single label conditions, participants made their selections using radio buttons. In the multiple label selection conditions, participants made their selections using checkboxes. In the rank ordering conditions, participants made their initial selections using checkboxes, and above each selected label, participants indicated their rank ordering. Finally, in the probability conditions, participants were given 100 tokens and were asked to assign these tokens to the provided annotation labels to indicate how likely each label is applicable to the displayed image.

The results of our experiment indicate that the most fine-grained questions tended not to result in the highest data quality. First, we found that for a single annotation, asking a crowd worker to select multiple applicable labels from the perspective of other crowd workers who ostensibly completed the same task outperformed all other approaches. In other words, this elicitation method resulted in the smallest Wasserstein Distance between the estimated answer distributions and the gold standard answer distributions created using traditional data collection methods.

We also found that when comparing different sizes of annotation sets (sample sizes of 2, 3, 10, 15, and 30 annotations), for all group sizes there was a significant main effect of annotation granularity, such that selecting multiple applicable labels yielded the smallest Wasserstein distance. However, when examining the effect of estimation perspective (self or others), there was only a statistically significant performance boost for sets of two annotations. As the annotation set size increased, we did not detect significant differences between the annotation method on the Wasserstein distance. Overall, when comparing the different answer elicitation methods, asking people to select all applicable labels from the perspective of other crowd workers was most effective for smaller sets of annotations.

Next, we examined which approach would result in the smallest Wasserstein distance if we held time spent on task constant across conditions. First, we calculated the median amount of time it took participants to complete the annotation task by condition. We found that as the elicitation condition became more fine-grained, the time crowd workers spent on the task increased. However, we wanted to know what would happen if we set a time budget, and assessed how many annotations could be collected in a particular timeframe and what the quality of these annotations would be (in terms of Wasserstein difference between the estimated answer distribution and the gold standard answer distribution). We set the time budget to the amount of time it takes to collect 4 single annotations (median task time = 1 second per annotation, which means the budget for 4 single label annotations is 4 seconds) and calculated how many annotations in the other elicitation approaches could be collected in this timeframe. We then computed the corresponding Wasserstein distance for each condition. We found that the multiple labels conditions were the only conditions with a significantly lower Wasserstein distance compared to the baseline single best label conditions. The multiple labels conditions also had significantly lower Wasserstein distances when compared to all other elicitation approaches. We did not find a significant effect of estimation perspective.

To follow up on these analyses, we assessed how many annotations need to be collected in the multiple labels conditions in order to achieve the same performance. The goal is to understand how efficient the multiple labels conditions are in comparison to the baseline approach. We set the baseline number to 10 annotations collected in the single label condition and compared this to the multiple label conditions ranging from 1-10 annotations. This allows us to see where in this range of number of annotations collected the multiple label conditions meet or exceed the single label condition in quality (i.e., Wasserstein distance). For multiple labels selected from the first-person perspective, 8 annotations were needed to achieve a lower Wasserstein distance than 10 baseline annotations, which required 20% fewer workers and 3.2% less human time. Multiple labels from the perspective of other crowd workers performed even better, with only 6 annotations needed to achieve a lower Wasserstein distance than the

single label condition, which required 40% fewer workers and 21.4% less human time. These results provide evidence that selecting multiple labels from the perspective of other crowd workers is the most efficient approach compared to all other elicitation approaches included in this research.

We also assessed whether or not data ambiguity would influence the effectiveness of our elicitation approaches. To do this, we tested if there was a correlation between the Wasserstein distance of the elicitation approach and the baseline and the data ambiguity. We measured data ambiguity by computing Gini coefficients for gold standard annotation distributions. In our study, a higher Gini coefficient indicated that the answer distribution is skewed toward one label, meaning the data was not ambiguous and the answers were consistent across annotators. A low Gini coefficient indicated that the answer distribution is more evenly distributed across labels, meaning the data were ambiguous and answers varied across annotators. We conducted linear regressions and found statistically significant negative correlations for all conditions. These results indicated that as the Gini coefficient gets larger (meaning that the data is less ambiguous and annotations are skewed toward a single label), the Wasserstein distance between the elicitation approach and the gold standard decreased. In other words, when the data were not ambiguous, all elicitation approaches yielded answer distributions that were very similar to the gold standard distributions.

Across multiple analyses, we found that selecting multiple relevant labels from the perspective of other crowd workers was more efficient and more accurate than the finer grained annotation approaches we tested. This prompted us to run additional analyses to understand why the more fine- grained approaches did not yield the most accurate annotation distributions as we expected. We found that when comparing the Gini coefficients of answer distributions from the real-valued probability weights condition with the gold standard distributions, the Gini coefficients were skewed irrespective of the gold standard Gini coefficient. This suggests that in fine-grained elicitation approaches, ratings were restricted to a smaller number of labels compared to other elicitation approaches.

Recommendations

Collective workflows need to account for false positives and slower reaction times in crowdsourcing environments where there is a high likelihood of risk aversion, such as in driving environments in which people need to decide whether or not there is the potential for danger. Under time pressure in a rapid annotation task, people are likely to engage in heuristic-based decision making. In the context of driving and being asked to focus on and flag any object that could potentially cause a car crash, it is plausible that people are overly cautious in their predictions in order to avoid any risk of a collision. People may think by default it is better to proactively flag objects as dangerous to avoiding any chance of a collision, even if the annotation turns out to be unnecessary and inaccurate as time progresses. However, the fact that overly cautious autonomous systems suffer from erratic and situationally inappropriate vehicle maneuvers may not be common knowledge or an intuitive correction for human annotators to make. In future predictive crowdsourcing approaches, accuracy could be emphasized in the instructions or

rewarded in order to redirect human annotator's motivation and reduce the incentive for prematurely indicating danger in a driving scene.

Annotation input methods can also play a role in facilitating how quickly people can provide an answer. In the least complex input method, in which people press the spacebar to indicate that there is a dangerous object in the scene and do not need to take any action when there isn't any danger present, people are able to provide answers quickly and accurately. However, we found that accuracy and precision were not maintained when aggregating the data in groups of four crowd workers, rather than considering data individually for each crowd worker. Future work in this space should consider the trade-offs of speed and accuracy with different annotation input methods. Considering the rate of false positives and input methods could also improve annotation methods. Making it easier for people to provide input rapidly could provide them with more time and a greater mental capacity to focus on providing accurate answers, or it could enable focusing on providing answers rapidly without needing to focus on providing a deliberate answer. For instance, social cognition research uses reaction time tasks as a method for detecting heurtistics, biases, and stereotypes, by facilitating quick reaction times in settings with high time pressure (Nosek & Banaji, 2001; Greenwald, McGhee, & Schwartz, 1998). Future work could explore how different input methods for rapid can facilitate overcoming heuristic based decision making while leveraging collective task distribution workflows that keep prediction latency low.

Additionally, we conducted work exploring how effective it is to use superpixel computer vision algorithms to pre-segment objects in driving scenes and had groups of crowd workers evaluate the segmented images outputted by the system. Though the superpixel algorithm accurately segmented images with a high number of superpixels, it performed poorly with low numbers of superpixels and included irrelevant parts of the image. Similar to the previous two studies we conducted, this work also relies on high accuracy output. Our future work in this space will devise a workflow that can overcome the trade-off between the number of pre-segments and the accuracy of pre-segments in hybrid intelligence systems that combine the efforts of people and computer vision algorithms.

We also propose that annotation answer elicitation methods is one potential way to generate all possible future states in driving scenarios and car crash prediction. Future work could explore the effectiveness of peer prediction and aggregated answer distributions in predictive crowdsourcing workflows. While we find that people struggle with overly cautious predictions when asked to identify dangerous objects in driving scenarios, future work could explore a new phrasing of the query which asks crowd workers to identify which object the driver thinks is the most dangerous. Though we find that asking crowd workers to select all relevant labels allows us to estimate an accurate data distribution at scale, in a complex driving scenario where responses are needed rapidly, this approach may not be the best situation fit since it could encourage false positive answers. Future work is needed to understand how these annotation elicitation methods can be effectively leveraged in ambiguous driving scenarios where both time and accuracy need to be prioritized.

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Outputs, Outcomes, and Impacts

Synopsis of performance indicators Part I and II (attached)

- We ran a study that investigated the human prediction capability in predicting objects that would cause a car crash.
- Conducted research and wrote a manuscript describing the remote driving approach, research roadmap, and nationwide workforce estimates of the number of required remote drivers.
- Submitted a workshop paper to the Association for Computing Machinery (ACM) Conference on Human Factors in Computing Systems (CHI), which was accepted and presented in May 2019.
- We submitted a poster paper to the ACM Symposium on User Interface Software and Technology (UIST), which was accepted and presented at the conference in October 2019. The paper investigated how quickly people can offer inputs regarding recognizing the possible danger of an object.
- We categorized rare dangerous scenarios that autonomous vehicles experience so that we can understand and address the most significant problem with the current state of the art systems. We characterized that autonomous vehicles not being able to recognize new or rare objects as a significant problem, because the system cannot make the optimal maneuvering decision for the object (e.g., trying to avoid a plastic bag on the road).
- We conducted a formative experiment to understand how a hybrid workflow with computer vision algorithms and crowd workers could help autonomous vehicles understand rare objects in the environment. We identified a challenge with this approach, which is that superpixel algorithms do not perform accurately on low pixel data.
- We submitted a research paper to the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), which was accepted and presented at the conference in November 2019. The paper accepted examined peer prediction in annotation settings, which would be related to how to coordinate people to participate in a real-time task of handling exceptional cases that autonomous vehicles confront.

Outputs

Publications, conference papers, or presentations (from major conference or similar event)

John Joon Young Chung, Fuhu Xiao, Nicholas Recker, Kammeran Barnes, Nikola Banovic, Walter S. Lasecki. May 2019. Accident Prevention with Predictive Instantaneous Crowdsourcing. Workshop on Looking into the Future: Weaving the Threads of Vehicle Automation at CHI 2019 https://johnr0.github.io/assets/publications/CHI2019-Workshop-PreCog.pdf

John Joon Young Chung, Fuhu Xiao, Nikola Banovic, Walter S. Lasecki. October 2019. Towards

Instantaneous Recovery From Autonomous System Failures via Predictive Crowdsourcing. In Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2019). New Orleans, LA. https://doi.org/10.1145/3332167.3357100

John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo (Ray) Hong, Juho Kim, and Walter S. Lasecki. November 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 62 (November 2019), 25 pages. **Best Paper Honorable Mention**
https://doi.org/10.1145/3359164

Other outputs

Press Coverage

- TechXplore, February 2019: https://techxplore.com/news/2019-02-air-traffic-driverless-cars-deployment.html
- CMU, Mobility 21, February 2019: http://mobility21.cmu.edu/air-traffic-control-for-driverless-cars-could-speed-up-deployment/
- Mlive, February 2019: https://www.mlive.com/news/ann-arbor/2019/02/self-driving-cars-could-deploy-sooner-with-instantaneous-crowdsourcing-um-researchers-say.html
- University of Michigan, News, February 2019: https://news.umich.edu/air-traffic-control-for-driverless-cars-could-speed-up-deployment/
- University of Michigan, Twitter, April 2019: https://twitter.com/umsi/status/1117804521206571009
- Stateside, Michigan Radio, April 2019: https://www.michiganradio.org/post/stateside-mi-iraqis-face-looming-deportation-theater-talk-making-self-driving-cars-safer#cars

List and electronic copies of outcomes from the project.

- Our work increases the understanding and awareness of transportation issues in the domain of autonomous vehicles.
- This research increases the body of knowledge on how hybrid intelligent systems that leverage complementary abilities of humans and AI agents can be used to address current weaknesses in autonomous vehicle systems.
- We introduce a new instantaneous crowdsourcing workflow which has the potential to reduce latency in systems identifying and reacting to ambiguous and unfamiliar objects, which can potentially be dangerous, by calling on the crowd to rapidly predict which objects in a scene are likely to be dangerous.
- We identified challenges with false positives and risk aversive judgments among human annotators evaluating objects in driving scenes, which can inform future approaches in this domain.

- Our research introduces new annotation elicitation methods for creating data distributions at scale, which has the potential to revolutionize training ML systems to better understand ambiguous and subjective data.

List of impacts from your project.

- The results of this work improves the operation and safety of the transportation system.
- The work we have done contributes to the body of scientific knowledge by uncovering new challenges when using look ahead approaches in crowdsourcing in a more complex and realistic autonomous vehicle context. We find that false positives and premature prediction of danger are not uncommon and need to be accounted for in instantaneous crowdsourcing workflows. We also find that dealing with a large number of potential futures with many variable objects is another challenge for instantaneous crowdsourcing prediction workflows.
- The characterization of the failures of autonomous vehicles contributes to generalizable knowledge on how we can deal with the failures of other autonomous systems (e.g., a robot in physical environments). Our published paper has a broad impact that not only informs other researchers and crowd-system builders how to leverage peer prediction in the system, but also has the potential of guiding better coordination in the real-time recovery from failures of autonomous vehicles by leveraging answer and annotation distributions.

References

Bigham, J. P., Lasecki, W. S., & Wobbrock, J. (2015). Target Acquisition and the Crowd Actor. *Human Computation, 1*, 101-131.

Chung, J. J. (2019). Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing*, (pp. 1-25). Austin, TX.

Gabriel Jr, V., Peng, B., Lasecki, W. S., & Taylor, M. E. (2015). Generating Real-Time Crowd Advice to Improve Reinforcement Learning Agents. *Association for the Advancement of Artificial Intelligence.* Austin, TX.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. . *Journal of personality and social psychology, 74*(6), 1464.

Lasecki, W. S. (2011). Real-time crowd control of existing interfaces. *In Proceedings of the 24th annual ACM symposium on User interface software and technology*, (pp. 23-32). Santa Barbara, CA.

Lundgard, A., Yang, Y., Foster, M. L., & Lasecki, W. S. (2018). Bolt: Instantaneous crowdsourcing via just-in-time training. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* , (pp. 1-7). Montréal, Canada.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social cognition, 19*(6), 625-666.