Matthew Soza, Qingyuan Lu

Proficiency of Machine-Vision Systems in Identifying Road Signs in Non-Optimal Conditions

Introduction

In the past decade and a half, the abilities of machine-vision systems have improved exponentially. In 2010, winners of the ImageNet Large Scale Visual Recognition Challenge had an error rate of 28.2%. Only four years later, in 2014, that number was reduced to 6.7%. (Russakovsky et al., 2015). These advances in machine-vision have enabled an equally exponential amount of use-cases for machine-vision, one of the most radical and sensational in recent times being autonomous vehicles that are intended to be used on public roads and freeways.

An autonomous vehicle is defined as an "automobile that employs driver assistance technologies to remove the need for a human operator"(*Autonomous Vehicle | Definition, History, & Facts | Britannica*, 2023). There are six generally accepted stages of vehicle autonomy that range from 0-5, where 0 is no autonomy, and 5 is fully autonomous. Though no vehicles currently operate at stage 5, there are vehicles that operate at stage 4, which is full autonomy but only in certain geofenced regions where the vehicle is permitted to operate (Anderson, 2020). This is the case with autonomous taxi companies such as Waymo or Uber. Even more common are stage 3 vehicles, which are theoretically capable of operating in most conditions a human driver would, but require the attention of the driver at all times to ensure the vehicle is operating correctly, as not all bugs have been removed from the systems.

These autonomous vehicles generally use some combinations of LIDAR, radar, GIS/GPS, and machine-vision systems to navigate a space and to identify people, animals, other

vehicles, and, very importantly, traffic signs (Campbell et al., 2018). Traffic signs are typically identified in a two-stage process . The first stage is identifying that a road sign exists. Once a road sign has been identified, a bounding box is drawn around it, and a cropped image of the sign is made and fed to stage two. The second stage takes the cropped sign image then identifies and categorizes the sign, whether it be a stop sign, yield sign, crosswalk sign, etc (Chen et al., 2022). Using this two stage process, autonomous vehicles have the capability to identify signs in their environment. As opposed to a one-stage process, this two-stage process both allows machine vision systems to be more interpretable and restricts the domain of each step to make it easier to solve each individual subproblem.

Working to make these systems robust is a major step in achieving level 5 autonomous vehicles. Though many vehicles use GIS data as an augmentation to their understanding of traffic laws in their environment such as speed limit, because of areas which may lack GIS data or new signs that are put up due to construction and changing infrastructure, vision systems are still a critical component to autonomous driving. However, as it is now, many autonomous vehicles are tested and deployed only in optimal conditions. For example, Waymo, Alphabet's autonomous vehicle company, has autonomous taxis only deployed in Phoenix, Arizona and San Francisco, California (*Waypoint - The Official Waymo Blog*, n.d.-b). These two cities are known for their clear weather year-round, with Phoenix having only around 7.62 inches of rain total a year over the past decade and no snow, and San Francisco getting above average sunny days and less snow than the rest of the US (US Department of Commerce, n.d.-b, n.d.-a). Testing in optimal conditions may leave gaps in machine-vision systems' ability to detect road signs that are obscured due to weather, and it is likely that these cities are chosen and geo-fenced for their optimality. In fact, facing the difficulty of bad weather poses a significant challenge to

autonomous vehicles, as inclement weather can muck up sensors with fog and dust, or form droplets of water or ice on sensors making them struggle to produce readings for obstacles and signs. This problem is so difficult that Waymo has added small weather stations to its vehicles to get more accurate readings of weather in San Francisco where fog can be thick and sudden (*Waypoint - The Official Waymo Blog*, n.d.-a).

It is not only these weather patterns that can make it difficult for autonomous vehicles to navigate their environment and identify signage. Any change to how a sign looks under optimal conditions can wreak havoc on a machine's ability to accurately categorize these signs. This can be observed through so-called "adversarial examples". Though much research has revolved around deliberately-made adversarial examples, recent research has explored natural ones which perturbate stimuli. A 2022 paper explored how simple geometric shadows projected onto a road sign can cause models to mis-identify not only sign categories, but also to misidentify speed limits, sometimes changing a "30 mph" sign into an "80 mph" sign (Zhong et al., 2022).

Given this, it is important to test autonomous vehicle vision systems in non-optimal conditions. Furthermore, it is important to consider non-optimal conditions on a variety of axes: weather, lighting, overgrowth, graffiti, etc. Though previous work has considered obscuration on these axes within lab settings, we want to understand how these obscurations play out in real-world examples.

We aim to understand whether systems are robust in real-world, non-optimal conditions. This will give us insight into the feasibility of these models and whether sign identification is expected to be a struggle moving forward, as the world is filled with non-optimal signs. If signs are not able to be accurately identified at a rate that closely matches those of humans, then it is worth considering how to detangle road signs from a world that is inherently filled with noise. To

investigate this issue, we compiled a natural dataset of images in which road signs were obscured by shadow or overgrowth in an urban setting, and compared computer-vision categorization ability on these images to that of humans.

Methods

Data collection

Because our goal was to study classification performance in realistic non-optimal conditions, we chose to collect a custom real-world dataset for our experiment. We chose to find real signs because the distribution of the collected dataset would be more realistic than if we were to set up fake signs in a controlled indoor or outdoor environment. Data was collected by going around the Cambridge suburbs and looking for signs that were obscured by shadows, foliage, or other obstacles. We collected a total of n=64 images, of which n=29 were pedestrian crossing sign images (of 5 distinct signs), n=9 were left turn only | forward/right turn (1 distinct), n=6 were stop ahead (1 distinct), n=15 were stop (4 distinct), and n=5 were speed limit 20 (1 distinct).

Figure 1: Two example images from our custom dataset. On the left, a stop sign is partially obscured by a shadow. On the right, a pedestrian crossing sign is obscured by branches.

Human participants experiment

Subjects for this study were sourced from MIT dormspam. We sourced 47 participants, each for a 3-5 minute online experiment. We told participants they would perform a task involving classifying traffic signs and instructed participants to use a tablet or laptop to prevent small image sizes from affecting their performance (device type was not verified due to the online nature of the experiment). We included 4 attention check trials with large, centered signs from commonly-seen categories to exclude participants who are not paying attention or are incapable of performing the task.

Participants complete 64 total trials. In each trial, they are shown an image containing a traffic sign for 500 ms, before they must select the correct category of traffic sign out of a list of 17 total categories. Images will be downsized to a width of 1376 px (and proportional height) to match the resolution of images given to the machine vision model. We give participants 500 ms, a relatively longer timeframe, because street signs may occupy only a small region of the image and because models get to see the images twice while human participants see it only once. Accuracy and timing data was collected from each trial.

Machine vision experiment

The classification task was split into two phases of traffic sign detection and traffic sign classification, as is standard for real-world traffic sign identification systems from visual data (Chen et al., 2022).

For the detection phase, we trained two different detection models using the YOLOv5 object detector,[1] an architecture with about 7 million parameters in 214 layers (Bochkovskiy et al., 2020). This architecture does provide classifications for detected objects, but only its bounding boxes were used as the detection datasets used to train the models do not contain all the sign classes we were interested in testing. The first model was trained to detect traffic signs using the Traffic Signs Dataset[2] (Stallkamp et al., 2011), which contains a subset of signs from the German Traffic Sign Detection Benchmark. The dataset consists of 900 images in total of width 1376px, of which 407 images were used for training and 98 images were used for validation. The model was trained for 30 epochs with a batch size of 16. The model achieves a final error of 0.01794 on the bounding box task and achieves about 95% precision and recall on its classification component.

The second model was trained using the Tiny LISA Traffic Sign Detection Dataset[3] (Lopez-Montiel et al. 2021), which contains a subset of signs from the LISA Traffic Sign Detection Dataset. The dataset consists of 900 images in total of width 704px, of which 430 images were used for training and 80 images were used for validation. The model was trained for 30 epochs with a batch size of 16. The model achieves a final error of 0.02155 on the bounding box task and achieves 0.684 precision and 0.834 recall on its classification component. The lower classification performance may be a result of this dataset having 9 distinct classes instead of 4 like in the Traffic Signs Dataset, some of which (such as speed limit 25 and speed limit 35 look more similar than the Traffic Sign Dataset classes. The classification performance was not expected to be a limiting factor, as this model was also only used for its detection component.

---

[1] https://doi.org/10.5281/zenodo.3908559, https://github.com/Balakishan77/yolov5_custom_trained_traffic_sign_detector/blob/main/yolov5_custom_traffic_sign_data_detecter.ipynb

[2] https://www.kaggle.com/datasets/valentynsichkar/traffic-signs-dataset-in-yolo-format

[3] https://www.kaggle.com/datasets/mmontiel/tiny-lisa-traffic-sign-detection-dataset

For the classification phase, we used a pretrained CNN model made available by Zhong et al. (2022). The architecture of the model is defined in the Cleverhans library (Papernot et al. 2016). The model has 3 convolutional layers and about 700 thousand parameters. It is trained on the 17 most common classes from LISA, a dataset of U.S. traffic signs containing 47 different classes of road signs (Mogelmose et al., 2012), using Zhong et al. (2022)'s adversarial training procedure to make the system more robust to interference from shadows. The adversarial model achieves 99% accuracy on the "clean" LISA dataset and 59% accuracy on the adversarial dataset.

Images from our custom dataset were downsized to a width of 1376 px to input into the detection model to match the size of the largest training images the models were trained on, then given to the classification model in whatever size they were reduced to after cropping to the bounding box found by the detection model.

Results

Results of Human Trials

Human performance on the obscured-sign identification task was high, with human participants achieving an overall accuracy of 92.9%. This performance level remained high both with signs of varying types, as well as with varying levels of obscuration from both shadow and brush.

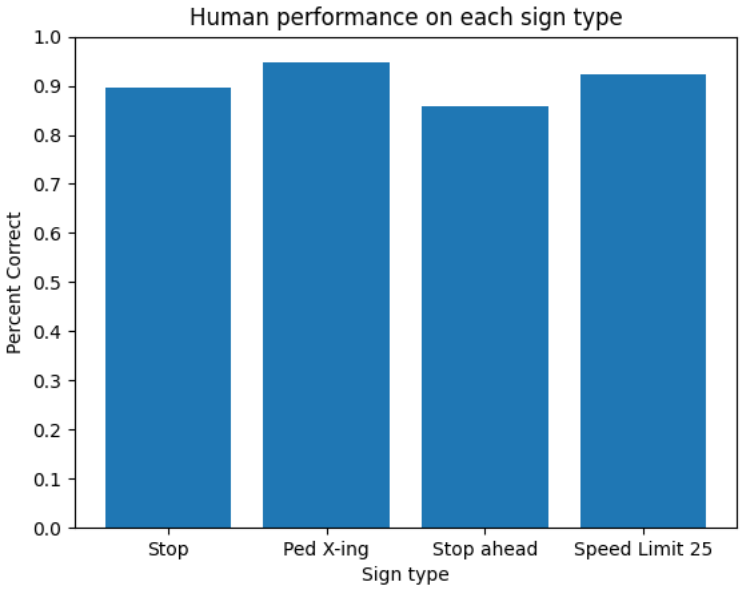
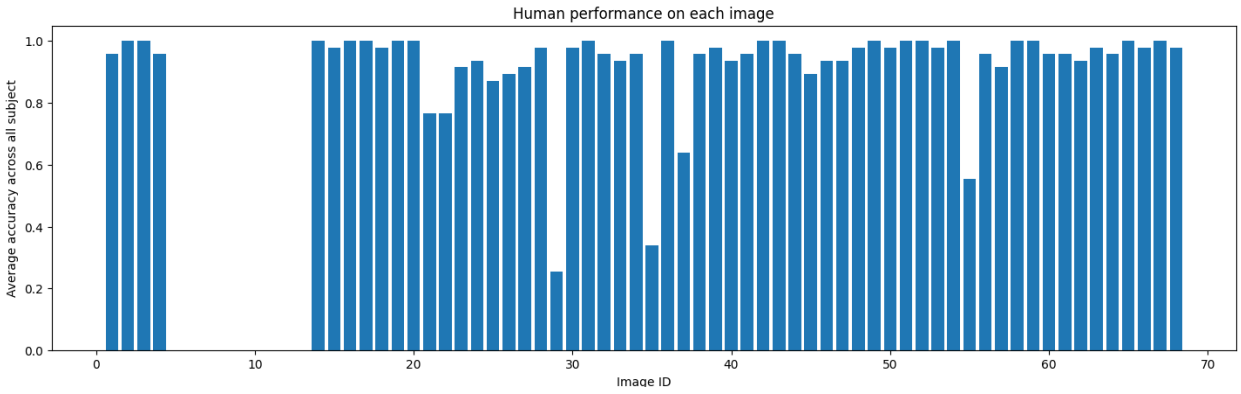


Figure 2:

Above: All sign types were well identified by human participants. Differences in performance are likely due to differences in images in the dataset; "Stop Ahead" was a sign that was most obscured in the dataset.

Below: Average performance for humans on each individual image. Images 5-13 were cut from the study due to the model not being trained on signs of this type.



We also performed an analysis on each individual image that human subjects were tested on. Performance remains very high across nearly all images tested. However, there are few cases that stand out in particular as difficult for humans to parse; images 29, 35, 37, and 55. These images in comparison to the rest of the dataset featured images containing signs that were both far away

from the camera as well as obscured by brush. Other images in the dataset that contained only one of those attributes, far away or heavy brush obscuration, were still easy for humans to identify as an aggregate.

Figure 3: Image 55. This image was particularly difficult for humans to identify the stop sign in when flashed quickly on the screen

Results of Machine Trials

Machine-performance on identification tasks was quite high when the sign was properly identified. Machine-accuracy on sign-identification was 88.9% for when the sign was identified. However, the machine-vision model struggled to identify every sign, meaning it couldn't categorize every sign. In total, the model identified 72% of signs it was meant to identify within images.

Since a two-stage process for sign categorization was used, we manually cropped signs that weren't able to be identified and fed them into the categorizer. In this case, all cropped signs were able to be properly categorized. When counting manual cropping, 92% of all signs were able to be properly identified by the categorizer. It is not currently possible to conclude whether

the model's difficulty with identification is from the difficulty posed by obscuration or from confounding variables. The images containing signs that couldn't be classified generally either contained signs that took up a large amount of space in the image or had signs that were far and obscured.

With regards to the machine's confidence, confidence on incorrect signs is 32.8%. Confidence on all signs that were correctly categorized, even with the manual crops, is 60%, though individual values range from 21% to 96% confidence in these correct classifications.

Interestingly, confidence in images that required manual crops and those that did not require manual crops are strikingly similar, being 60.3% and 58.7% respectively.

Discussion

Sign identification through both shrubbery and shadow proved to be a very possible task not only for humans but also for machine-vision systems. The output of the categorization machine-learning model achieving 92% is on par with that of human participants in this task. However, categorization was not the only bar that the machine-vision models needed to clear. Identification was also a problem and the machine-vision model performed poorly on this task, with only 72% confidence. Given this, it is possible that obscuration poses a significant problem to machine-vision models.

Further work should investigate this problem using a larger data-set that includes more sign types, different weather conditions, and more closely resembles a vehicle's point of view. Though vehicles do use video to understand the world around them rather than stationary pictures, there are still situations in which an on-vehicle camera may have access only to stationary footage for a period of time. For example, sitting at a traffic light or stop sign provides

only static input into a vehicle as the environment of signs around it is not changing, and depending on how the vehicle chooses to integrate that data may make it so that a mis-classified or mis-identified sign is an issue. Additionally, a vehicle parked on the side of the road which has just been turned in will likely have no additional data to integrate over time about its surroundings, and the identification of the sign it makes in this static position will likely help decide its next movements. This is especially important as autonomous vehicles move from being taxis that start and end at the same place everyday to being things that people own, as the places they will start and stop from will be almost infinitely larger.

Limitations

Several issues may limit the conclusions we can draw from the results of this experiment. It is less inherently problematic to draw from WEIRD populations for this experiment because we are specifically testing people's ability to identify US traffic signs, a task which only holds practical relevance for people living in the industrialized US, and not making general statements about the workings of human vision and/or psychology. However, drawing from the MIT student body means we are sampling from an especially small minority within WEIRD populations, and the results may not be representative of the performance of the general US population on this task.

The distribution of our dataset is also limited by the signs we were able to find within a reasonable amount of time. The distribution is heavily skewed toward pedestrian crossing signs because they are more common in the walkable suburban areas in which we collected our dataset, and the dataset only covers 4 different sign classes (after cutting the signage the model could not classify properly). The images were also all taken during the day in springtime in the

northeast. An improved dataset would cover a wider range of geographic regions, road types, and seasonal/lighting conditions.

For the machine experiment, our results are also limited by the power of the models we were able to train. In particular, the fact that the detection model was only trained on a subset of all possible traffic sign classes notably interfered with its ability to generalize and identify signs which were not represented in its training set. Future work would benefit from spending time to assemble a more diverse dataset and training a more sophisticated model, and potentially a newer release from the YOLO line of models, compute capabilities permitting.

Another possible bottleneck was the input size of the CNN model. The CNN model resizes all inputs to 32x32 before applying convolutional layers. This resizing did not hinder it from reaching over 99% accuracy on datasets of clean, clearly-visible signs, but it may not be sufficient for recognizing obstructed signs. If, for example, a sign is nearly 50% obstructed by leaves, a higher input resolution may be what is needed to preserve enough of the sign's contents for it to be recognizable to the model. Lightweight models are preferable for speed in autonomous vehicle applications and models should be able to run on an onboard computer, but there should be some leeway to use slightly larger models for better performance in adversarial conditions.

Figure 4: An example of an obscured sign in its original size (left) and resized to 32x32 (right).

Finally, we must consider whether the data we gave to the machine-vision models was a fair and accurate representation of what it would see on the road. It is difficult now to parse whether the difficulty in identifying signs was due to signs being obscured by shrubbery and shadow being difficult for machine-vision models to find, or whether it is because the pictures were out of distribution, having appeared in ways that were not present in the training set. Two varying features in particular could be the angle at which the sign is presented and how much space the sign takes up in the image. Most images fed into the training model were gathered from on vehicle cameras, which provide a consistent angle on most signs and only could only get so close to them.

References

Anderson, M. (2020). The road ahead for self-driving cars: The AV industry has had to reset expectations, as it shifts its focus to level 4 autonomy - [News]. *IEEE Spectrum*, *57*(5), 8–9. https://doi.org/10.1109/MSPEC.2020.9078402

Autonomous vehicle | Definition, History, & Facts | Britannica. (2023, April 4). https://www.britannica.com/technology/autonomous-vehicle

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection* (arXiv:2004.10934). arXiv. http://arxiv.org/abs/2004.10934

Campbell, S., O'Mahony, N., Krpalcova, L., Riordan, D., Walsh, J., Murphy, A., & Ryan, C. (2018). Sensor Technology in Autonomous Vehicles: A review. *2018 29th Irish Signals and Systems Conference (ISSC)*, 1–4. https://doi.org/10.1109/ISSC.2018.8585340

Chen, J., Jia, K., Chen, W., Lv, Z., & Zhang, R. (2022). A real-time and high-precision method for small traffic-signs recognition. *Neural Computing and Applications*, *34*(3), 2233–2245. https://doi.org/10.1007/s00521-021-06526-1

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). *Adversarial Examples Are Not Bugs, They Are Features* (arXiv:1905.02175). arXiv. http://arxiv.org/abs/1905.02175

Lopez-Montiel, M., Orozco-Rosas, U., Sánchez-Adame, M., Picos, K. and Ross, O. H. M., "Evaluation Method of Deep Learning-Based Embedded Systems for Traffic Sign Detection," in IEEE Access, vol. 9, pp. 101217-101238, 2021, doi: 10.1109/ACCESS.2021.3097969.

Mogelmose, A., Trivedi, M. M., & Moeslund, T. B. (2012). Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey. *IEEE*

Transactions on Intelligent Transportation Systems, *13*(4), 1484–1497. https://doi.org/10.1109/TITS.2012.2209421

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011). The German Traffic Sign Recognition Benchmark: A multi-class classification competition. *The 2011 International Joint Conference on Neural Networks*, 1453–1460. https://doi.org/10.1109/IJCNN.2011.6033395

US Department of Commerce, N. (n.d.-a). *Climate*. NOAA's National Weather Service. Retrieved April 25, 2023, from https://www.weather.gov/wrh/climate?wfo=mtr

US Department of Commerce, N. (n.d.-b). *Phoenix Rainfall Index*. NOAA's National Weather Service. Retrieved April 25, 2023, from https://www.weather.gov/psr/PRI

Waypoint - The official Waymo blog: A fog blog: Understanding a challenge inherent to driving in San Francisco. (n.d.-a). Waypoint – The Official Waymo Blog. Retrieved April 25, 2023, from https://blog.waymo.com/2021/11/a-fog-blog.html

Waypoint - The official Waymo blog: Expanding Waymo's testing to the city that keeps it weird. (n.d.-b). Waypoint – The Official Waymo Blog. Retrieved April 25, 2023, from https://blog.waymo.com/2023/03/expanding-waymos-testing-to-Austin.html

Zhong, Y., Liu, X., Zhai, D., Jiang, J., & Ji, X. (2022). Shadows can be Dangerous: Stealthy and Effective Physical-world Adversarial Attack by Natural Phenomenon. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15324–15333.

https://doi.org/10.1109/CVPR52688.2022.01491

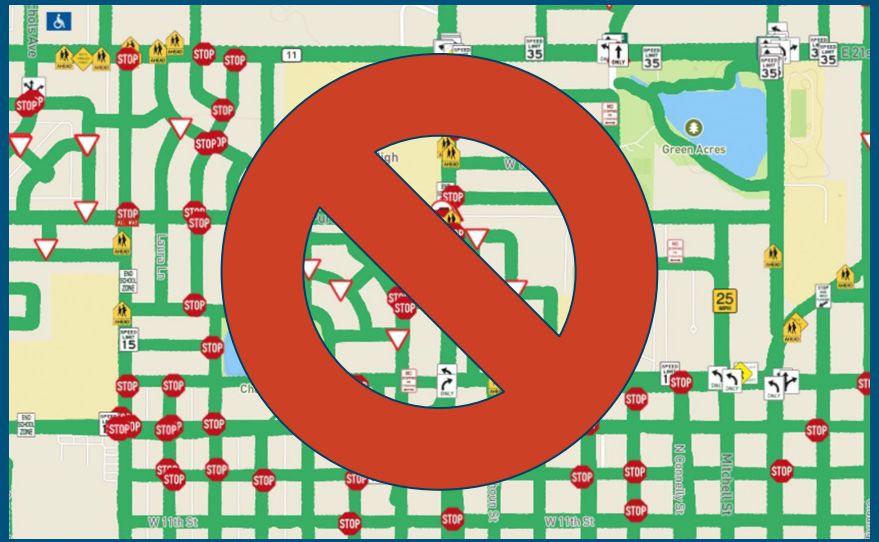
How well do machine-vision systems and humans identify obscured traffic signs on approach?

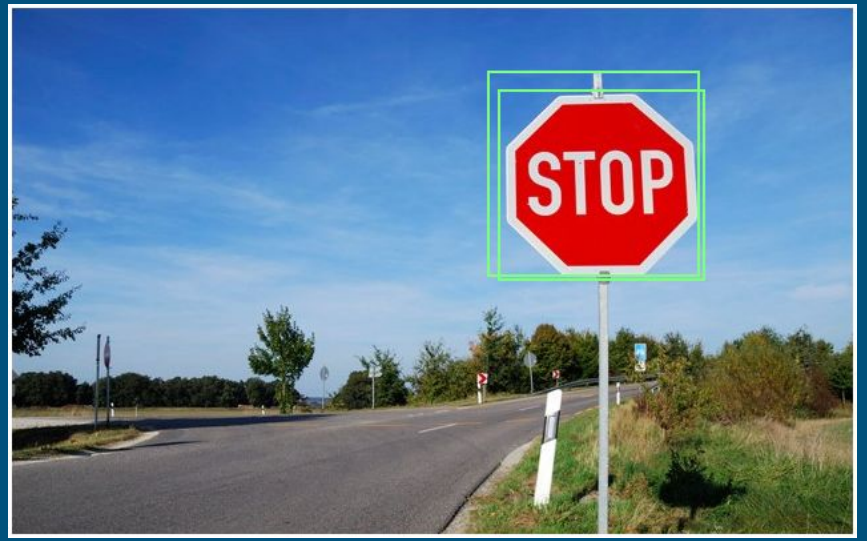
Matthew Soza and Qingyuan Lu

Background and Aims

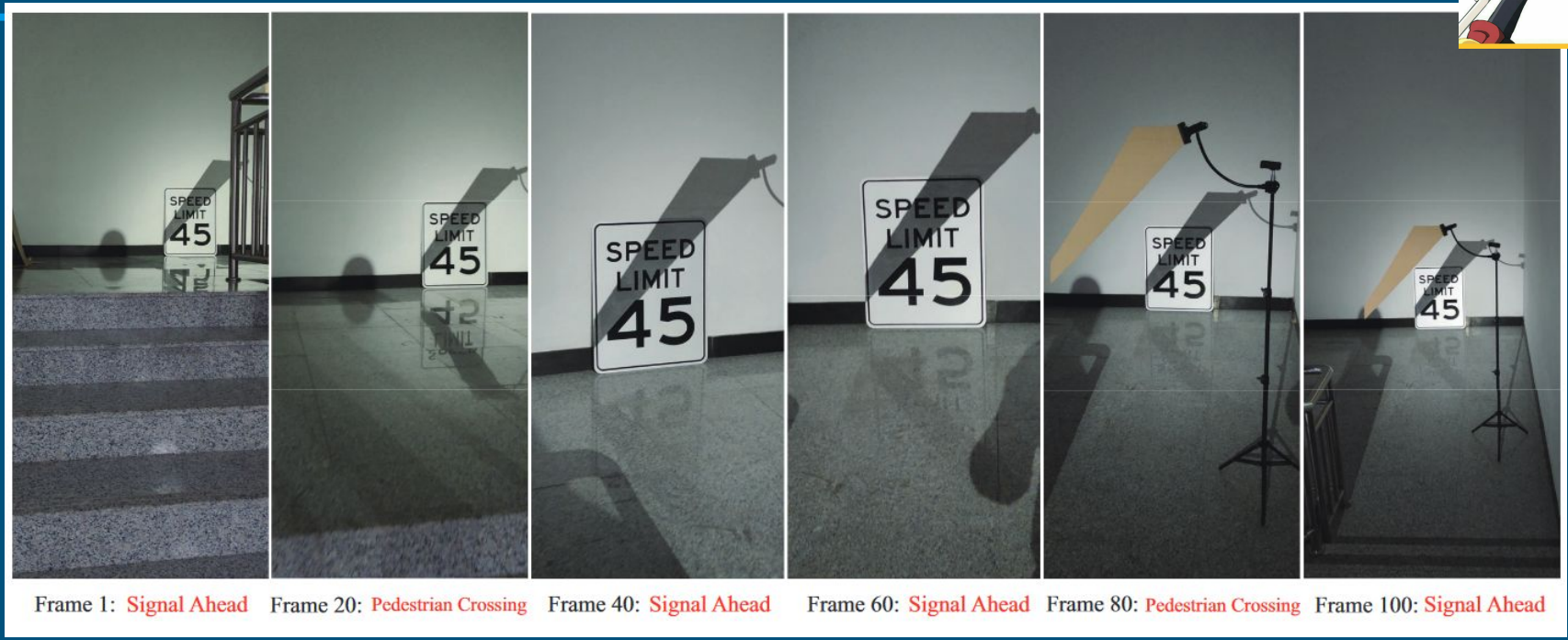
Current Machine-Vision in vehicles

- Identification of traffic signs relies on both GIS data and machine-vision





What's the problem? Aren't machine-vision systems good?



Zhong, Y., Liu, X., Zhai, D., Jiang, J., & Ji, X. (2022). Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15345-15354).


STOP
LOCKJAW
TO-BE
HATED
VOTE
ACME
KRK
STAYS CHI'D UP

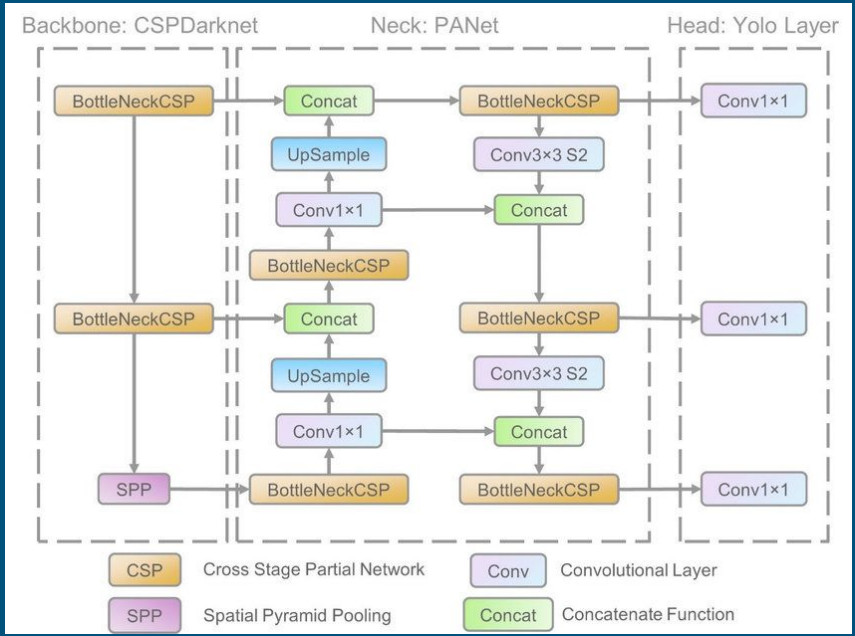

STO


QUINTARD
END
ROAD WORK

Our goal: Natural obstructions on approach

Methods

Detection model



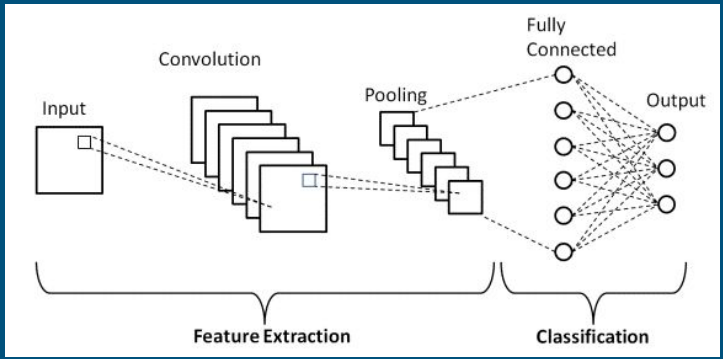
YOLOv5 architecture

figure:https://www.researchgate.net/figure/The-network-architecture-of-Yolov5-It-consists-of-three-parts-1-Backbone-CSPDarknet_fig1_349299887

A sample detection run

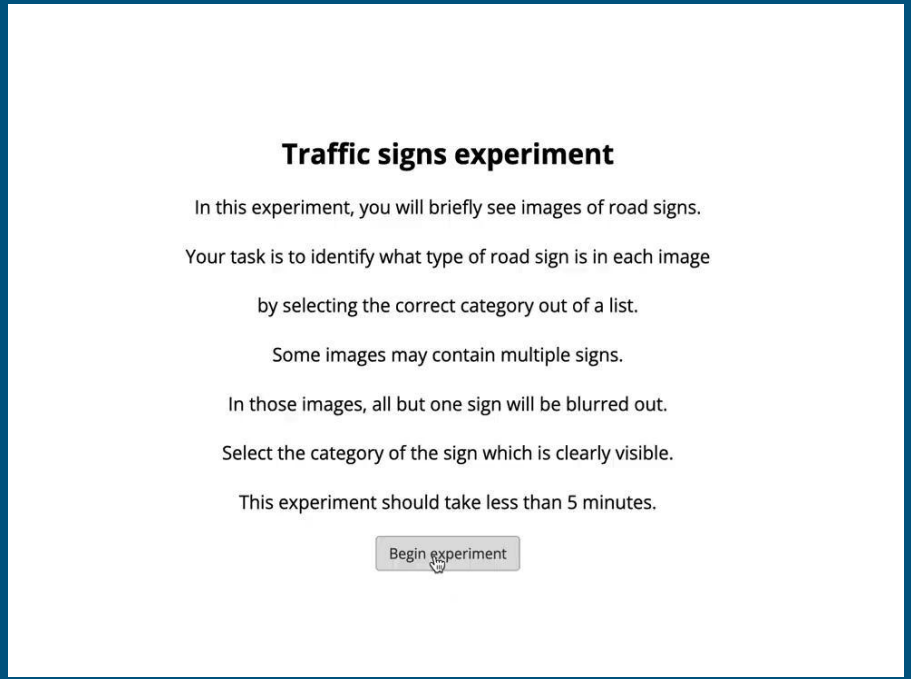
Classification model

- Pre-trained CNN model with adversarial training for shadow attack
- Operates on cropped inputs



Human trials

- Given full uncropped image, classify traffic sign
- No assumption that humans do detection and classification separately
- Participants: initial dormspam round



Results

92.3%

Human average accuracy (n=30)

83.3%

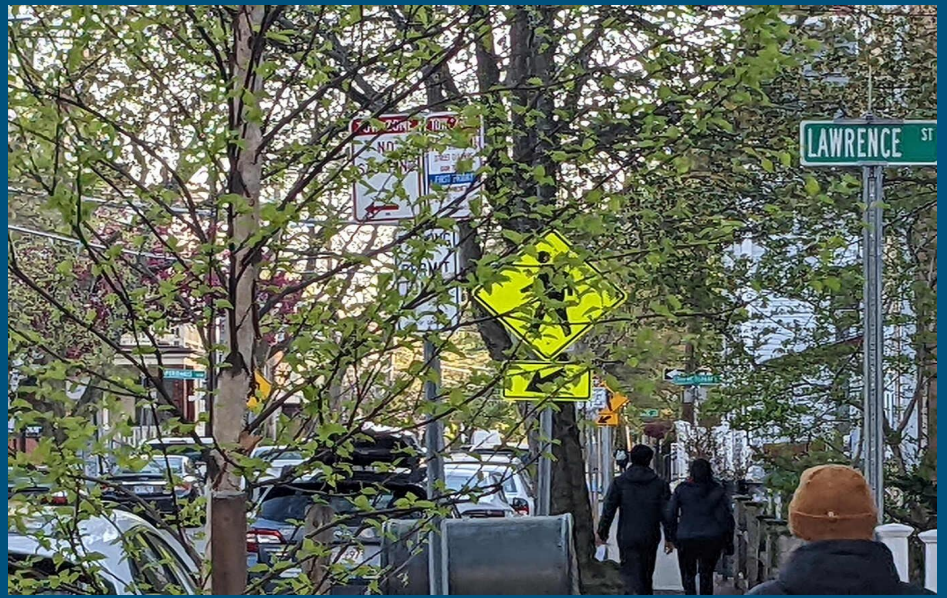
Machine-vision accuracy (20/24 images correctly classified when it identified a sign)

Other preliminary data

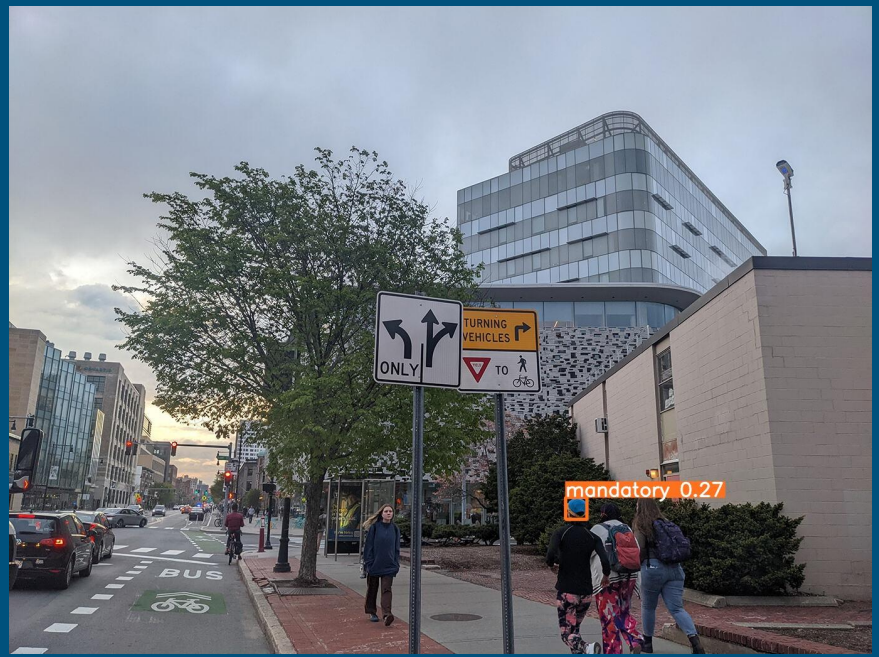
- Performance including when it did NOT identify a sign: 31.3%
- Average correct confidence: 75.8%
- Average incorrect confidence: 44.3%

Limitations and Next Steps

Detection model performance



Detection model performance



Next steps

- Participant recruitment on Prolific
- More in-depth data analysis
  - What are humans most likely to misclassify, i.e: a stop sign for a yield sign?
  - What images is the model most likely to misclassify, or, conversely, classify correctly?

Questions?