2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Collaborative Perception for Automated Vehicles Leveraging Vehicle-to-Vehicle Communications

Ryan Yee, Ellick Chan and Carmine Senatore
Exponent, Inc.

Bin Cheng and Gaurav Bansal
Toyota InfoTechnology Center, U.S.A., Inc.

Abstract— **Currently, many automated vehicle systems primarily perceive the environment from a single perspective and as a result are unable to leverage additional scene information from the viewpoint of other vehicles on the road using vehicle-to-vehicle communication technologies. We study how increased data sharing can improve the perception capabilities of automated vehicles. Our methodology shares sensor measurements and objects detected by state-of-the-art deep learning networks between vehicles to increase the automated driving systems confidence of detecting objects using a 3D sensor fusion algorithm. This approach can benefit scenarios where an object may be occluded (fully or partially) or located too far away to classify accurately by a single vehicle alone.**

I. Introduction

Accurate perception of the environment is a key design consideration for automated vehicles. For such vehicles to operate safely, they must be able to sense the world accurately to avoid obstacles, stay on their intended course and plan a safe route. Although automated vehicles may be outfitted with many sensors working together in concert, these vehicles still fundamentally perceive the world from a single point of view. In this paper, we explore how vehicles may share data from different viewpoints to improve perception.

Several challenges currently limit the efficacy of single-viewpoint systems. (a) Physical occlusions that prevent the system from seeing objects (e.g., buildings or other vehicles etc.), (b) Sensor limitations (e.g., resolving power, lighting, etc.), and (c) Computer vision system limitations (e.g., uncertain detection, mis-classification, etc.). Examples of ensuing misdetections or false detections are depicted in Fig. 1.

A collaborative perception system would use multiple viewpoints to improve the performance of detection systems in these cases by using vehicle-to-vehicle communication technology. Such technology is not limited to vehicles, and the system can more generically work with data shared by cameras mounted on traffic lights, buildings, and aerial sources. By sharing and fusing the sensor information from multiple viewpoints, it is expected that the perception of a single vehicle can have improved range and field-of-view (even beyond line-of-sight conditions). This is because its sensing region is effectively extended by aggregating perception information from all neighboring road users [1], [2]. With such enhanced perception capabilities, a single vehicle's situational awareness as well as that of its peers can be largely improved. To maximize the benefit of collaborative perception, a significant number of vehicles must be equipped with the exteroceptive sensors which enable perception such as cameras, LIDARs, and radars, as well as some form of vehicle-to-vehicle communication technology. To that end, twenty automakers representing 99% of the U.S. market pledged to voluntarily equip new passenger vehicles in the U.S. by September 1, 2022 with a low-speed AEB system that includes forward collision warning (FCW), a technology that uses camera, LIDAR, and radar to mitigate rear-end collisions [3], [4].

We devise a methodology, depicted in Figure 2, that considers the input from multiple viewpoints to reconcile a global view of the world that leverages the individual views of each vehicle. The guiding principles of a stereo camera are similar. In this work, a computer vision object detector identifies bounding boxes for objects detected in the scene and also reports the confidence of object detection. These bounding boxes are then considered from a global 3D perspective that aggregates the information from multiple viewpoints. Detecting an object from more than one viewpoint reveals additional information about the position of the object and helps to resolve object identity. Measurements from this system can be used to *confirm* the identity of a weakly detected object, or if only one observer is able to detect an object, the information can be *shared* with other participants. Although this paper considers camera vision inputs only, the methodology can be extended to other exteroceptive sensors. The current focus is primarily in improving the detection confidence for previously undetected objects, but future work may address the other described scenarios in more detail.

This paper is organized as follows. Section II explores related work in the context of perception systems and vehicle-to-vehicle sharing. Section III describes our fusion method in detail. Section IV describes our experimental setup and data collection methodology. Section V evaluates our results on some real-world datasets and discusses the significance of this work. Section VI provides concluding remarks and possibilities for future work.

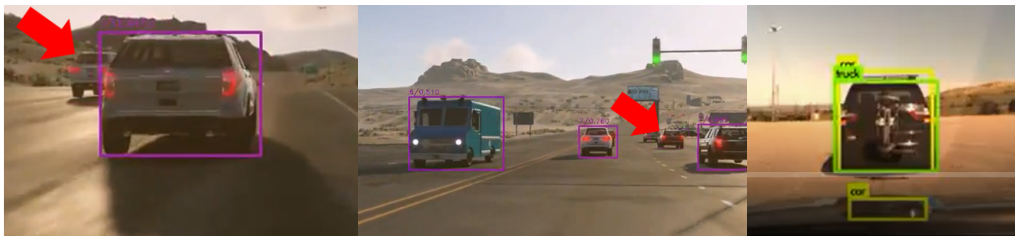
II. Background

The subsections below discuss vehicular communications and object detection as they are the two technologies that enable the implementation of the novel methodology proposed in this paper.

A. Vehicular communications

Dedicated short-range communications (DSRC) is a state-of-the-art wireless technology that enables vehicle-to-vehicle

978-1-5386-4451-5/18/$31.00 ©2018 IEEE



(a) Occlusion: vehicle indicated by the red arrow was not detected

(b) Limited range: vehicle indicated by the red arrow was not detected

(c) Uncertain detection: vehicle is classified as both a car and truck

Fig. 1: Perception challenges

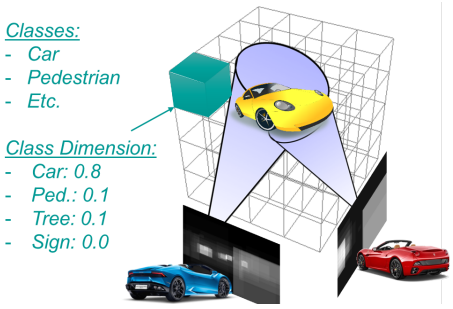


Fig. 2: 3D fusion algorithm

(V2V), vehicle-to-infrastructure (V2I), or even vehicle-to-everything (V2X) communication systems [5]. DSRC's primary use cases are driving safety related applications. In these applications, each vehicle periodically exchanges its status information (e.g. driving speed, GPS location, heading, etc.) with other road users (e.g., pedestrians, bicyclists, even smart traffic lights). With shared information, road users can achieve better situational awareness and thus improve driving safety [6]–[8]. However, although DSRC enables basic message exchange among road users with a range of up to 1000m (in near-ideal conditions), the maximum data rates of DSRC are normally as low as 2-6 Mbps.

An improvement in terms of bandwidth is provided by broadband cellular network technologies. The third Generation Partnership Project (3GPP) standard defines a V2X communication system based on the device-to-device (D2D) mode in fourth generation (4G) cellular systems [9]. This system is capable of 100 Mbps in high mobility scenarios.

Finally, an emerging technology for high bandwidth vehicle-to-vehicle communication is based on millimeter wave (mmWave) communication [10]. mmWave is not new to the automotive world since current automotive radars already use the mmWave spectrum [11] and vehicular communication using the mmWave band has already been tested [12].

In the near future, vehicle-to-vehicle capabilities are expected to increasingly penetrate the market with several automakers already offering such technology on today's vehicles [13], [14].

B. Object detection using deep learning

In computer vision, object detection is the task of finding and classifying a variable number of objects in an image. Recently, deep learning algorithms have significantly outperformed classical approaches [15] and have become state of the art. A brief summary of the leading object detection algorithms is presented below. We start with R-CNN [16] which was able to boast a near 50% improvement on the PASCAL VOC 2010 object detection challenge. This work inspired a large body of research investigating methods for region proposal and region-based feature extraction. Similar to R-CNN, Fast R-CNN [17] used a technique called Selective Search to generate a region proposal. Compared to R-CNN, Fast R-CNN is faster and easier to train due to the added Region-of-Interest (RoI) pooling layer for the end-to-end training.

YOLO [18] proposed a simple convolutional neural network that can achieve both high accuracy and efficiency, allowing real-time object detection for the first time. SSD [19], built upon YOLO, was able to achieve better performance and speed by utilizing multiple sized convolutional feature maps.

YOLO and SSD were considered for the work presented in this paper to provide object detection for single viewpoints (more details provided in Section V). The proposed approach then seeks to improve the detection performance by fusing the detection performance from the individual viewpoints.

III. Methodology

A. Overview

Our algorithm constructs a 3D map of objects identified from multiple viewpoints in a global reference frame as depicted in Fig. 2. Each viewpoint utilizes a camera (or other exteroceptive sensors) to detect objects, and after an object detector such as Tensorflow Single Shot Detector (SSD) [19] or YOLO [18] is run bounding boxes are proposed. We fuse together these observations in 3D space using the position, heading and detected object boxes from each viewpoint. We assume that the bounding boxes and positions are shared with other nearby vehicles using V2V technology. Vehicles share information on detected objects, not the actual sensor data, hence only requiring a low-bandwidth link or mobile network. Vehicles may optionally share full data streams with

other vehicles, but this possibility was not considered in the work presented in this paper.

The algorithm works in 4 steps: (1) Imaging and localization, (2) Object detection, (3) 3D fusion in the global reference frame, and (4) Backprojection to each viewpoint in the local reference frames.

B. Imaging and localization

The first step is common to any automated vehicle technology with the vehicle acquiring exteroceptive (camera in our case) and proprioceptive (GPS and gyros in our case) data from the on-board sensors. From this data, a position and heading estimate for the vehicle are calculated. We assume that sensors are properly calibrated, synchronized via timestamps and tuned to work well in the deployment environment. We also assume low localization error.

C. Object detection and heatmaps

Next, we process captured image frames to detect objects. Each detected object is represented by a bounding box with the proposed class type and confidence. This is depicted in Fig. 3. Normally for single viewpoint systems, the detections are thresholded to suppress uncertain predictions that produce spurious boxes. Our system aims to improve such uncertain detections by observing from many viewpoints, so we intentionally do not filter boxes based on confidence. Instead, we produce confidence heatmaps which show the degree of confidence for each object position.

Fig. 3 center and right represent degrees of confidence where brighter areas are more confident and darker areas are less confident. The central image shows heatmaps for the car class while the right image shows heatmaps for the truck class. The maximum confidence value in the map is reported in the title of the subfigure.

These heatmaps tend to be very confident at the center of an object, while dropping off near the edges. We leverage this additional information in later steps to help resolve uncertain detections. We assume that object detectors perform reasonably well in the deployment environment for the objects of interest.

Each heatmap image is a W×H×C multidimensional array which represents width, height, and the number of object classes respectively. Class is encoded as a one-hot vector where the vector is strongest for the most likely class. The classifier makes a prediction over the probability distribution over likely classes of a pixel.

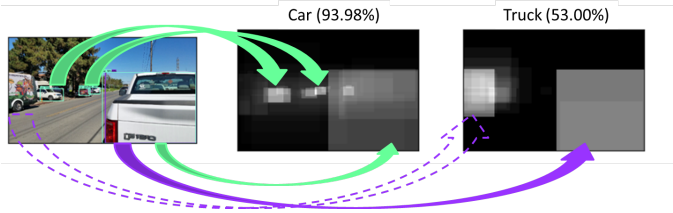


Fig. 3: Car (center) and truck (right) class heatmaps are generated from the camera data which has been processed using the TensorFlow object detector (left)

D. Voxel projection

This step fuses together measurements from multiple viewpoints into a unified 3D space in the global reference frame. The basic model is inspired by ray tracing used in computer graphics [20]. Ray tracing simulates rays of light interacting with the environment by tracing the path of a ray of light from light sources bouncing off surfaces until the rays intersect with an image plane.

In our application, multiple viewpoints observe objects in the scene. Physically, real objects in the scene interact with rays of light to generate images of the objects onto several image planes of observation corresponding to each viewpoint. We invert the problem by using the bounding boxes of detected objects on each image plane and solve for potential locations of the objects in the scene that could have produced such a heatmap from each viewpoint. Since our cameras do not directly capture depth information, we assume that each object may lie within the constraints of a conical shape in space with the axis of the cone aligned with the heading of the camera. This is depicted in Fig. 2. A distance estimator was also implemented, with more details provided in Section III.H.

More specifically, we assume that space is divided into voxels (3D pixels) and we calculate which voxels may be responsible for producing an image on the image plane. The solution of this is a cone shape, or a frustum shape if we assume constraints on where an object may be placed.

To compute this, our algorithm casts rays through each voxel in 3D space and labels the voxel with the label of the pixel from the heatmap that corresponds to the ray that passes through the image plane into voxel space. Each viewpoint produces a set of labeled voxels corresponding to class confidence, and the fusion algorithm aggregates the measurements from multiple viewpoints.

E. Fusion

Once each viewpoint projects its heatmap into voxel space in the global reference frame, we make inferences into real 3D space by reconciling the individual voxel maps produced by each viewpoint. This step produces a voxel map in the global reference frame from the perspective of each viewpoint. Voxel maps are subsequently aggregated using several fusion strategies designed based on statistics and machine learning principles to generate a global perspective. We evaluated three fusion strategies: (a) mean value, (b) product, and (c) Treisman [21].

The mean value strategy takes the average of the confidence values from individual viewpoints. The advantage of this strategy is in its simplicity. For example, if vehicle 1 sees an object with 0.4 confidence and vehicle 2 observes the same object with 0.6 confidence, the mean value is simply 0.5. The drawback of this approach is that the confidence is always somewhere between the least and most confident observers.

The product strategy considers the joint confidence of two independent observers. We multiply all the confidence values together to get an aggregate value. The benefit of this

approach is that false positives are suppressed and the two observers can confirm an observation jointly. This strategy seems to be the most appropriate when the classifier may be experiencing a high rate of false positives and some level of confirmation is needed. The drawback to this strategy is that it can never produce a joint confidence that is higher than any individual confidence value.

The third strategy we considered is based on a methodology to combine the observations from two observers using partial probability summation [21]. The joint confidence of two independent observers is expressed by: $p_{12} = p_1 p_2 + bp_1(1-p_2) + bp_2(1-p_1)$, where p_{12} represents the confidence of both observers and p_1, p_2 represents the confidence of observers 1 and 2 respectively, and b is a biasing parameter. The advantage of this approach is that it evaluates the joint confidence in an additive way where the joint confidence can be higher than any individual confidence. In our example, the combined value for $b = 1$ would be $0.4 * 0.6 + 0.4 * (1 - 0.6) + 0.6 * (1 - 0.4) = 0.76$. The disadvantage of this strategy is that there is no way to disconfirm an observation as all terms are positive. We empirically observe that confidence levels are higher with this approach and background noise can become an issue if detection thresholds are not adjusted to compensate. Mathematically, this approach contains elements of the previous two approaches that blends a sum and a product together. In the Evaluation section we present results for all three strategies including a sensitivity analysis for the the biasing parameter b.

F. Backprojection

After the scene has been computed in the global voxel space using an aggregation strategy, we must convey this information back to the individual viewpoints to assess the performance of our methodology in terms of the individual vehicle object detection performance. To accomplish this task, we invert the ray tracing problem once again and this projects the fused information back on each individual vehicle local reference frame.

G. Overall workflow

The overall workflow is depicted in Fig. 4 using a computer simulation developed to highlight the main steps of the algorithm. Fig. 4a shows an overview of the scene with two viewpoints V1 and V2 observing 3 vehicles in the scene: a black truck, a red car and a blue car. The rays cast for V1 and V2 are shown in red and blue respectively in Fig. 4a.

Step 1: Figures 4b and 4c represent the scene perceived from the perspective of each viewpoint. This is simply the projection of the objects in the scene onto a flat image plane as the camera on each individual vehicle would see them. We depict the image captured by the camera as well as the heatmap shown in the bounding box.

Step 2: Object detection is performed on the collected images and darker colors represent object detection with lower confidence while brighter colors indicate higher confidence. Note that some objects may be partially occluded as shown in the center of Figure 4c with one object (i.e., the black truck detected with low confidence, hence gray) partially obstructed by another object (i.e., the red car detected with high confidence, hence white).

Step 3: Rays are projected from each viewpoint through the image plane into voxel space. This is shown in Figure 4d, which represents the fused heatmap. Areas where intersecting rays agree on object class exhibit stronger object detection confidences shown roughly where the objects lie.

Step 4: The back projection for V1 and V2 are presented in Figures 4e and 4f showing fused heatmaps in the local reference frame. These are similar to the individual viewpoints (Fig. 4b and 4c) but with adjusted confidence due to the additional information from the other vehicle.

H. Distance approximation

To more precisely locate objects in space, we approximate the distance to the detected object or vehicle within its field of view for the purposes of limiting the length of the cone of confidence. As noted in Fig. 4d, long cones cast at oblique angles can lead to ghost artifacts shown in both the V1 and V2 backprojection heatmaps. This also causes the appearance of a ghost object in the fused heatmap where the cones intersect without an object at the intersection. This is due to uncertainty in the position of the object detected, and a long cone is just a first-order approximation of inferring the vehicle's location. Introducing an object distance estimator allows cones to be truncated into frusta, hence reducing the ghosting effect.

Algorithms to estimate vehicle distance from mono camera input have been developed for Adaptive Cruise Control and Automatic Emergency Braking applications [22]. The development of a robust distance estimator was beyond the scope of the current work and therefore distance estimates were approximated with basic trigonometry complemented with assumptions on vehicle dimensions.

IV. Experimental setup and data collection

Current and future designs for automated vehicles typically include a range of sensors that can perceive the environment from the front, back and sides of the vehicle. This may include as many as 6 to 8 cameras around the vehicle with overlapping fields of view. In this work, we evaluate our approach using two front-facing cameras, but the approach is general enough to work with any number of cameras, where the accuracy of perception is expected to increase with additional cameras.

A GoPro Hero 5 Black was selected for this work to capture perception (camera) and localization data. Video data can be collected at a number of different resolutions and frame rates, but 1920×1080 resolution at 24 frames per second was found to be adequate for this work without being burdened by large file sizes. The camera also has a number of different Fields of View (FOV) that can be selected, with larger FOVs providing more scene information at the cost of increasing optical distortion. The "Linear" FOV was chosen as this produced images with the least amount of optical distortion.

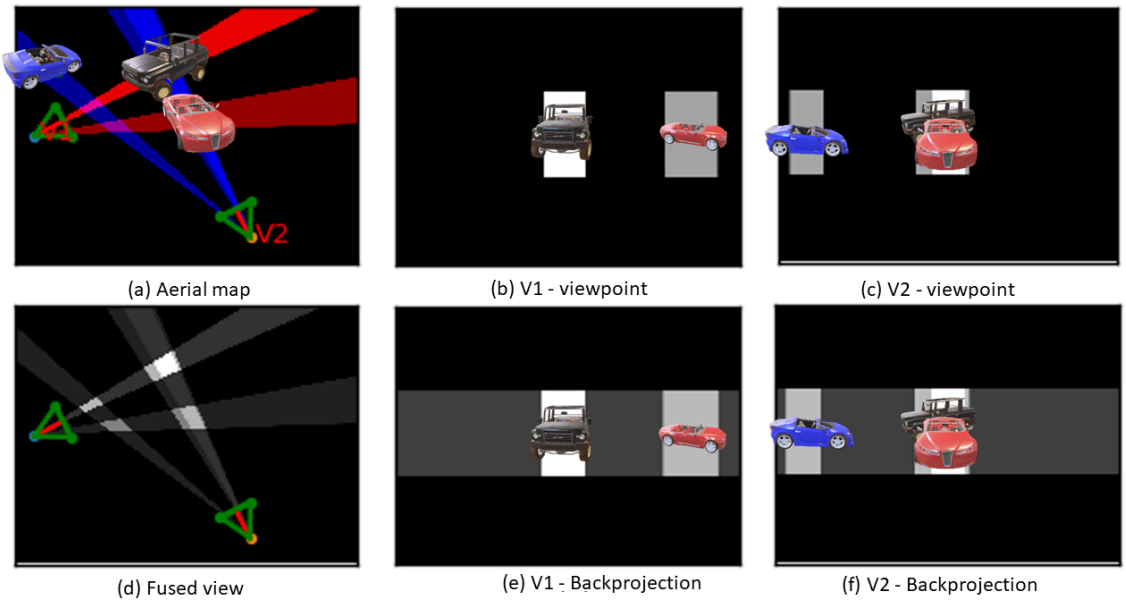


Fig. 4: Workflow overview

The GoPro camera also includes a triaxial accelerometer, a gyroscope, a temperature sensor, and a GPS receiver. Each of these sensors samples data at different rates, with the GPS collecting data at approximately 18 Hz, the triaxial accelerometer at 200 Hz, the gyroscope at 400 Hz, and the temperature sensor at 1 Hz. Metadata was extracted from the GoPro along with each video frame, but were post-processed to ensure that the data collected was aligned in time with identical step sizes.

To facilitate data collection for the collaborative perception proof-of-concept measurements, two GoPros were mounted on a single vehicle windshield at a known distance of approximately 40 inches apart. This not only simplifies the task of collecting perception data from multiple viewpoints, but allows for: (1) an assessment of the relative GPS error between two GoPros as they are mounted a known fixed-distance apart, and (2) straightforward synchronization of two GoPro data sets (mimicking two vehicles) as the GPS speed on each GoPro can be matched and aligned in time. In initial experiments involving two vehicles, a single GoPro was placed near the center and top of each windshield.

A. GPS error evaluation and data synchronization

Localization accuracy is a key aspect for improving detection through collaborative perception. Errors in GPS and inertial measurements from any one vehicle that participates in the detection process can affect the rate of false positives or negatives. Therefore, in order to realize the full potential of collaborative perception, a minimization of these errors must occur. Although this is beyond the scope of the current proof-of-concept discussed in this paper, consideration is given to evaluate the relative localization error in this work, using the single vehicle equipped with two GoPros data. Generally, it is anticipated that a number of methods and instrumentation options such as differential GPS can be implemented to ensure accurate vehicle localization in the context of collaborative perception among a number of vehicles.

Fig. 5 illustrates the relative localization error between the two GoPros, which is determined as the difference between the Euclidean distance defined by the two GPS coordinates and the known physical distance between the GoPros of approximately 1 meter. The maximum error observed was on the order of 4 meters, with a mean error of approximately 2 meters.

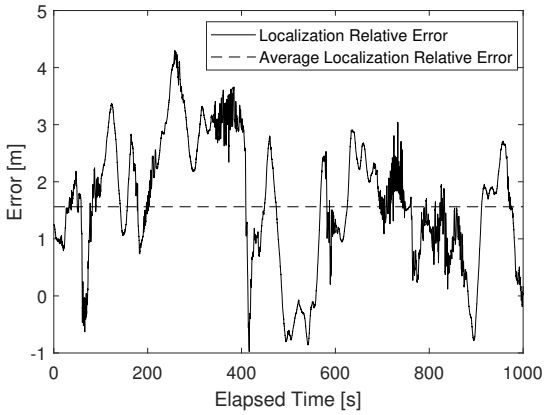


Fig. 5: GPS relative error measured between two GoPros mounted a fixed distance apart on the windshield of a single vehicle

To synchronize the video and GPS data between each GoPro mounted on a single vehicle, the speed profiles measured by each unit were matched in time. The quality

of the synchronization was assessed by analyzing the residuals, that is the difference in speeds measured by the left and right GoPros. It was determined that the absolute and relative errors did not exceed approximately 3 km/h and 5%, respectively.

For the two vehicle experiment, data synchronization was achieved by observing a common event between both camera frames (e.g., a brake light on a vehicle observed by both cameras). Obviously common event matching may not be practical on production level systems where data synchronization can be achieved through GPS and network time-stamping.

V. Evaluation

Typically, perception work in the automated vehicle space has been evaluated using a standard dataset such as KTTI [23] or CityScapes [24]. These datasets primarily perceive the world from a single point of view and as such, they are not suited to this work. We instead evaluate on two datasets we collected as described in Section IV.

At the time of writing, the collected data was not entirely labelled. However, our carefully designed experimental setup affords us an opportunity to evaluate our results via self-consistency. Self-consistency in this case means that objects detected by the left and right cameras on a single vehicle agree when run through the full fusion system and back propagated. Weakly detected objects should be amplified and noisy object detections should be disconfirmed by the approach. Additionally, objects which may be occluded or partially occluded through the field of view of one camera might be detected with the additional information provided by the other camera.

We evaluate the predictions made by our algorithm by comparing the original heatmaps to the new predictions made by the mean, product and Treisman fusion methods. We expect that under some of these conditions, some weakly detected vehicles may be detected more accurately. We evaluate this qualitatively for the two main use cases of detecting weakly/not detected and distant objects. The evaluations were conducted on a machine with an Intel Xeon E5-2670V3 (12 core) CPU and 128 GB Memory.

A. Initial Qualitative Results

Figure 6 compares the results for various fusion strategies where multiple viewpoints are expected to improve detection where there is either a weak or missed detection. In this particular example, the case of a single vehicle detected by one camera but not detected by the other is analyzed. In Figure 6, the lower left and middle quadrants present the output confidence for different fusion strategies. Each row in these quadrants refers to a different fusion strategy for the objects detected in the image above. Given the two-dimensional nature of the approach proposed, objects on the back projection are characterized by their azimuthal location with respect to the camera centerline and their width. The label "Original" indicates the output confidence provided by the object detector with the rows below representing the confidence achieved after fusion. Areas in red indicate increased confidence while areas in blue indicate decreased confidence. As expected, using the "Mean" as a metric increases the confidence for one camera but decreases the confidence for the other. Additionally, using the "Product" as a metric decreases the overall confidence for both. However, it can be observed that by using the Treisman fusion strategy (and by selecting particular values for the biasing parameter b), the overall confidence for detecting the vehicle identified with the purple arrow increases for the right camera (as the areas for the "Treisman blend" are shaded in red), where it was previously not detected (the absence of a detection is indicated by the white shade "Original" metric and the lack of a bounding box drawn by the right camera). These maps show that our algorithm is able to guide confidence levels accurately using information from multiple viewpoints.

Figure 7 illustrates preliminary data collected using two vehicles, each equipped with a single camera and GPS. In this particular example, the lead vehicle is driving on a single lane roadway which has a slight incline followed by a slight decline over a bridge and the trailing vehicle is following several car lengths behind. The lead vehicle detects an oncoming vehicle in the opposing lane of traffic identified by the bounding box. As one might expect, the trailing vehicle does not detect this vehicle, either due to resolution (which can be a result of a range issue) or geometry (the vehicle to be detected is partially occluded by the inclination of the roadway). Using the Treisman blend as a metric, it can be observed that the detection confidence increases for the trailing vehicle, while the confidence remains the same for the lead vehicle (as expected, because the trailing vehicle does not contribute to the measurement). Therefore without any fusion, the trailing vehicle would not have any increased confidence (from zero confidence) of detecting the vehicle approaching in the opposite lane.

Although it is currently unclear which fusion strategy might result in an optimized receiver operating characteristic, initial results here demonstrate that the preliminary fusion algorithm can improve the detection of partially occluded, distant, or weakly detected vehicles. Further evaluation of the fusion strategies discussed here (as well as others) can occur with a ground truth data set.

B. Classifier selection

Tensorflow's object detection module is currently hosted in the research models repository on GitHub [25]. As such, we understand that it is meant for research and testing purposes rather than being released as production-level code. We leverage this research code as the basis for our prototype. Other models such as YOLO or production models may offer lower error rates on object detection. Although the object detector is imperfect and not always completely confident in detection, we show that our approach allows predictions from multiple weak classifiers to be aggregated in the spatial dimension to provide more confidence much in the way that machine learning ensembles enables additional strength in classification.

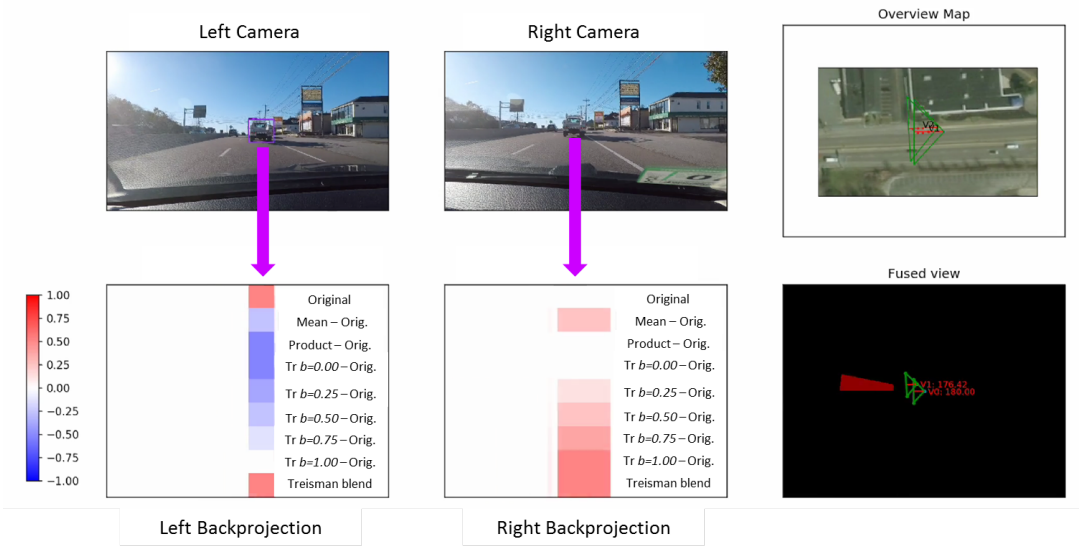


Fig. 6: Fusion algorithm result for a single frame from data collected using a single vehicle with two cameras

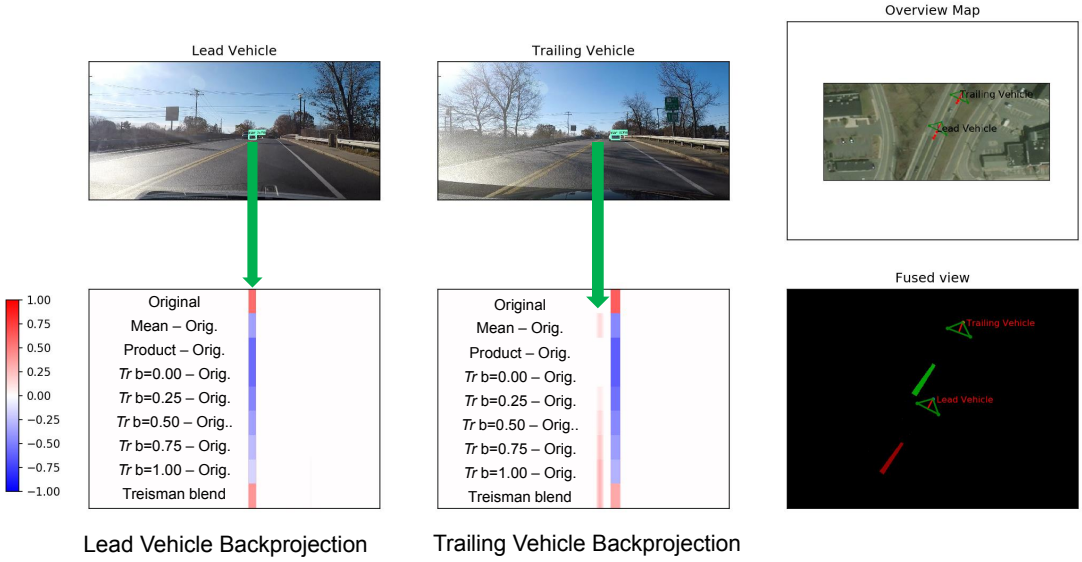


Fig. 7: Fusion algorithm result from data collected using two vehicles, each equipped with a single camera and GPS

For the purpose of rapid prototyping, we used the COCO-based pre-trained Mobilenets SSD model. The models listed on the model zoo trade off speed for performance and we found that the Mobilenets detector provided adequate performance for our task. Generally, more accurate models are slower but given limited processing hardware, a faster but less accurate model may offer better perception of the environment due to a better frame rate and general ability to react to key events faster, such as applying the brakes if needed. Also, more accurate models may require more computation power which may consume extra energy on electric vehicles, thus limiting driving range and also may need extra cooling for the electronic components.

One observation we make is that these object detectors were trained on a wide variety of object classes. Therefore, performance is balanced for classes that may not be beneficial for driving such as identifying kites or coffee cups. An object detector tuned for the classes such car, truck, person, street sign and other specific classes may attain better performance for this particular task.

VI. Concluding Remarks and Future Work

This paper proposes a collaborative perception algorithm for automated vehicles by leveraging V2V communications. Single-view systems can suffer from several challenges such

as occlusions, limited range, and low-confidence of detection. We have proposed a novel data fusion algorithm, by implementing the latest deep learning techniques and aggregating object detection information from multiple viewpoints, to improve detection performance. We have conducted proof-of-concept experiments to demonstrate the technique using both a single vehicle equipped with two cameras and GPS units, as well as two vehicles equipped with a camera and GPS unit in each. The results presented are evaluated using multiple fusion strategies, where they appear to be consistent with expectation. Preliminary results indicate that by sharing data among vehicles and using partial probability summation, the confidence of a previously missed detection can be improved.

Currently, there are no labeled data-sets for cooperative driving. As a future work, we plan to collect such data-sets, and analyze and improve the performance of our algorithm over a larger number of frames. We also plan to collect more diverse data incorporating various traffic scenarios (urban, suburban, highways, etc.), lighting conditions, and classes of dynamic objects (pedestrians, vehicles, bicyclists, etc.). In addition, we plan to perform additional sensitivity analyses on the GPS measurements, as they provide a source of error in the ultimate accuracy and performance of the collaborative perception technique that is proposed. It is also anticipated that as this work continues to evolve, implementation of V2V technology enabling collaborative perception will become clearer.

REFERENCES

[1] Vutha Va, Takayuki Shimizu, Gaurav Bansal, and Robert W. Heath Jr., "Millimeter Wave Vehicular Communications: A Survey," *Foundations and Trends in Networking*, vol. 10, no. 1, pp. 1–113, 2016.

[2] S. W. Kim, B. Qin, Z. J. Chong, X. Shen, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "Multivehicle Cooperative Driving Using Cooperative Perception: Design and Experimental Validation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 663–680, April 2015.

[3] "NHTSA-IIHS Announcement on AEB," `https://www.nhtsa.gov/press-releases/nhtsa-iihs-announcement-aeb`.

[4] "Where Automakers Stand on Automatic Emergency Braking Pledge," https://www.consumerreports.org/car-safety/where-automakers-stand-on-automatic-emergency-braking-pledge/.

[5] J. B. Kenney, "Dedicated Short-Range Communications (DSRC) Standards in the United States," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1162–1182, July 2011.

[6] F. Ahmed-Zaid, F. Bai, S. Bai, C. Basnayake, B. Bellur, S. Brovold, G. Brown, L. Caminiti, D. Cunningham, and et al. H. Elzein, "Vehicle Safety Communications-applications(VSC-A) Final Report," Tech. Rep., National Highway Traffic Safety Administration (NHTSA), 2011.

[7] "Vehicle Safety Communications Project: Task 3 Final Report: Identify Intelligent Vehicle Safety Applications Enabled by DSRC," Tech. Rep., National Highway Traffic Safety Administration (NHTSA), 2005.

[8] Fan Bai, Daniel D. Stancil, and Hariharan Krishnan, "Toward Understanding Characteristics of Dedicated Short Range Communications (DSRC) from a Perspective of Vehicular Network Engineers," in *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2010, MobiCom '10, pp. 329–340, ACM.

[9] H. Seo, K. D. Lee, S. Yasukawa, Y. Peng, and P. Sartori, "LTE Evolution for Vehicle-to-Everything Services," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 22–28, June 2016.

[10] T.S. Rappaport, R.W. Heath, R.C. Daniels, and J.N. Murdock, *Millimeter Wave Wireless Communications*, Prentice Hall Communications Engineering and Emerging Technologies Series from Ted Rappaport. Pearson Education, 2014.

[11] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-Wave Vehicular Communication to Support Massive Automotive Sensing," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 160–167, December 2016.

[12] K. Sato and M. Fujise, "Propagation Measurements for Inter-Vehicle Communication in 76-GHz Band," in *2006 6th International Conference on ITS Telecommunications*, June 2006, pp. 408–411.

[13] "Toyota Bringing Advanced ITS Technology to Mass market Models," `https://newsroom.toyota.co.jp/en/detail/9676551/`.

[14] "Volkswagen Will Enable Vehicles to Communicate With Each Other as From 2019," `https://www.volkswagenag.com/en/news/2017/06/pwlan.html`.

[15] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun 2010.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich Feature Hierarchies For Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Neural Information Processing Systems (NIPS)*, 2015.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016.

[20] A. S. Glassner, *An Introduction to Ray Tracing*, Academic Press Inc., 1989.

[21] Michel Treisman, "Combining Information: Probability Summation and Probability Averaging in Detection and Discrimination," *Psychological Methods*, vol. 3, pp. 252–265, 06 1998.

[22] Gideon P. Stein, Ofer Mano, and Ammon Shashua, "Vision-based ACC with a Single Camera: Bounds on Range and Range Rate Accuracy," in *IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No.03TH8683)*, 2011, pp. 120–125.

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[24] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] "GitHub TensorFlow Detection Models," `https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md`.