Multiple Computer Vision Engines: How Mobileye and Tesla are Tackling 3D Perception

By Mohamed Isse

During Tesla Autonomy Day early last year and during the Mobileye press conference early this year, we got to peek behind the curtain at the approaches and ideas underlying the technology of the two most prominent ADAS suppliers in the world. These two companies have mirrored each other in a lot of ways, with both looking to take on the Robotaxi nominal leaders by developing their own robotaxi solutions, from the comfortable position of having significant cash flow and growth potential coming from their sales in the ADAS and level-2 worlds. Both also overlap in that they emphasize cameras above all else. This article covers the Autonomy Day presentation, the Mobileye press conference, and how it relates to these companies' different takes on solving the problem of creating 3D models of other road users.

Mobileye's Approaches

Mobileye's presentation was largely on the components and approaches behind their all-camera self-driving car. It shed light on their desire to use a two-approach perception system to develop a high confidence environmental model that their vehicle can use to operate safely in superhuman margins of error. This two-approach perception system includes end-to-end independent self-driving systems, one comprised of camera only, and the other comprised of Radar and LiDAR. This independent dual system where camera never interacts with LiDAR and Radar is unique in an industry that has embraced sensor fusion at the system or even sensor level. Mobileye reasons that by getting near-human level capabilities out of both systems, the two systems together will multiply to get super-human levels of performance.

Mobileye focused on their camera-only approach during their press conference and even released a 30-minute driving demo of their camera-only self-driving car. They also went into great detail about the camera-only system during this presentation, and we learned that it mirrors the modular approach of Mobileye's general self-driving car, with multiple computer vision perception engines running simultaneously to develop an environmental model for the vehicle. These independent perception engines are then fused together at a late stage to give robust modeling of other road users. These independent computer vision engines include an engine for 3D neural-network-based object detection, neural-network-based scene segmentation, classical computer vision

based full-image object recognition, neural-network based wheel detection, and neural-network based VIDAR. These independent engines are added together after the fact to deliver a high confidence environmental model. These engines are pictured below.

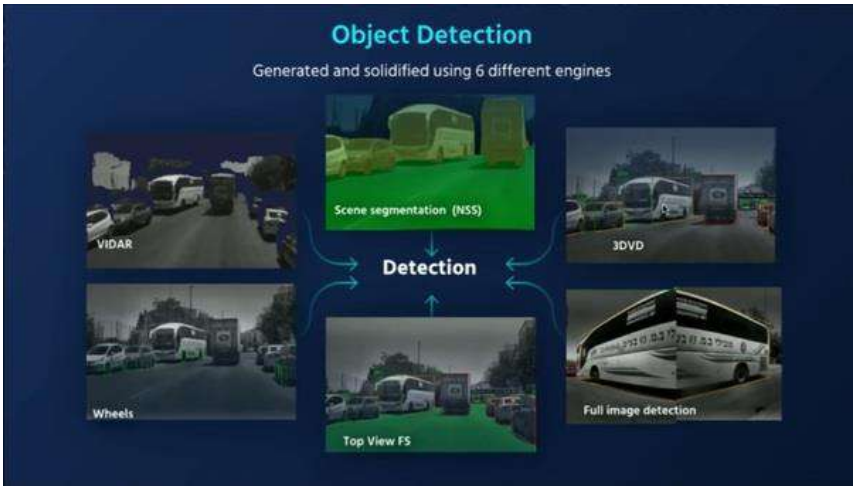

Image Source: Mobileye

Aspects of the System:

- **3DVD or 3D Object Detection**: is the common approach of predicting a 3D bounding box using a neural network from the 2D image.
- **Scene Segmentation**: is a neural network-based approach to break the scene down into free space and semantic objects (vehicle, sign, pedestrian, etc.), the pixels containing an object are treated as a detection.
- **Full Image Detection**: deals with the common scenario where an object is extremely close and can only be partially seen. Object classification using visual signatures is used to track close up vehicles in this scenario.
- **Wheel Detection**: is an engine that detects wheels of vehicles. This appearance-based engine detects vehicles from this reliable feature.
- **VIDAR**: is an approach that uses multiple views of the environment to triangulate and create a depth map. That depth map is converted into a point cloud and LiDAR based processing algorithms are used on that point cloud.

Mobileye also described the different methods used to attain the depth of an object, after it has been detected and classified. This is pictured below.

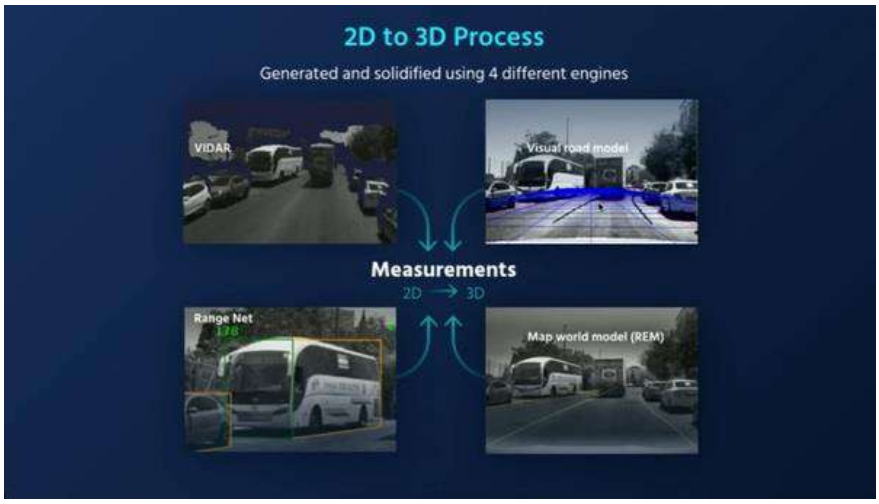

Image Source: Mobileye

The four approaches mentioned are:

- **VIDAR**: This approach mentioned earlier under object detection naturally gives a 3D estimation of an object's position. This is achieved using triangulation from multiple views of the scene.
- **Visual Road Model**: Using classical computer vision to estimate the ground plane and extract depth from how far vehicles are on this ground plane.
- **Range Net**: A neural network-based approach to inferencing the range of any object from its appearance in an image. This neural network was trained using a range-based sensor like LiDAR, radar or ranging (stereo, RGBD) camera.
- **HD Map (REM)**: Using road models from the HD maps generated by Mobileye REM to estimate the ground plane, and extract depth from how far the vehicle is on the ground plane.

Tesla's Approaches

Tesla's autonomy day presentation showcased aspects of their system in a similar way, with an in-depth technical look at Tesla's approaches to forming a high confidence environmental model. Tesla also used this presentation to boost their robotaxi efforts, with the announcement of a Tesla ridesharing network planned in 2020. A very intriguing part of the presentation however was the portion allocated to Andrej Karpathy, their Autopilot Computer Vision director.

As background to the presentation, Tesla has had problems as of late with their 3D environmental modeling. Two of their major accidents that received national attention were of the Autopilot system impacting an 18-wheeler semi-truck. These accidents in Florida and Ohio were eerily similar in that the Autopilot system failed to detect a truck trailer positioned laterally across the road.

Image Source: Electrek

This is a major challenge for a computer vision and radar system to detect. Notice the similarities to a billboard positioned above the highway. A radar has no vertical resolution in most cases, so it is unable to distinguish between this sort of object and a billboard above the highway.

What this task comes down to is how well a computer vision pipeline can estimate an object's position and dimensions in 3D. This explained why Karpathy's presentation focused heavily on 3D perception in general and depth perception in particular. Given that a high performing camera-only perception system is vital in the semi-truck trailer case mentioned above (as radar can't distinguish it from a billboard), Andrej broke down at a high level the different engines that Tesla was looking at and using to create 3D image-based models of the world. The approaches mentioned were stereo-based depth estimation, structure from motion, monocular depth estimation and unsupervised depth estimation.

- **Stereo based depth estimation**: An approach of using multiple cameras to triangulate the 3D position of objects in the environment.
- **Structure from motion**: An approach using subsequent images from the same camera to triangulate in the same manner as with stereo based depth estimation.
- **Monocular depth estimation**: This estimation is created from using the radar to annotate depth on camera images, so the labeled images can then be used to train a monocular depth neural network.
- **Unsupervised depth estimation**: This is an approach to train a neural network to predict depth by using the consistency of depth estimations from subsequent frames as a training signal.

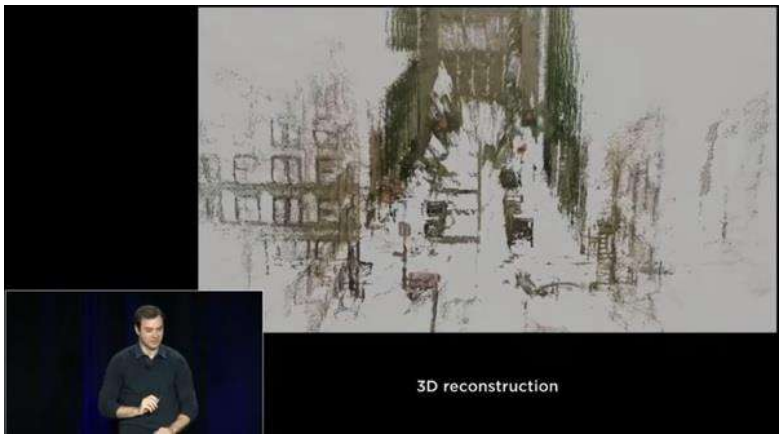

Image Source: Tesla

Comparing Approaches between Tesla and Mobileye

From a technical standpoint, the more interesting of the two presentations in regard to 3D object detection was certainly Mobileye's, which went into depth on the different computer vision engines that their team has integrated into their all-camera self-driving car platform. Tesla's presentation didn't get to the bones of their system and was more aimed at making the case for cameras as a ranging sensor replacing the use of a LiDAR.

Regardless of whether Tesla mentioned many of the computer vision engines Mobileye discussed, it is very possible that Tesla is either researching or has already incorporated similar techniques into their Autopilot product. Another important note is that some of the techniques Mobileye mentioned may not actually end up being used in their current or future autonomous platforms. Mobileye was also presenting these ideas in the context of their autonomous all-camera platform, and not their common ADAS EyeQ series.

Tesla's Autopilot problem areas are well-known given that hundreds of thousands of people are actively testing out the new features they are releasing. One of the Autopilot platform's greatest weaknesses seems to still be in estimating the depth of uncommon road users. Even into 2020, there have been crashes reported with the Tesla Autopilot system even in well-understood situations like Adaptive Cruise Control.

Image Source: Twitter User AlanZavari

Already in 2020, a crash was reported in which a Tesla Autopilot system accelerated into a carrier style of vehicle. This recent crash seems to be attributed to the Autopilot system improperly estimating the size and depth of a carrier vehicle. We have no strong insights into how exactly the object detection within the AP system is functioning, but some educated guesses are that this collision was an issue with determining the range of objects hovering above the ground plane, it's also likely there were some elements of the system having difficulties detecting and classifying objects from very close up (such as can be seen in figure 7, another image of the collision). Both will be discussed further below.

Mobileye's presentation on their multiple semi-independent object detection engines struck at the heart of some of the challenges Tesla's Autopilot has been facing.

To tackle the issues with perceiving the depth of objects that are hovering above the ground such as the 18-wheeler trailers (from Figure 3) in the fatal collisions in 2019 and 2016, Mobileye presented their use of depth perception using ViDAR. This a technique where stereo-based depth estimation or structure from motion is used to create a depth map, and the depth map is converted into a point cloud (very similar structurally to a LiDAR point cloud) which is then fed into LiDAR processing algorithms. Shashua explained that this ViDAR approach is tailored for exactly the use case where there is a hovering or protruding object of some sort (see below).

Image Source: Mobileye

Mobileye CEO Shashua also discussed the difficulties inherent in detecting objects at extremely close ranges, which may also have been a contributing factor in the recent 2020 crash. Any given camera on the vehicle may not get a full view of the object. The Tesla 2020 crash video showed the moments before the crash, that vehicle was seen from an up-close perspective which would pose a challenge for any usual appearance-based object detection system. Also from this image (Figure 7) below, one can see a lot of overlap with the rear-protruding objects, which Shashua mentioned VIDAR would be a perfect use-case for. This carrier vehicle in Figure 7 has several protruding limbs just like the example vehicle in Mobileye's presentation in Figure 7.

Image Source: Twitter User AlanZavari

Mobileye presented an independent computer vision engine to deal with the up-close vehicle scenario as well which is the full image detection module mentioned earlier.

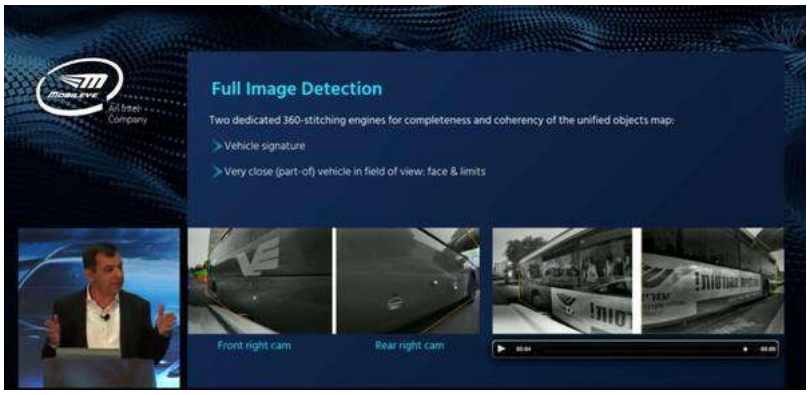

Image Source: Mobileye

This full image detection module uses an appearance-based method that breaks an object detection down into a 'signature' vector which represents its features in a compact way. This is what is used to do near object detection when no one camera can capture more than a small part of the object. The footage from the Tesla which crashed showed the inherent challenge of detecting an object from an extremely up-close sensor perspective, and it is unclear whether or not Tesla Autopilot uses VIDAR or near image object detection approaches like those mentioned by Mobileye.

Further Discussion

We cannot say conclusively whether Tesla incorporates most of what the Mobileye CEO presented. It's very possible that within Tesla Autopilot there is some aspect of VIDAR-based approaches and full image detection. The fact of the matter is that in the 2020 crash, the Autopilot version on the car was not capable of properly detecting the 3D position of the carrier vehicle. If Tesla aims to become a robotaxi maker, they will need to address this shortcoming while maintaining CEO Elon Musk's promise of not incorporating any new sensors into future versions of their platform. There are some aspects of the Mobileye system that are unique from anything Tesla has within their platform. For example, the use of HD maps to estimate road geometry. This was mentioned by Mobileye as one of the ways that they ascertain the depth of an object within their environmental model. Late last year, Musk confirmed that Tesla was not using HD maps, precluding them from using this sort of depth estimation approach.

There is also the approach Mobileye is using to differentiate themselves. Tesla has emphasized that they want to take the current Tesla sensor complement, mostly without change, to robotaxi levels of autonomy. Mobileye on the other hand, wants a fully capable self-driving system using camera alone, and another independent system using LiDAR and radar only. The assumption is that the sum of these two systems is another system that gets to a significantly higher level of performance. This is not necessarily how things will work, and the devil is in the details for any kind of system-level

fusion by Mobileye. They didn't go into detail on their LiDAR-Radar only system nor how it would be fused with the camera-only system. The Tesla approach to Robotaxis will almost certainly not involve LiDAR in any capacity in the foreseeable future, and this is another way that Mobileye and Tesla have diverged in approaches.

Mobileye released a 20-minute unedited ride in their autonomous vehicle where it drove on crowded regular streets in Jerusalem. The ride was certainly impressive and featured various complex driving scenarios including unprotected left-hand turns, highway merges and complex yielding situations. It is very important to stress is that this all-camera autonomous system has never been subjected to the sheer scale of situations that Tesla Autopilot has, and it is very likely that the Mobileye system in its present form would encounter many situations in which it would currently break down. In any case, the Mobileye press conference demonstrated the careful thought towards redundancy and performance Mobileye has put into their all-camera autonomous system and gave useful insights into how this sort of system could be done. Whether anything in that presentation was of interest to the Tesla team is pure speculation, but there is no doubt that these companies will certainly be watching each other closely.

About VSI Labs

Established in 2014 by Phil Magney, VSI Labs is one of the industry's top advisors on AV technologies, supporting major automotive companies and suppliers worldwide. VSI's research and lab activities have fostered a comprehensive breakdown of the AV ecosystem through hands-on development of its own automated vehicle platform. VSI also conducts functional validation of critical enablers including sensors, domain controllers, and AV software development kits. Learn more about VSI Labs at https://vsi-labs.com/.

How to Engage with VSI

VSI Labs offers subscription research packages to meet your needs:

- VSI Insights – High level technical analysis of CAV technologies and the future of automated driving.
- VSI CAV Technology Databases – Deep insights on the products and technologies that make up automated driving.
- VSI Pro – Decomposition of an AV's functional domains and time saving instructions on how to build an AV.

Learn more about our portal services or contact us to get started!

Conditions of this Report

In no event shall VSI Labs (a.k.a. Vision Systems Intelligence, LLC.) be liable to you or any other party for any damages, losses, expenses or costs whatsoever (including without limitation, any direct, indirect, special, incidental or consequential damages, loss of profits or loss opportunity) arising in connection with your use of this material. The information in this report is assembled on a best-efforts basis and VSI Labs shall not be responsible for errors or omissions and any claims arising from this.