Self-driving vehicles: Key technical challenges and progress off the road

Michael Milford, Sam Anthony, and Walter Scheirer

MINE SHAFT—©ISTOCKPHOTO.COM/POWEROFFOREVER, TRUCK—©ISTOCKPHOTO.COM/YURIY BUCHARSKIY

In a period of fewer than 10 years, the quest for self-driving vehicles, also referred to as *autonomous vehicles* (*AVs*) or *driverless cars*, has become one of the biggest technology races in the world, with tens of billions of dollars poured into companies and start-ups. The goal is an on-road, consumer-driverless car: whether owned by individuals or part of a centralized ride-sharing fleet, this is the area where the majority of investment has occurred. However, AVs have been around for much longer in other fields, such as mining, which share some but not all of the same technical challenges faced by on-road AVs. In this article, we provide an overview of the key technical challenges and solutions for both on- and off-road AVs, with a focus on one of the key unsolved challenges—interaction with vulnerable road users (VRUs).

Key technical competencies: Hardware

A typical AV contains a number of key components: the physical platform and a suite of sensing and onboard computational hardware.

Platform

The type of AV platform affects the viability of different technological

Digital Object Identifier 10.1109/MPOT.2019.2939376
Date of current version: 31 December 2019

solutions to the problem of autonomy. Larger vehicles are typically heavier and harder to stop—and more damaging when they hit something—but they can carry a greater number of onboard sensing and computing components. Energy storage also generally scales favorably with vehicle size, an important consideration that can enable better up-time percentages and utilization of more power-hungry computing.

Sensor suites

AV platforms have access to a range of sensing technologies (Fig. 1).

FIG1 A typical sensor suite on top of a car, with multiple cameras and lidar sensors. (Source: QUT; used with permission.)

Lidar- and laser-based range sensors provide accurate long distance range-to-object information and can also use reflectance information to detect lane markings. However, in adverse weather, such as rain or smoke, their capabilities can be significantly degraded.

Modern camera technology provides very-high-resolution imagery of the environment, with good dynamic range (revealing detail both in bright and dark areas of an image simultaneously) and high frame rates (Fig. 2). The information present in a camera image is much richer than that produced by any other sensing modality, provided it can be successfully extracted—the widely quoted proof of concept here being that humans can drive very well with primarily visual sensing alone. Cameras are often less expensive and require less power than lidar, but they are sensitive to changes in environmental appearance caused by factors such as day–night cycles (Fig. 3).

Radar's primary purpose in most current AV applications is collision

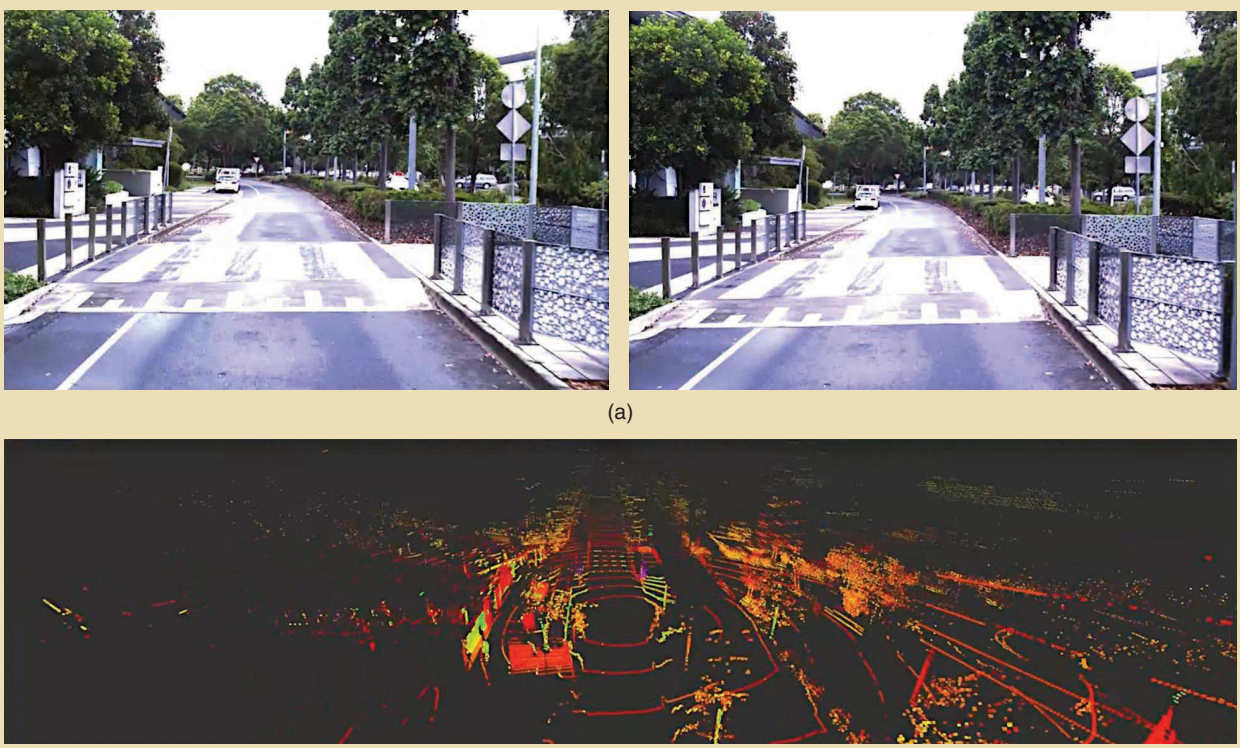


FIG2 What an autonomous car sees: (a) a range of camera views and (b) range scans. (Source: QUT; used with permission.)

avoidance: although it does not have good acuity and, as a result, struggles to distinguish small objects, it is relatively resilient to environmental conditions, such as adverse weather, and can see through smoke and fog quite well. Finally, sensors such as GPS receivers provide positioning information (which can be disrupted by tunnels or tall buildings), while internal sensors deliver information such as linear acceleration, rotational rate, steering angle, and wheel speed.

Computational hardware

Computer hardware provides the processing power to perform all of the onboard autonomy-related tasks, such as scene understanding, navigation, and high-level control. To maximize electric-vehicle range, recent hardware trends have focused on power usage per computing unit. Nvidia is a good example of a key player in this space, with power-efficient, highly capable systems such as its Jetson AGX Xavier. Offboard computing still has a useful role to play in AV applications—for example, in the consolidation and merging of the massive amounts of data uploaded by thousands of cars in a city on a daily basis.

Key technical competencies: Software

The software operating on AVs performs a number of key technical competencies, including localization, planning, decision making, and scene understanding.

The information present in a camera image is much richer than that produced by any other sensing modality, provided it can be successfully extracted.

Mapping and localization

Mapping and localization are key pillars of AV operation. There are several subtypes of localization, each of which plays a different role in enabling autonomy on a vehicle (Fig. 4).

Simultaneous localization and mapping has long been a major research field in robotics: how does a robot move through an environment, building up a map of that environment, while simultaneously localizing itself within that everchanging map? Approximate localization—what you get on your phone's GPS—is typically used for overall route planning and is obtained from GPS or onboard localization systems. Automation-enabling higher-precision localization is typically provided by onboard localization within existing maps of the environment or, in the case of some autonomous mining vehicles, high-accuracy GPS.

Relative localization is also important—for example, knowing that the vehicle is currently located 0.73 m from the edge of the road. Accurate relative positioning (and velocities and accelerations) with respect to moving objects, such as an oncoming car, is critical for safe vehicle planning and control.

Planning, decision making, and control

Just as critical to an AV's viability as sensing and mapping is what is then done with that information: how does the vehicle plan and then act, whether to accelerate, brake, turn, or activate a turning indicator? These processes play a critical role in safety; the planning system must continually plan safe actions, such as slowing down or suddenly changing lanes to avoid an unexpected obstacle when braking is not an option.

The planning and decision-making process also changes significantly for on-road delivery vehicles that carry goods rather than people, such as those used by Nuro. In accident situations with these AVs, there is no tension between protecting humans inside and outside the vehicle, so the safety of humans outside can be entirely prioritized.

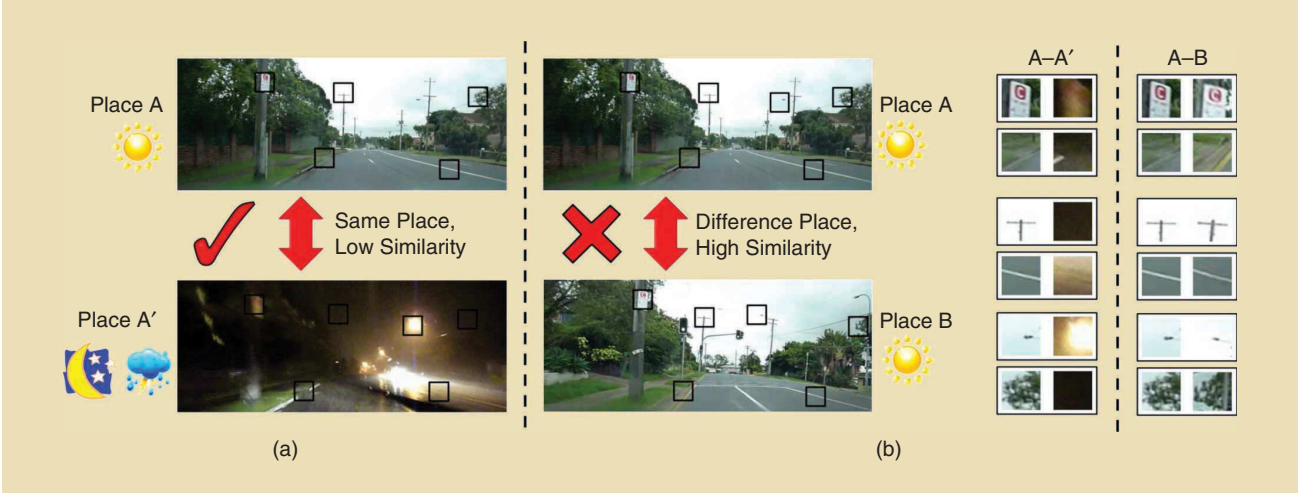


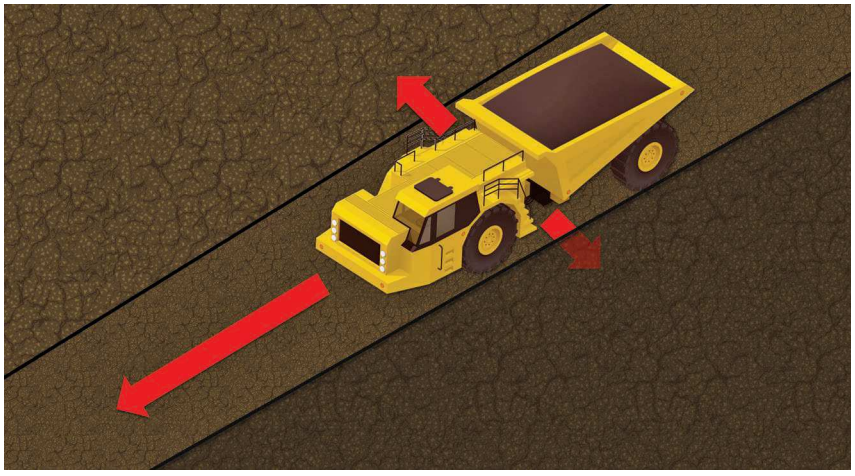
FIG3 Environmental change poses a significant challenge for camera-based technologies. (a) The images are from the same place under radically different environmental conditions, whereas in (b), the images are from different places in the environment.

VRUs are sometimes difficult to distinguish: from the waist up, a cyclist, pedestrian, and scooter rider all look very similar to a computer vision system.

Interaction with VRUs

Vehicles that have reached Society of Automotive Engineers level 3 autonomy and above will have to know how to interact with humans. This includes human drivers, who—even in a world where AVs are rapidly adopted—will be on the roads for the foreseeable future. Bicyclists, pedestrians, motorcyclists, scooter riders: these categories of VRUs have enduring claim to their share of the urban pavement. Interacting safely, explainably, and politely with VRUs is likely to remain an essential part of the AV's task (Fig. 5).

Pedestrians and cyclists are not predictable with standard techniques, such as Kalman filters. Simply stopping every time a VRU could potentially enter the vehicle's path results in vehicles that perform excessive and unnecessary emergency maneuvers. Overall, 86% of documented incidents with AVs are either rear-endings or sideswipings that result from a human's misunderstanding of an AV's behavior. Understanding VRUs is key to eliminating this failure mode.



TRUCK © 0MELA/TELES2/123RF.COM

FIG4 All errors are not created equal. For example, for second-to-second control in a mining tunnel, minimizing lateral error is more important than downtrack (along the length of the tunnel) error since the immediate risk is hitting the wall.

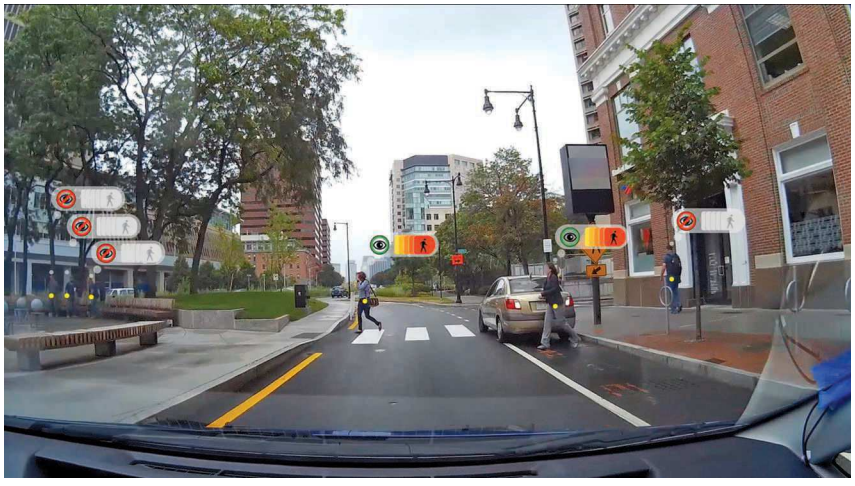


FIG5 Detecting and predicting the intent of VRUs, such as cyclists and pedestrians, is a critical challenge for AVs. (Source: Perceptive Automata; used with permission.)

Moving away from the trolley problem mind-set

Much of the attention devoted to interactions between AVs and VRUs has focused on ethical dilemmas. A famous thought experiment is the "trolley problem," where, in the eponymous problem, a trolley driver is forced to choose which of two actions, both of which cause someone's death, is more morally acceptable; this has been held up as a model for the kinds of decisions AVs will have to make. Although it may someday be the case that AVs are sophisticated enough and have good enough information about the world that the primary concern with VRU interaction is how to behave ethically in the unlikely event that there is no option but to cause catastrophic bodily harm to a human, there are several reasons why this is not currently a primary concern to AV makers.

First, the starting goal for many vehicle makers is finding a motion plan that provably minimizes or eliminates any chance of a harm-causing interaction. The Intel division MobilEye has published work attempting to formalize risk analysis in motion planning to develop behavior plans where a negative interaction is impossible. Second, the types of ethical dilemmas discussed in most trolley-problem research rely on very fine-grained categorization of VRUs—an old person versus a young person, a pregnant woman versus a helmetless cyclist, and so on—that are largely out of reach for current perception systems in AVs. Third, much of the current focus in AVs is on minimizing harm in general, and one way to do that is to plan around the level of damage likely to be caused. For these reasons and others, the ethical considerations raised by the trolley problem are increasingly not being considered as the most immediate practical challenge for AVs.

Key technical breakdown

VRUs are sometimes difficult to distinguish: from the waist up, a cyclist,

pedestrian, and scooter rider all look very similar to a computer vision system. Here, we provide a technical breakdown of the relevant technologies that address this challenge.

If you know the location and trajectory of the pedestrian, how well can you extrapolate his/her future trajectory?

Detecting VRUs

When we use the term *detection*, we mean automatically detecting that something is in the way of the vehicle. Radar, lidar, and various other non–red/green/blue sensors are very capable for this task, but they have limitations around discriminatory resolution (radar) and current cost (lidar). Camera sensors represent a less expensive option that has caught on with several car companies producing self-driving vehicles.

Recognizing VRUs

When we use the term *recognition*, we mean automatically determining what exactly is in the scene as sensed by the vehicle (Fig. 6). Current recognition algorithms have been largely developed and tested in the laboratory by using standard benchmark data sets, such as ImageNet and Common Objects in Context (better known as *COCO*), but a troubling disconnect exists between laboratory experimentation and real-world operation. For example, a well-known limitation of all machine-learning-based algorithms is their poor handling of inputs from classes outside the training set. This is known as the *open set recognition problem*, and it is a common problem in autonomous driving. A new class of open-set-tolerant machine-learning algorithms is being developed to address this issue.

Recognizing VRU actions

After detection and recognition of a VRU comes activity recognition—what the VRU is doing. Take one scenario: traffic officers signaling cars to follow a detour by waving in a certain direction. With a correct determination of what the officer's action means, the vehicle can alter its course and safely proceed as directed.

This is a nontrivial sequence of events that must unfold within seconds and be executed with a level of accuracy that matches that of a human driver. As with other areas of visual recognition within computer vision, great strides have been made in action recognition, but current approaches are not as robust as human drivers.

Predicting VRU actions

Arguably, the most important aspect of interacting with VRUs is prediction. A motor vehicle that is traveling straight at 25 mi/h on a road and does not have its brake lights illuminated can be assumed to continue traveling at approximately 25 mi/h, at least momentarily. Compared with vehicles, VRUs have many fewer constraints in terms of traffic signals, other traffic, and rules of the road and, therefore, have much more variability in potential paths.

Much work on the prediction of VRU actions has relied on fundamentally physics-based models: If you know the location and trajectory of the pedestrian, how well can you extrapolate his/her future trajectory? Elaborations have included the use of cues, such as the presence of relevant context like crosswalks, and the integration of information regarding the pose of the person. These approaches have proven to be relatively robust on very short time scales, but they have not been able to provide useful predictions outside of a time window of about 1.5 s. At normal urban driving speeds, that's not enough. One proposed solution is to model the dynamics of all of the actors at an intersection, which critically relies on being able to accurately model every agent in the scene.

Almost all of the current approaches have another shortcoming: that the drivers with which VRUs are most comfortable interacting—humans—do nothing like either of these approaches. Humans have a finely tuned and remarkably high-functioning facility called *theory of mind*, which allows them to make behaviorally useful assumptions about the internal mental state of another human. A human driver isn't trying to guess the trajectory of a pedestrian; instead, he or she is making sophisticated inferential judgments about what that pedestrian's goals are and how that pedestrian might interact in a social process with the vehicle. Approaches that model this concept look promising.

Communicating car intent to VRUs

The interaction between VRUs and human-driven vehicles begins when

FIG6 Reliably detecting and recognizing VRUs, such as cyclists, is difficult enough under normal conditions but is compounded in poor visual conditions and when the VRU (indicated by the red box) is partially obscured by other objects in the environment.

The pedestrian wants to know that the car recognizes that he or she is there, the car seeks to know what the pedestrian wants, and so on.

either the driver or the VRU first notices the other and ends when the vehicle has proceeded out of the VRU's field of view. It is bidirectional: the pedestrian wants to know that the car recognizes that he or she is there, the car seeks to know what the pedestrian wants, and so on. Companies, such as Jaguar/Land Rover, have experimented with mounting large, cartoon eyes on vehicles to communicate information about how the AV is distributing its "attention." Former startup Drive.ai designed its AVs to feature interactive screens that can communicate more complex messages, such as "I'm waiting for you to cross." These systems have limited grammar, but actual interactions between human drivers and VRUs also rely on very limited grammar. To communicate with limited grammar, the ability of both VRUs and vehicles to understand the intentions and state of mind of other road actors is essential.

Current technical issues

With the field maturing over the past 15 years since the DARPA Grand Challenge AV competition in 2004 (which is widely credited as catalyzing the modern AV technology race), it has become relatively clear that some key technical issues remain unsolved, and these are generally widely acknowledged by both industry and researchers working in this area. One of the most significant topics is interactions with VRUs, which we have already covered. Here, we briefly highlight some of the other challenges.

The problem of corner cases

Corner cases, as they have become known, are situations that rarely occur and, as a result, are hard to predict, anticipate, and react to appropriately. A person dressed in a chicken suit is one example of a corner case. For self-driving cars, the problem is particularly difficult because the current artificial intelligence techniques behind these systems do not generalize as well as a human driver and have difficulty coping with these highly unusual situations. Consequently, much effort is being invested in coming up with ways to deal more effectively with these corner cases by gathering ever-larger amounts of data from the real world, simulating billions of miles of driving, and repeatedly testing pathologically difficult scenarios.

Simulation versus real-world testing

A key issue for AV developers is that cars are already quite safe: there is approximately one fatality for every 100 million miles of driving. Consequently, it is very difficult under normal conditions to obtain sufficient mileage on a limited number of development vehicles to prove the safety of a system. Therefore, developers have turned to simulation as a critical tool in their autonomy arsenal. High-fidelity simulation environments enable researchers to target specific weather conditions and pedestrian configurations and run much-higher-throughput simulation and evaluation than is possible in the real world.

A key challenge in using simulation arises from the transferability problem: how do you show and prove that the system you have developed in simulation will work as well in the real world? Simulation is never a perfect replication of reality. Many resources and much effort have consequently been invested in improving the utilization and transferability of development in simulation environments.

Sometimes versus anytime: Weather and other environmental conditions

The real world is a constantly changing environment, which presents major challenges for AVs. First and foremost, the environment can change in both appearance and physical structure due to day–night cycles; seasonal change; and weather conditions, such as rain, snow, and fog.

Figure 3 illustrates some of the key challenges that a changing world can cause. The same place [shown in the left column of Fig. 3(a)] can appear completely different at night during a tropical storm versus clear weather in the daytime. The problem is further complicated by the natural environmental aliasing that can also be encountered, shown in the left column of Fig. 3(b): these are two places that are completely different locations but look highly similar.

These problems can be partly solved by using advanced methods or sensors that are not as sensitive to appearance change, such as lidar. However, visual sensing is critical for the rich, nuanced understanding of the world around an AV, and, consequently, the problem of operating in challenging visual conditions remains relevant and unsolved.

Provability, explainability, and self-characterization

A significant shortcoming of the present generation of self-driving vehicles (and deep learning in general) is the difficulty in describing the properties of their underlying deep-learning models in a rigorous manner. In essence, the learning problem during training is one of function approximation, where the approximated function cannot be recovered in an exact manner afterward. (This is why neural networks have a reputation of being black boxes.) We would like to be able to enforce explainability for any output of a deep-learning model, but since we cannot examine any learned functions directly, we can only turn to the observable output of the system—the same situation psychologists find themselves in when studying the human brain. One possibility, then, is to test the deep-learning models in a manner similar to how psychologists test the brain.

For some applications, pausing and handing off control to a human operator

is feasible, but only if the system is able to assess its own performance reliably. To do this, probabilistic outputs reflecting uncertainty are required. For deep-learning-based systems, this can be accomplished with strategies, such as making small perturbations to the weights of the network, dropping out units of a trained network at test time, using a probabilistically calibrated readout layer, or examining statistical distributions of the data sampled by the sensors. The choice of distribution is important: underestimating the occurrence of rare events can be dangerous, but overestimating them may be problematic for usability.

Underestimating the occurrence of rare events can be dangerous, but overestimating them may be problematic for usability.

AVs beyond the road

Beyond the road, AVs have been or could be deployed in a range of other domains, including mining, logistics, agriculture, and defense. Here, we briefly cover the key deployment domains and their unique problems and opportunities.

Mining

Mining, in general, has several of the key characteristics that facilitated its early adoption of AVs: it is large enough to support the capital-intensive development of AV-related technology, its existing remote operation workflows are more easily automated, and there are fewer latency-critical scenarios, meaning that occasional handover to a remote operator is feasible. One example milestone in AVs in mining is Rio Tinto's autonomous haulage system, which recently hauled its one billionth ton autonomously.

Mining is a challenging environment (Fig. 7). Underground, there is no access to satellite-based GPS, so alternative technological solutions are required: some involve installation of additional infrastructure, local Wi-Fi networks, or on-vehicle camera- and laser-based localization solutions. Onboard camera-based solutions encounter a range of challenging perceptual conditions: dust, smoke, water, and highly varied lighting conditions. Range-sensor-based solutions encounter a different set of challenges, including the highly aliased geometry of many underground tunnel systems.

Logistics

It is possible to design an entire logistics center to facilitate higher levels of automation. Amazon's fulfillment centers, built on top of its acquisition of Kiva Systems, are a prime example of this: the autonomous robots move shelving around rather than attempt to pick things off static shelves. Other approaches, such as Ocado's, involve a rigid square lattice on which robots move around, picking up and dropping off grocery loads. In both cases, humans are restricted to certain areas of the environment, so human safety issues are significantly reduced as a technological concern.

Agriculture

Farms generally have relatively controlled access and minimal to no human presence in the operational

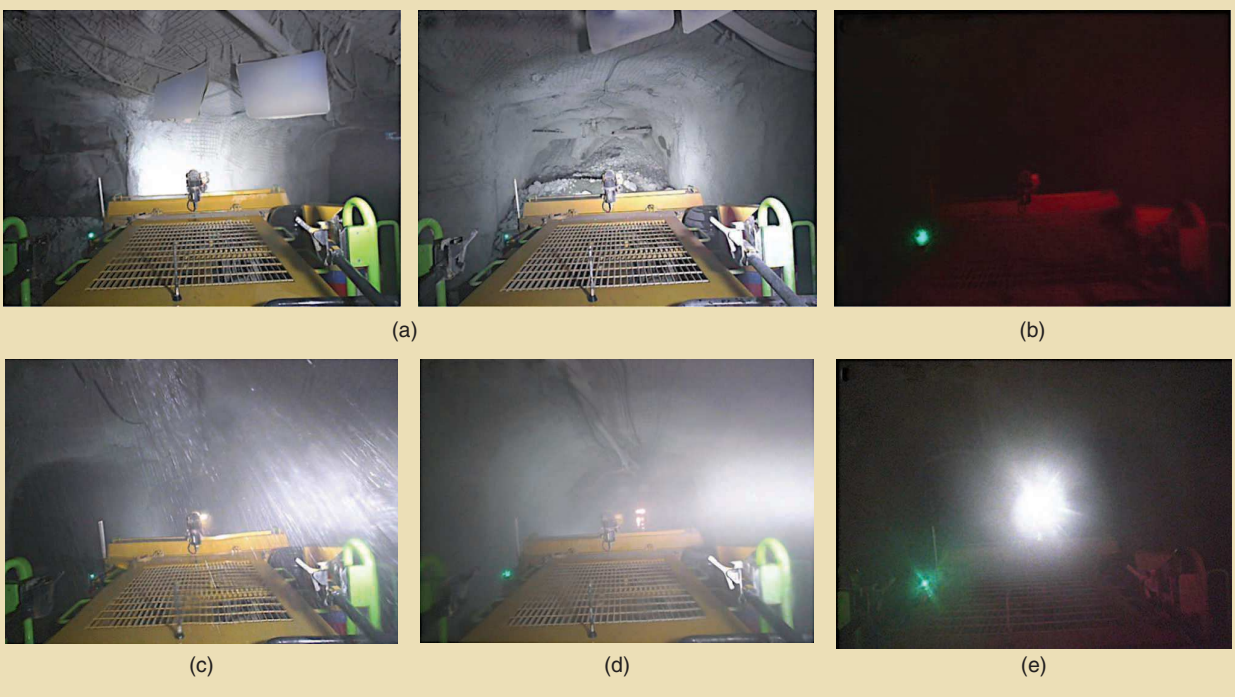


FIG7 In underground mining environments, a range of challenging perceptual conditions are encountered by AVs, including huge lighting changes, darkness, water, and dust: (a) clear images, (b) low light, (c) water, (d) dust, and (e) glare. (Adapted from Zeng et al.)

Beyond the road, AVs have been or could be deployed in a range of other domains, including mining, logistics, agriculture, and defense.

zone of an AV. In addition, it can sometimes be hard to find people to fill some labor roles, further motivating the case for developing AVs. Autonomous farming vehicles can perform a range of activities, including sowing and planting crops, killing weeds, and the long-term holy grail: harvesting crops. Progress has been slow: although there have been dozens of AV trials, there are few long-term commercial deployments (Fig. 8). Most of the more capable platform demonstrations have been announced only in the past 2–3 years.

Defense

In defense, as in mining, the cost per unit of many vehicle types is typically far larger than that of a normal consumer car, enabling the use of more capable sensing and computing. Much modern defense theory assumes that there will be a complete blackout on both communications and GPS-based positioning technologies (similar to the conditions imposed on underground autonomous mining trucks), meaning that on-vehicle autonomy will have to shoulder the bulk of the decision making rather than relying on outsourcing to a human at a remote command post.

The environments that these vehicles might deploy into, such as ruined, dusty, or smoking urban landscapes and thickly vegetated forests, pose a range of challenging mobility, perception, planning, and control challenges. Finally, there are also the ethical considerations around autonomy in any defense application, which are receiving significant sustained attention.

Other fields

There are almost 40 marine ports that are at least partly automated globally, and some of those autonomous components involve AVs, for example, shifting shipping containers around. Other areas of AV deployment include sidewalk-based delivery vehicles, such as Amazon's Scout program and Starship technologies. These vehicles are typically relatively small and inexpensive, and they move at relatively low speeds, radically reducing their danger profile compared to on-road larger vehicles moving at higher speeds.

Conclusion

AV-enabling technology has matured and advanced significantly over the past decade in a range of domains, including on-road passenger-carrying or delivery vehicles, mining, and logistics. In some application areas, such as logistics and mining, these vehicles already form a commercially critical part of the companies that operate them, whereas in others, most notably on-road AVs, widespread commercial deployment has not yet occurred.

Much of the core technology is likely to continue benefitting from steady progress in sensing and computing capabilities (along with a corresponding decrease in price) and the associated progress in vital technical capabilities, such as general scene understanding and VRU interaction. In fields where safety is not directly involved, such as those where humans are physically absent from the operating environment of AVs, future progress will likely be determined by simple commercial calculations based on the cost and efficiency of AV systems. However, there remain key technical hurdles to overcome with respect to safety for widespread on-road deployment, which will make for interesting years ahead.

FIG8 Agriculture shares many of the same motivations for AV use as mining, but widespread commercial deployment has lagged. (Source QUT; used with permission.)

Read more about it

- L. Alter, "$80 billion has been spent on self-driving cars with nothing to show for it," *TreeHugger*, Mar. 12, 2019. Accessed on: June 12, 2019. [Online]. Available: https://www.treehugger.com/cars/praise-dumb-transportation-self-driving.html
- O. Bawden et al., "Robot for weed species plant-specific management," *J. Field Robotics*, vol. 34, no. 6, pp. 1179–1199, 2017. doi: 10.1002/rob.21727.
- T. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, and W. J. Scheirer, "Learning and the unknown: Surveying steps toward open world recognition," in *Proc. AAAI Conf. Artificial Intelligence (AAAI 2019)*., vol. 33, pp. 9801–9807.
- J. Torresen, "A review of future and ethical perspectives of robotics and AI," *Front. Robot. AI*, Jan. 15, 2018. Accessed on: June 12, 2019.

[Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2017.00075/full

- G. Barber, "Shark or baseball? Inside the 'black box' of a neural network," *WIRED*. Accessed on: June 12, 2019. [Online]. Available: https://www.wired.com/story/inside-black-box-of-neural-network/
- J. Diaz, "Jaguar Land Rover is putting googly eyes on its autonomous cars," *Fast Company*. Accessed on: June 12, 2019. [Online]. Available: https://www.fastcompany.com/90231563/people-dont-trust-autonomous-vehicles-so-jaguar-is-adding-googly-eyes
- M. J. Milford and G. F. Wyeth. "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. 2012 IEEE Int. Conf. Robotics and Automation*, May 2012, pp. 1643–1649. doi: 10.1109/ICRA.2012.6224623.
- B. RichardWebster, S. E. Anthony, and W. J. Scheirer. "Psyphy: A psychophysics driven evaluation framework for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, 2019. doi: 10.1109/TPAMI.2018.2849989.
- "Rio Tinto's autonomous trucks pass one billion tonne milestone," Australian Mining, Jan. 30, 2018. Accessed on: June 5, 2019. [Online]. Available: https://www.australianmining.com.au/news/rio-tintos-autonomous-trucks-pass-one-billion-tonne-milestone/
- F. Chu, S. Gailus, L. Liu, and L. Ni, "The future of automated ports," McKinsey and Company, Dec. 2018. Accessed on: June 5, 2019. [Online]. Available: https://www.mckinsey.com/industries/travel-transport-and-logistics/our-insights/the-future-of-automated-ports
- M. Baram, "Why the trolley dilemma is a terrible model for trying to make self-driving cars safer," *Fast Company*, Feb. 19, 2019. Accessed on: June 12, 2019. [Online]. Available: https://www.fastcompany.com/90308968/why-the-trolley-dilemma-is-a-terrible-model-for-trying-to-make-self-driving-cars-safer
- F. Zeng, A. Jacobson, D. Smith, N. Boswell, T. Peynot, and M. Milford, "Enhancing underground visual place recognition with Shannon entropy saliency," in *Proc. Australasian Conf. Robotics and Automation*, 2017. [Online]. Available: https://eprints.qut.edu.au/116887/1/pap120s1-file1.pdf

About the authors

Michael Milford (michael.milford@qut.edu.au) is a professor at the Queensland University of Technology, Brisbane, Australia, and a chief investigator at the Australian Centre for Robotic Vision. His research interests include the neural mechanisms in the brain underlying tasks such as navigation and perception and how they can be used to develop new technologies in challenging application domains, such as all-weather, anytime positioning for autonomous vehicles.

Sam Anthony (santhony@perceptiveautomata.com) is an ex-hacker, vision scientist, expert in human cognition, and veteran of tech booms and busts. He is the chief technical officer and cofounder of Perceptive Automata, Inc., an early-stage company building safety systems for autonomous vehicles.

Walter Scheirer (walter.scheirer@nd.edu) is an assistant professor in the Department of Computer Science and Engineering at the University of Notre Dame, Indiana, and a cofounder of Perceptive Automata, Inc. His research interests include the fundamental problem of recognition, including the representations and algorithms supporting solutions to it.

P