Onboard Computer with a Sixth Sense

Emergency braking systems already prevent quite a few traffic accidents, but electronic assistants still have no proper overview of what’s happening on the road. Bernt Schiele, Director at the Max Planck Institute for Informatics in Saarbrücken, wants to change this. He teaches computers to anticipate the routes of vehicles and pedestrians.

Text Tim Schröder

They’re called Forward Alert, Pre-Safe or PreSense and can be found primarily onboard expensive vehicles. The systems with the pithy names are otherwise known as emergency braking assistants. If a child suddenly runs into the street or a vehicle speeds through a red light at an intersection, the assistant brakes autonomously – even when the driver is still paralyzed with shock. These systems scan the scene on the road in front of the car with small infrared cameras or radar sensors. The emergency braking assistant detects objects on a collision course in fractions of a second – for example, a pedestrian who is lost in thought when stepping into the street. The microprocessor analyzes the situation in milliseconds: How fast is the object? Will there be a collision? Caution. Stop!

Vehicle manufacturers emphasize that the emergency braking assistants prevent thousands of accidents every year in Europe alone. The technology works and is faster than drivers when they are in a state of shock, but at present it is still comparatively stupid. Emergency braking assistants and similar sensor systems in vehicles can react only in the final instant. They concentrate fully on an approaching object – but nothing more. Well-rested car drivers, in contrast, have the whole scene in their sights. The ambulance approaching from the right some distance away, for example, or a fast car that is briefly hidden by a truck at a multi-lane intersection and then races across the intersection. In short, emergency braking assistants can intervene at the last moment, but they can’t drive with anticipation.
Alert but stupid: Today’s conventional emergency braking assistants scan the road, as sketched here, and react to obstacles that suddenly appear. But they can’t anticipate the movements of other road users. The system developed by the researchers in Bernt Schiele’s group, in contrast, monitors the traffic in an anticipatory way.
The same applies to other automatic systems – robotic assistants, for example. They are now quite capable of taking a tray from the kitchen table and carrying it into the living room, avoiding all obstacles. But a more complex scene, where a dog is frolicking about and children are playing in front of the kitchen cabinets, still throws the metal helpers off their stride. Scientists are aiming to teach their machines to take in complete scenes and especially to relate the objects to each other. If one vehicle brakes, then the next one brakes, as well. Modern automatic systems can’t usually master such simple logic. This is why there are still no camera assistants for the blind. The interrelationships are too complex. Pedestrians avoid each other, then suddenly cross one’s own path. And pedestrians and vehicles that are half hidden from view are regularly overlooked by automatic systems.

A CAR SHOULD NAVIGATE THROUGH TRAFFIC AUTONOMOUSLY

Bernt Schiele, Director of the Computer Vision and Multimodal Computing Department at the Max Planck Institute for Informatics in Saarbrücken, is working toward making assistant systems more intelligent. He wants to teach computers to understand a scene just like humans do – and to act accordingly. Schiele’s research group has developed sophisticated computation rules that analyze a street scene completely, registering all objects, pedestrians, vehicles and trucks regardless of whether they are easy to see or half hidden. If Schiele and his colleagues were allowed to do what they wanted to, they would have a vehicle with their software navigate through traffic autonomously.

A pedestrian coming from the right: The software developed by the computer scientists from Saarbrücken analyzes the scene. It recognizes the vehicles and the pedestrian, monitors the movements of all road users, and computes whether a collision is imminent. If danger arises, it acts autonomously.
Photos and graphic: MPI for Informatics

How a computer understands confusing traffic situations: A classifier determines the probability that a particular pixel cloud is a pedestrian. Detectors enter the results into a kind of map (left), where there is a high probability that the green areas represent pedestrians. The software frames the identified road users in a certain color (right) in order to monitor them and anticipate their movements.

They are already able to register complete street scenes in the images from a video camera mounted behind the windshield, observing the traffic. “Autonomous driving will definitely be here in the future, but of course it isn’t allowed – yet,” says Schiele. Understanding scenes – it sounds so simple. Thanks to our experience, we humans understand the situation at an intersection immediately. Traffic light red: everybody stops. Traffic light green: I can go. We don’t care how many people are hurrying to and fro. What a computer reads from the camera image is something completely different. It sees thousands of pixels, bright ones, dark ones, red and green ones, and it first has to learn what these actually mean.

Analyzing a scene thus requires a whole bundle of ingenious algorithms that the computer uses to analyze what is happening bit by bit. First, algorithms that recognize certain structures. Pedestrians are elongated and have a certain height. They have two arms and two legs. The front of a car is flat, truck fronts are high. The next step is for the software to find out whether and in which direction the objects are moving. And third, the computer must draw logical conclusions: if vehicle one stops at a traffic light, then in all likelihood, vehicle two behind it will stop as well.

The basis of all these analyses is the probability computation. The program executes billions of arithmetic operations per second to query the probability that a cloud of pixels really is an object. Schiele calls this complex form of computer understanding “probabilistic 3-D analysis of a scene.” It’s worth mentioning that Schiele uses only one camera for the three-dimensional – that is, spatial – analysis. Humans have two eyes in order to be able to see in three dimensions. “We compute the three-dimensional information from the two-dimensional computer image,” says Schiele.

THE SOFTWARE LEARNS WHAT CARS AND PEOPLE LOOK LIKE

But first, the researchers in Saarbrücken had to teach the assistant their algorithms: They fed the computer with training data – with hundreds of images of pedestrians, cars and trucks. The software thus learned step by step what vehicles and pedestrians look like. These recognition programs, which search through the clouds of pixels in the camera image for specific objects, are called classifiers. They detect outlines of objects with the aid of abrupt changes in the color or brightness of neighboring pixels. At the end, they output scores that state how probable it is that a certain pixel structure really is an object.

Schiele needs a whole range of different classifiers for the complex street scenes: for instance, those that recognize environmental structures such as a street or a tree, and those that scan the pixel chaos for discrete objects such as cars, trucks and pedestrians. Classifiers that recognize only complete objects aren’t sufficient here. Special classifiers that Schiele and his colleagues have trained with parts of objects – an arm, half of a person’s back, a hood – are used as well, because this is the only way that half-hidden objects can later be detected with certainty. The results of the classifiers, the scores, are evaluated by larger algorithms, the detectors. For each individual image in a video sequence, the detectors create a kind of map, a score map, that records the probability that each pixel belongs to a specific object.

In order to determine whether the detector result is plausible, the computer compares the score map values with its knowledge of the world. It has learned with the aid of its training data what a street or a vehicle looks like. And there is also the three-dimensional knowledge: The further away approaching vehicles are, the smaller they appear. Furthermore, vehicles that are further away are closer to the top of the
camera image than vehicles that are nearby. Its real-world knowledge tells it that a large vehicle couldn’t appear at the top edge of the image, for example. A street lamp, on the other hand, isn’t a pedestrian because it is much taller. The computer thus checks the plausibility of its analysis for each video camera image: how probable is it that the objects in this scene actually correspond to the real scene?

Trained in this way and equipped with real-world knowledge, the software had to be put to the test. The researchers played it real video sequences that they had recorded in a moving vehicle and that showed pedestrians hurrying through the streets of Zurich, for example. ETH-Loewenplatz, ETH-Linthescher and ETH-PedCross2 are the names of the sequences of images taken by researchers from ETH Zurich, where Schiele spent quite some time.

THE SYSTEM ANTICIPATES THE MOVEMENTS OF OBJECTS

It turned out that the classifiers and detectors were often wrong when they evaluated each image individually, one after the other. Hidden objects, in particular, were often overlooked. This changed when the algorithms compared around five successive images with each other. A non-flickering film sequence comprises at least 24 images per second. Moving objects change their position only minimally from image to image, just like in a flip-book. If the algorithms take several successive images into account, their recognition especially of hidden objects is better. “The image recognition became much more robust,” says Schiele. He calls these analytically fused short successions of images “tracklets.”

A crucial difference with respect to conventional vehicle assistant systems is the fact that the software permanently monitors the movement of the objects. Emergency braking assistants detect dangers that arise in a flash. In Schiele’s system, by contrast, the objects “propagate”: when the software detects an object on the screen, it surrounds it with a colored frame. The colored frame moves with the object from tracklet to tracklet until it disappears from the scene. If the street is full, dozens of these frames move across the video image. Thanks to the real-world model, the system can anticipate very precisely how an object moves. A pedestrian doesn’t suddenly accelerate to the speed of a vehicle when the traffic light turns to green. And a vehicle that briefly disappears behind an object in the foreground continues to move in the memory of the software if all other vehicles also continue their journey in the same direction.
The speed of the analysis is remarkable. It’s not just the 24 images per second that are important. If you want to keep an eye on the flowing traffic and analyze it in real time, you have to be faster than that. The complete and complex probability computation is carried out in milliseconds. This permanent rapid analysis has the advantage of protecting against nasty surprises. A vehicle that races into the intersection hidden by a line of stationary vehicles could possibly be overlooked by emergency braking assistants. With the 3-D scene analysis, it should be discovered early when it briefly appears in the gaps between the stationary vehicles.

VIDEOS FROM A CAMERA IN THE REAR-VIEW MIRROR

Vehicle manufacturers are, of course, also interested in teaching vehicles an anticipatory understanding of the events happening around them. In fact, Schiele has been cooperating with such companies for many years. He and his fellow scientists were provided with a vehicle with a small camera on the rearview mirror to take the vehicle video sequences, for example. “But this isn’t just about vehicles,” says Schiele. On the contrary, the probabilistic 3-D scene analysis is more suitable for analyzing very different film sequences – for instance also for the images from the camera eyes of a robot in a domestic environment or in a factory.

Some of the software components developed by Schiele and his team will soon be used for the first time in the US in a vehicle driving autonomously. But first the pedestrian and object detectors need to show what they can do. Schiele wants to test how the detectors work together with radar and laser scanners under conditions that are as realistic as possible. One of the objectives is to make do with as few radar and camera systems as possible and, in particular, to use commercial ones like those already used in vehicles. The technical effort required to realize the system must be within reason if it is ever to provide a driver with a sixth sense for traffic, or even, at some point, take over the steering wheel completely.

TO THE POINT

- Emergency braking systems stop at the last moment when a child or a vehicle appears in front of an automobile. They can neither analyze nor anticipate the behavior of other road users.
- Automatic 3-D scene analysis utilizes a probability analysis to recognize other road users, even when they are temporarily hidden, and can compute their movements in advance.
- Anticipatory assistants can be realized with little technical effort and can control a vehicle autonomously; they also allow robots to move in a complex environment.

GLOSSARY

**Classifier:** Software that recognizes objects in traffic situations. The classifier decides on the basis of probability values, known as scores, whether a specific pixel accumulation in the image of a real scene is the object on which it was trained with images – of vehicles or pedestrians, for example. Each class of objects requires special classifiers.

**Detector:** A program that collates the results of the classifiers in a score map. For each pixel in an image, it records how high the probability is that it belongs to a specific object.

**Tracklet:** Sequence of about five images in a video sequence, which are taken together in order to evaluate a scene. Since this results in larger jumps in the movement from tracklet to tracklet, the detector’s recognition particularly of partially hidden objects is more reliable.