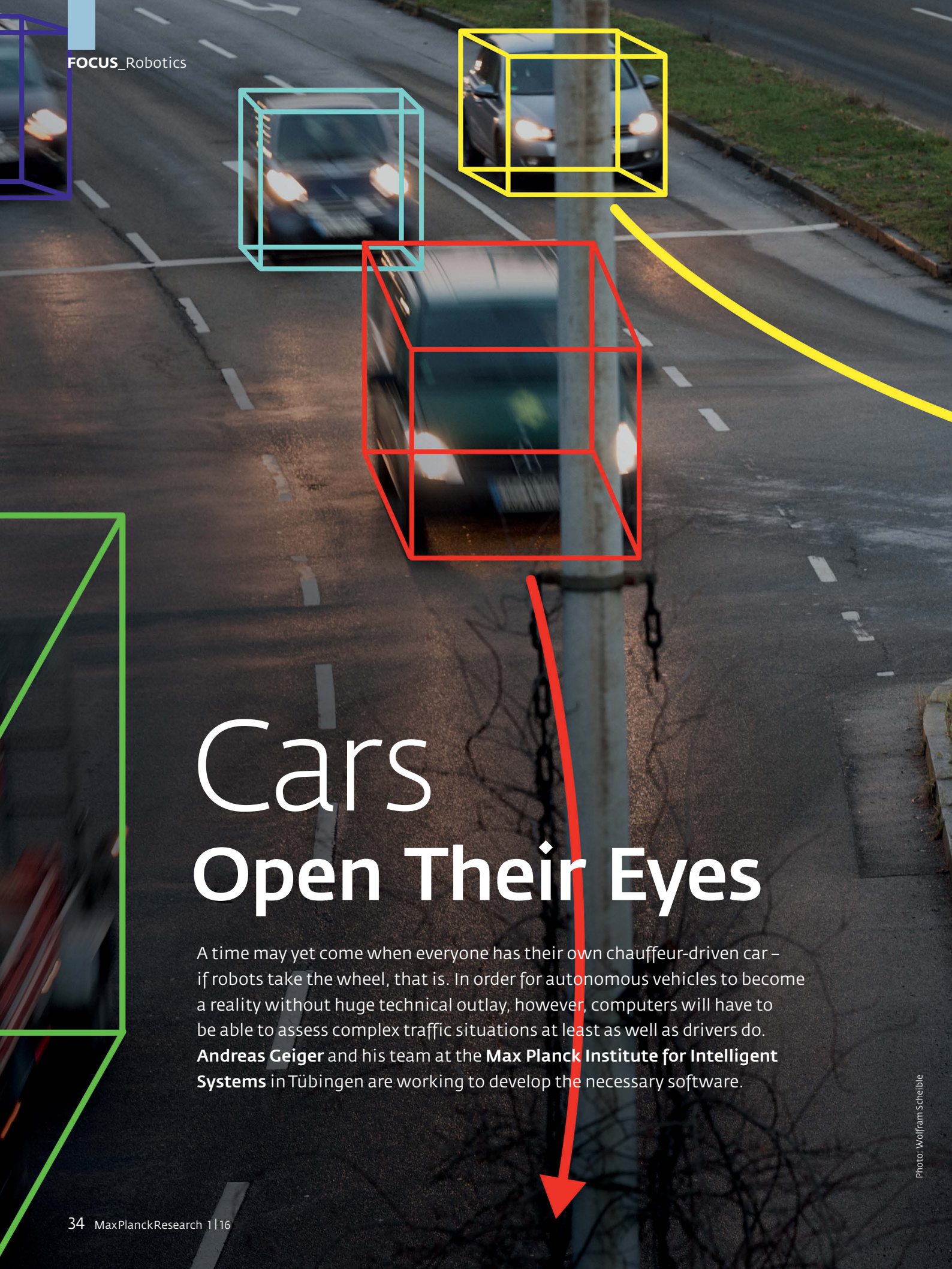


Cars
Open Their Eyes

A time may yet come when everyone has their own chauffeur-driven car – if robots take the wheel, that is. In order for autonomous vehicles to become a reality without huge technical outlay, however, computers will have to be able to assess complex traffic situations at least as well as drivers do. **Andreas Geiger** and his team at the **Max Planck Institute for Intelligent Systems** in Tübingen are working to develop the necessary software.

Photo: Wolfram Scheible

TEXT **CHRISTIAN J. MEIER**

In today's world, technology has eyes practically everywhere. Webcams can be had for just a few euros; smartphones often have multiple cameras, and stereo cameras in many luxury cars map their surroundings in three dimensions, not unlike humans. In this way, increasingly affordable image sensors are becoming an inescapable part of everyday life, and all kinds of life circumstances and situations are captured on photo or video. Every second, another 48 hours of video material is uploaded to YouTube, while Instagram, the online photo-sharing app, is growing at a rate of 20 million images per day.

For many, these ubiquitous cameras open new windows onto the world. But they mean even more to Andreas Geiger of the Max Planck Institute for Intelligent Systems in Tübingen: to him, cameras are the eyes of computer systems, enabling them to actually perceive and understand the world around us.

"Perception is an essential component of intelligence," explains the computer scientist, and illustrates his statement with an example: "We humans often give things striking shapes and colors to help us find our way in the world. Think of road signs, for in-

Object recognition: A kind of world knowledge helps software identify people and cars, even when they are partly obscured from sight. It also makes it possible for programs to predict the behavior of road users.

Stereoscopic images for modeling: In order to estimate distances, the program locates corresponding points on two images taken from different angles, and uses this information to reconstruct the scene with depth information. The white patches represent areas for which no information was available, as they were hidden from the camera.

stance." As it is hoped that computers will find their bearings in the human world more easily in the future and move autonomously in applications such as domestic robots and autonomous vehicles, they must first learn to perceive their environment as humans do.

There is a problem, though. Computers don't understand images, which they see as a chaotic mosaic of millions of varicolored pixels instead of a scene containing houses, trees, cars or curbs. People, in contrast, recognize objects and are able to grasp complex situations, anticipate movements and estimate distances. "Computers are still a long way from that goal," says Geiger. "Many treasures remain hidden to them for now, lurking in the deluge of images."

If a computer is to guide a driverless car through traffic, it must be able to assess whether the vehicle in front is going to turn or keep going straight, or whether a child on the curb is going to run onto the road. "This is why we're developing systems that can perceive situations as humans do and react accordingly," explains Geiger.

The process of teaching computers to detect objects and interpret scenes is an arduous one. "They have to convert the light that has been captured into meaning," as Andreas Geiger puts it. To this end, a program must first reconstruct the three-dimensional world that has been captured as images in just two dimensions. Geiger and his research group of four are developing the software required for this kind of task.

Objects such as cars, tables and even the human body with all its complex movements can now be represented in computer language. The virtual world contains three-dimensional models of people, monsters and Formula One racing cars. In computer games, such models meet, fight and compete with each other: in short, the computer simulates highly complex scenes within a 3D virtual reality.

AMBIGUITIES IN TWO-DIMENSIONAL IMAGES

Gamers, however, see only two-dimensional images as their graphic cards continually project the complex three-dimensional model world of the game onto their flat screens. "The software does an amazingly good job of converting the spatial model of the virtual world into a two-dimensional image," affirms Geiger. The challenge now is to achieve the opposite, namely to take two-dimensional camera images and compute a model of three-dimensional reality.

"One of the problems with this is the issue of ambiguity," says Geiger. An image that contains a thick tree trunk can be interpreted in different ways by the computer. The thick trunk could actually be a thin trunk that is close to the observer. Two different 3D models, one with a distant, thick trunk and another with a close-up, thin trunk, would generate a similar image on the camera.

As a two-dimensional image lacks depth, it is not possible to conclusively differentiate between the two options. This is why computers use stereo images, as humans do, to estimate distances and detect the spatial structure of a scene. But even then ambiguities may arise, as Geiger shows using two images of a residential street lined with old houses, with vehicles parked along both curbs. The images show the same scene captured from slightly different angles, as if seen using the right and left eye of a human observer. The

Graphic: Andreas Geiger/MPI for Intelligent Systems

human brain would then generate a spatial impression using the data from both perspectives.

Computer software can estimate distance in a similar way, by measuring the displacement of a feature such as a window frame in one image compared to the other. If the displacement is large, the object is close to the camera. If the images reveal only minor displacement of the feature, then it is located far away. This principle can be observed by looking at a close object and closing the left and right eye alternately. The object will appear to move back and forth in relation to the background. The computer converts this displacement data into actual distance values in meters.

The computer goes about this by comparing the individual pixels in both images. For each pixel in the first image, it looks for its counterpart in the second, meaning the pixel that represents the same point in the real scene. It does so by analyzing the color values of the pixels.

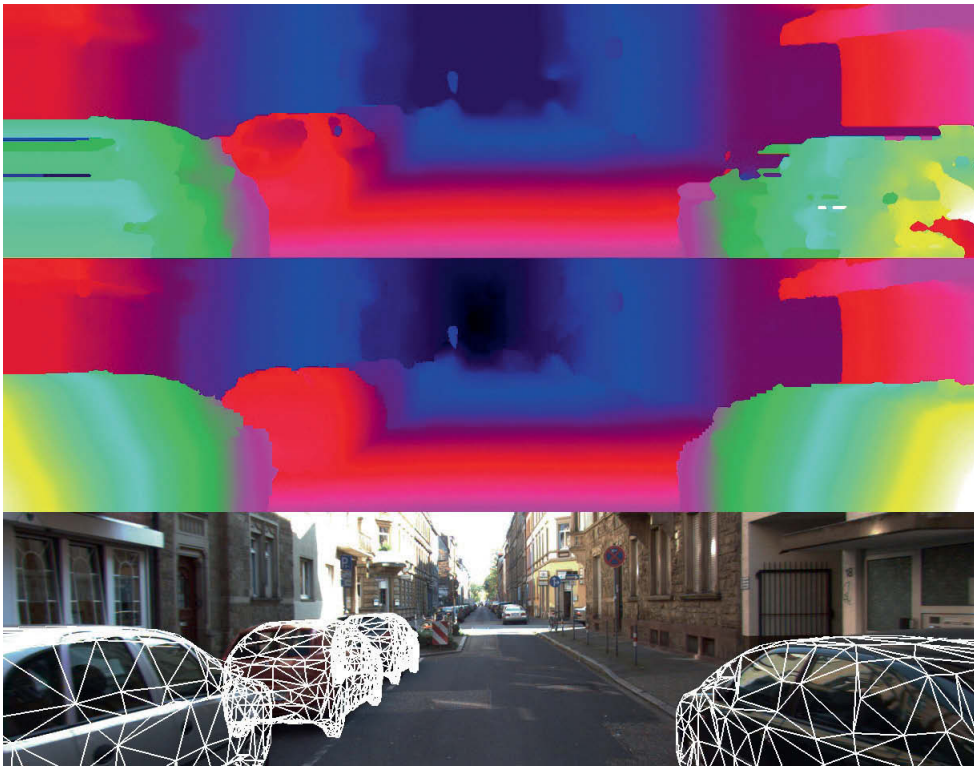
"Edges such as window frames are easy to pinpoint in this way," says Geiger, as they show an abrupt transition from one color to another, and this is easily recognizable in the second image. Paint on a car door, on the other hand, is generally monochrome and all pixels have a similar color value. This means that, for each pixel in one image, there are many candidates in the second image that would have to be considered as possible counterparts. Existing procedures for calculating depth maps are unable to handle this level of ambiguity. In the worst case, it leads to miscalculation of depth – and in a system that is relevant to safety, this could have fatal consequences.

Geiger illustrates the problem with the image of a scene in which depth is represented by false colors. Green dominates for close objects, with violet and red further back, and everything that is far off appears blue. Vehicle contours can be detected on this depth map, but many colored specks appear around car doors. "In those cases," explains Geiger, "the computer was either unable to assess the distance or miscalculated it."

OBJECT KNOWLEDGE HELPS MAKE SENSE OF DISTANCE

In order for their computer to estimate distance reliably in spite of these difficulties, the Tübingen-based researchers feed the software with information about the image, called object knowledge. In other words, they turn a collection of pixels into a scene with objects as a human would perceive them. Adaptive software can identify cars on the basis of multiple sample images, and then consistently mark the places in new images where cars are located. In this way, the computer learns to detect the presence or absence of cars in a given image.

Geiger describes object knowledge as mid-level knowledge, or "knowledge of a medium level of abstraction." It is helpful, he says, to build a scene up from pixel-based low-level features such as those window frames and divide it into different items, just as a person detects tables, chairs and cupboards within the home. >



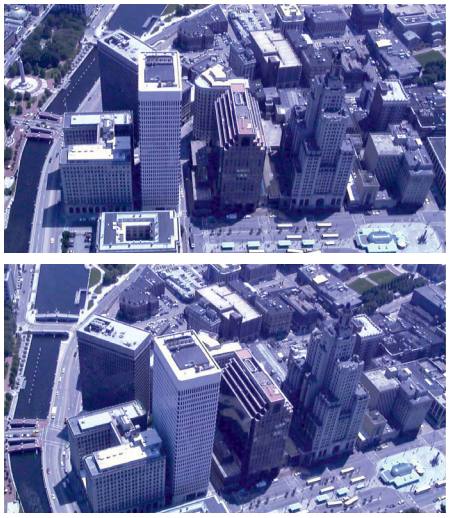
Top A depth map codes distances using different colors (yellow – near; blue – far).

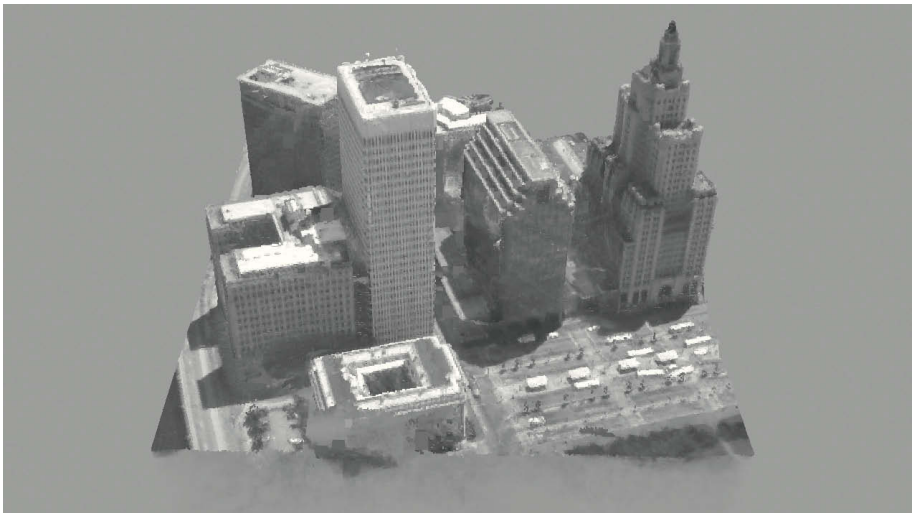
Bottom The software calculates distances with the help of information on the geometry of objects such as cars. Relevant models are stored in its memory.

Graphic: Andreas Geiger/MPI for Intelligent Systems

Right-hand page Built on probabilities: Osman Ulusoy, Joël Janai and Andreas Geiger (from left) discuss the algorithm they are using to reconstruct 3D models from stereoscopic images. The background image shows them the algorithm's confidence level in relation to depth information for the Capitol in Providence. White dots on the image mean the assessment is fairly reliable, unlike the black dots, which indicate that the algorithm has depended more on prior knowledge, for example about the general shape of buildings.

Below Downtown Providence poses: Osman Ulusoy uses aerial photos taken from different angles (left) to compute a 3D reconstruction of his hometown in Rhode Island (US). The reconstruction can then be used to observe the city center from different perspectives not provided in the original images (right).





Geiger's team uses software that reconstructs scenes virtually using geometric 3D models of cars, generating a 3D simulation with virtual cars in a row. Modern graphics cards accurately convert these scenes into depth maps without any data gaps around the car doors, as they are based on complete 3D models.

However, this still doesn't deliver a totally unambiguous result. It is not clear from the photos how many cars are parked along the curbs, or whether they are parked parallel to the curb or at an angle. Consequently, there are thousands of simulations with different numbers of cars and different parking orientations that reconstruct the image of the streetscape with varying degrees of accuracy.

The research team's program tests all these variants for consistency with the corresponding image data. So, for example, it compares the depth map generated by the simulation with that produced using only pixel comparison with no world knowledge. The software also measures how well the artificial image reproduces the areas where cars are located in the real image. "This allows us to filter out the most probable hypotheses," explains Geiger. The method doesn't deliver any hard and fast certainty, but it does achieve a more consistent and meaningful interpretation of the image.

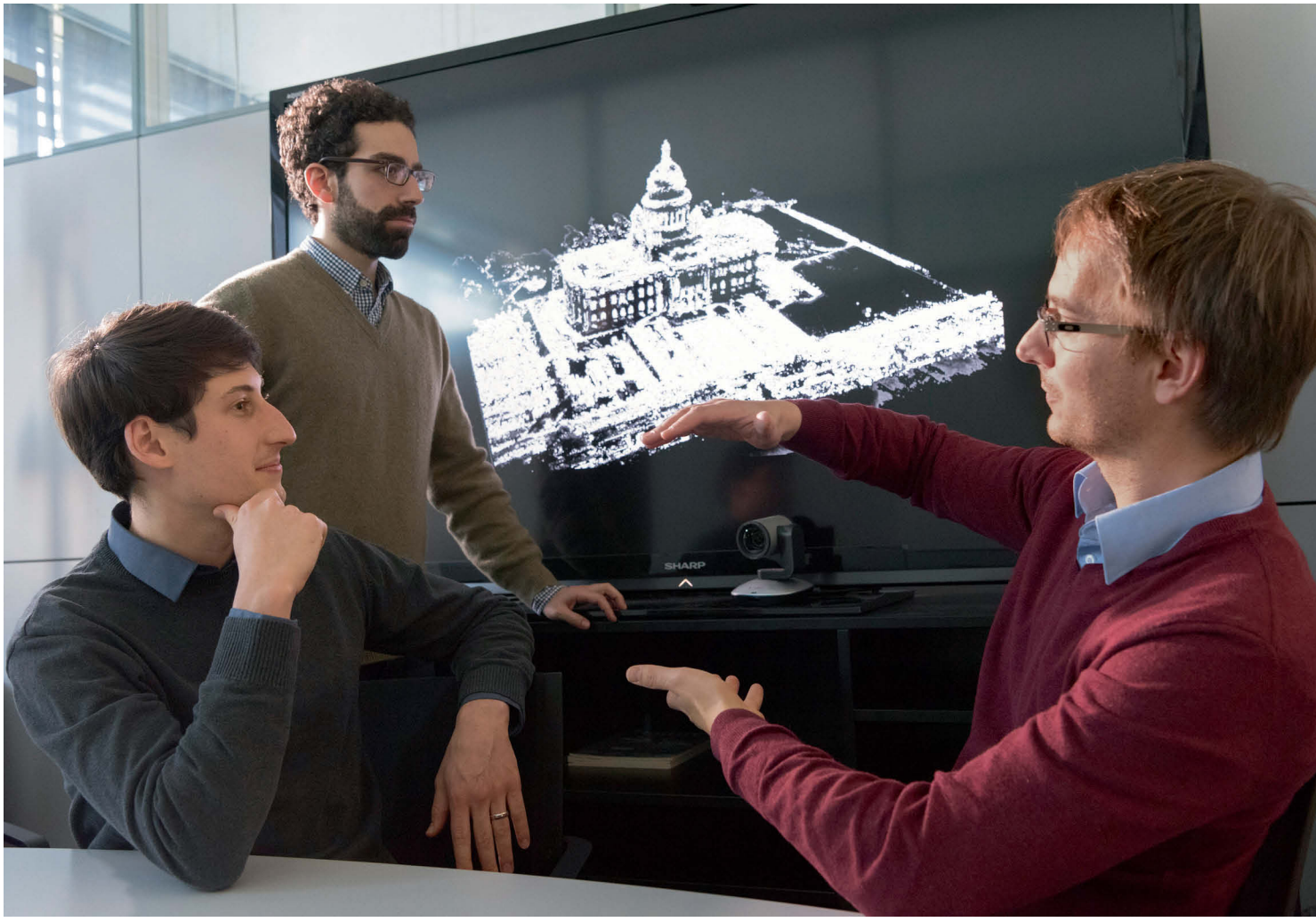
AERIAL PHOTOGRAPHS YIELD 3D CITY MODELS

Osman Ulusoy, one of Geiger's team colleagues, demonstrates a similar principle using aerial photographs of his hometown of Providence in Rhode Island (US). "Photographs taken from different angles can be used to generate a 3D model of the city center," he explains. However, reflecting surfaces are difficult for the computer to reconstruct, since reflections throw the calculation of distance into confusion.

"We feed *a priori* knowledge into the computer to close the gaps," says Ulusoy. This is a kind of world knowledge in terms of the characteristics and structure of things in general; for example, the fact that reflecting surfaces are generally smooth. This enables the software to complete the model in spite of ambiguous observations. "This could be of interest to urban planners," surmises the computer scientist. "It would enable them to document the development of the city in 3D."

Indoor scenes can be virtually reconstructed too, as Andreas Geiger shows using an image of a room with a bed, chair and cupboard. "The model recognizes the shapes and sizes of typical pieces of furniture," he ex-

Graphic: Ali Osman Ulusoy/MPI for Intelligent Systems



plains, adding that it can detect a chair even if the image shows only a side view of the chair back. Again, the researchers feed *a priori* knowledge into the system for the virtual reconstruction of the scene. "Cupboards, beds and sofa are generally positioned up against a wall," says Geiger. Furthermore, objects do not intersect each other. As in the scene with parked cars, this knowledge limits the number of possible hypotheses to a range that the computer can run through in a shorter time.

The virtual reconstruction of indoor areas may be useful for robots that need to maneuver safely in a domestic setting. It could also help architects and designers produce more realistic drafts and develop ergonomic designs, Geiger believes.

As the computer uses the object knowledge supplied, it learns to detect objects in new images. "Still, it's important to approach the problem as a whole and not just focus on the individual components," warns the group leader.

HIGH-LEVEL KNOWLEDGE FOR INTERPRETING THE IMAGES

The team in Tübingen relates the objects in an image to each other by inputting high-level knowledge into the computer, that is, knowledge involving a high degree of abstraction. This includes the above assumption that pieces of furniture do not intersect one another, or that they are generally positioned against a wall.

It is this high-level knowledge that enables the computer to assign meaningful interpretations not only to static images, but also to moving ones. Here, Geiger uses the term "3D scene flow," which means an estimate of the three-dimensional movement of all objects in the scene. His team attempts, for instance, to derive the best data from the rather limited perspective of traffic scenarios as captured by a car's on-board cameras, say at the junction of two busy streets in the city center.

A fixed bird's-eye view would be the best perspective for understanding this type of situation, as only the vehicles would move and it would immediately be clear which lanes they were driving in, what traffic lights are located at the junction, and how the traffic light phases alternate. "From a height of 1.60 meters, which would be typical for a car's stereo cameras, it's much harder to deduce that information, and it involves a greater degree of uncertainty," says Geiger. In fact, the built-in cameras are often unable to detect whether the traffic lights for their own lane are red or green. >

Photo: Wolfram Scheible

Two parts, one person: Andreas Geiger demonstrates a scene that computers do not initially understand. They do not grasp, namely, that there is only one scientist in the photo, not two. Geiger's team is working to train software to reach this conclusion on its own.

Despite this incomplete and unreliable information, the team in Tübingen hopes to make autonomous vehicles a reality by increasing the intelligence of on-board computers: they aim to teach them to accurately detect and interpret scene flow.

OBJECT RIGIDITY REDUCES THE NUMBER OF MODELS

The first problem: the identification of other road users. To a computer, a street scene is initially just a swarm of moving pixels. We humans, on the other hand, know that many of the scenes we observe, including road traffic scenes, consist of a small number of rigid objects. Cars don't suddenly change shape, but move as a compact whole.

Then again, there is only a small number of vehicles at a junction at any given time, and not hundreds. "So we tell the computer to break the scene down into the smallest possible number of rigid components," says Geiger. Rigid objects have less freedom to move than, say, the human body. They can move along three planes: forward and backward, left and right, up and down. They can also turn on three axes, while the complex movement of a human body involves hundreds of variables, including the rotational angles at each joint.

"So this assumption of rigidity greatly restricts the model of the scene," says Geiger. The computer has fewer variants to test for plausibility and resolves ambiguities with better results. Furthermore, the command to identify the smallest possible number of objects excludes many other hypotheses, for example that a car on the far side of a lamppost be misinterpreted as two separate objects. Rigidity, then, is a simple criterion with far-reaching effects.

Once Geiger's software has detected individual vehicles at a junction, it follows them for a time. Do they drive straight ahead? Do they turn right or left? This is where machine learning comes into play. Using many sample images, computers learn to recognize certain elements. If a computer is trained with thousands of images of human faces, it can ultimately detect faces in new photos by itself.

CAMERAS AND INTELLIGENCE REPLACE EXPENSIVE TECHNOLOGY

In similar fashion, the software designed in Tübingen is trained to use traffic flow data and road markings to detect the lanes for driving straight and turning right or left, and to infer traffic light sequences. "Different types of traffic light configurations are linked with different phase sequences," explains Geiger. "Our computers learn those sequences on the basis of large volumes of data, and then use that knowledge to improve how they relate road users to each other."

Photo: Wolfram Scheible

The junction surroundings are also subjected to scrutiny: the location of buildings, the orientation of the streets, etc. The computer uses all this information to reconstruct a digital map of the junction and run a virtual 3D film that reduces the scenery captured by the cameras to a bare minimum. The autonomous system can build on that to reach the right decision; and it does so ad hoc for each new junction it comes across.

"If autonomous vehicles were to combine cameras and intelligence, they would manage without the expensive technologies of today's prototypes, like laser scanners and radar," affirms Geiger. The highly accurate satellite navigation and laboriously produced digital maps such as those used for current systems would not be necessary, and in the transitional period when only a small number of autonomous vehicles are on the road, there would be no intelligent infrastructure available to support the new class of car.

There is currently still a problem with the software for analyzing complex scenes: relatively speaking, it still makes many mistakes. It mistakes a sofa for a bed, or a piano for a table. It slips up at junctions, partly because machine learning in the area of road traffic is more arduous than for facial recognition, for instance. A very high volume of data is required to train it, but there are far fewer video sequences containing cars than photos of faces. Not only that, but people have to overlay the training data with information, for instance by showing the computer where to find faces in the images. "This kind of annotation is very labor-intensive in the case of traffic junctions," says Andreas Geiger.

The pitfalls of digital photography pose a further obstacle to the researchers. Bright sunlight can blind the sensors, trees can obstruct a view, and large contrasts between light and dark can make it impossible to capture a scene on camera. In such cases, the accuracy of the virtual reconstruction suffers or is rendered impossible.

ACCEPTANCE OF THE TECHNOLOGY WILL COME

Again, the researchers intend to tackle these technical problems using *a priori* knowledge. "In the case of houses in a development, we can assume that they are similar to each other," says Geiger. This assumption of similarity helps the computer to virtually reconstruct an entire residential street, even if it is lined with trees or the camera is hampered by the sun.

Think of it like this: The system logs the facade of one house, the left external wall of another, and the right wall of a third. Since the houses are assumed to be similar in structure, these three pieces of the puzzle can be combined to generate a typical house for that street. "The model is flexible enough to allow the extrapolation and interpolation of different geometries," says Geiger. This means it can generate houses that have not actually been observed but fit perfectly into the development based on their appearance.

However, even if the software continues to become better at assigning meaning to billions of pixels, those interpretations of the images it captures are still just approximations. And even the most probable hypothesis is only a hypothesis and not a certainty. But surely certainty is the one thing needed in situations involving road traffic?

"Even a good driver can only estimate how the car in front will behave," counters Geiger. It is true, he admits, that computers are not yet as good as drivers in making such estimations. "Acceptance for this kind of technology will come as soon as the systems make significantly fewer mistakes than human drivers." ◀

TO THE POINT

- **Initially, computers perceive images only as a series of meaningless pixels. Andreas Geiger and his team at the Max Planck Institute for Intelligent Systems train them to understand images of complex traffic situations and to anticipate the behavior of road users.**
- **When a program uses two-dimensional images to construct a three-dimensional model of a street scene, ambiguities arise in areas such as the evaluation of distance. To resolve this issue, the team input data of a medium level of abstraction into the software, helping computers to recognize individual objects such as cars.**
- **The software uses knowledge with a high level of abstraction to understand how individual objects relate to each other. This specifies, for example, that solid objects do not intersect one another.**
- **When computers have analyzed many traffic situations using machine learning, they can anticipate traffic flow at junctions.**