Self-supervised monocular depth estimation for dynamic scenes

Popular Science Summary - Jonathan Lindberg

Supervisors : Amer Mustajbasic (Volvo Cars)

Anders Heyden (LTH)

Humans are able to view the world thanks to their eyes and we are able to move around different places. An important aspect of this, is to estimate the depth and distance to objects. Likewise, for cars to be able to move around the world and aid human in driving, it needs to be able to perceive depth. In this master thesis, the ability to train the computer to estimate depth from a single camera under different scenarios is studied.

To reduce accidents for all involved, assisted driving and active safety technologies are very important. To produce measurements against potential danger, the car needs to be able to see its surroundings. The eyes of the car are usually cameras mounted around its side. A camera, however, loses all information about depth when taking a picture or recording. To regain the depth information, a technique similar to how humans refer to depth can be used. The human eye can estimate depth such as the familiar size of objects or how the same object differs when viewed with different eyes. Machine learning can mimic this behaviour. If one constructs a neural network and trains it on a lot of data, one can estimate depth from a single camera frame. However, the data must be very specific. The data has to be recorded on a static world in which only the car with the camera is moving. Typically, this is not the case when recording natural data. Therefore, in addition to the static scenes, the network must also include moving objects. This master's thesis examined the methods to incorporate moving objects into training the network by removing all non-stationary objects. Using an auto mask and a backwards-forward optical flow consistency check, it was found that depth estimation improved significantly after moving objects were removed. The consistency ensures that the pixels that is moved to the next frame are moved back (by another approach) to the same position. If the pixels do not return to the same spot, they are removed and can therefore not interfere with the training of the network.