Automated visual traffic surveillance system for traffic analysis

Speaker: Dick Scholte, R&D Engineer ViNotion BV / Doctoral candidate TU/e

Project name: CONVOLVE - EU Horizon 2020 (Grant nr. 101070374)

Automated traffic surveillance systems support a range of tasks involving crowd management analysis, congestion, and (near-)accident observations. In these systems, high-level metrics such as the type of vehicles are generally determined by computer vision algorithms, which visually interpret the camera streams. In order to analyze the behavior of traffic participants, it is necessary to accurately detect, localize, and follow all objects in the camera scene in real-time. To achieve real-time analysis, computationally efficient algorithms and optimized embedded hardware are required. This talk presents the main components in such an analysis system and discusses the obtained metrics. Furthermore, two recent optimizations are highlighted. First, the use of dynamic neural networks for optimizing computational complexity and thereby lowering energy consumption. Second, the introduction of improved localization techniques using instance segmentation models instead of the typically employed 2D bounding boxes for localization.

Typical system description: A typical analysis system first performs object detection, classification, and localization on an incoming video stream. Typical techniques for object localization use 2D bounding boxes to represent the object location. The obtained localization results are combined over time in an object tracking algorithm to form trajectories. By using a camera calibration, the object trajectories can be converted to real-world (GPS) locations for further high-level analysis. High-level analysis involves measuring near-collisions, inter-vehicle distances, density metrics such as the number of vehicles per traffic lane, and traffic-flow analysis. These results can then, for example, be used for behavior analysis of traffic participants.

Dynamic neural network for efficient object detection: In traffic surveillance, a large percentage of the time no objects are present in the camera stream. However, the neural network for object detection is always fully utilized. We experimented with a dynamic neural network technique to implement an early exit in the YOLO object detection model architecture to reduce the computational load when no traffic participants are present in the scene. This approach reduced the inference processing time by 60 to 75%, and thereby significantly lowered the overall power consumption, making the dynamic model suited for battery-based systems.

Improving object localization with instance segmentation: Localization of objects has been improved by using instance segmentation instead of 2D bounding boxes by extending the YOLO model with YOLACT for predicting instance segmentation masks. The creation of large-scale annotated datasets, to allow training a neural network that can predict instance segmentation masks, is challenging and time-consuming. To this end, foundation models have been utilized to generate instance segmentations for an existing dataset automatically. Although the initial results are promising, the construction of an automated pipeline is not straightforward. The automated process to generate the dataset for training and the instance segmentation model that is trained on this dataset (YOLO + YOLACT) have to be adapted to enable training and to achieve real-time performance. The trained detection and instance segmentation model improves the tracking performance, ground-plane localization, and allows estimating accurate 3D bounding boxes and object dimensions.