Low-Power Embedded AI Camera System

Motivation

AI adoption is accelerating, but centralized inference is costly, power-hungry, and raises privacy concerns. With billions of cameras and other sensors producing ever-growing data, network transfer becomes a bottleneck. Modern embedded hardware offers rising compute power at low energy, enabling tasks to move to the edge—reducing data flow, preserving privacy, and allowing scalable, sustainable, even ambient-powered AI deployments.

Approach & Methods

Our research explored modern embedded AI accelerators that use integer arithmetic, combined with advances in quantization. These methods allow deep learning models to run efficiently on edge devices, far from traditional data centers. To exploit this, we co-designed a multi-stage hardware/software pipeline for an AI camera system, targeting state-of-the-art performance under strict power constraints.

The pipeline progressively reduces the problem in both space and time: early stages perform lightweight detection or classification to identify regions of interest or select only key frames, while later stages handle more complex tasks. This staged design minimizes energy use, improves throughput, and preserves accuracy.

A critical part of the system is the configuration of the camera's image signal processing (ISP) pipeline. Unlike traditional setups that optimize for human viewing, our ISP is tuned for machine perception. We deliberately maximize analog and digital gain to maintain sufficient signal levels, even at the cost of noise, while keeping shutter times short to avoid blur. Neural networks can learn to handle noise, but lost sharpness cannot be recovered. By tailoring imaging for AI, we provide input data that supports robust inference.

Overall, the combination of quantization, staged processing, and ISP tuning enables scalable, low-power, real-time AI operation in constrained environments.

Camera with CNN Accelerator

Modern vision sensors, such as the Sony IMX500, integrate on-chip accelerators capable of running 8-bit integer-quantized convolutional neural networks directly on the sensor. This enables real-time object detection and delivers already annotated frames to the next pipeline stage. The CNN is tuned to minimize missed detections, even at the cost of more false positives. As a result, the camera serves as a very low-power yet high-quality pre-detection stage, deciding whether the full pipeline runs or energy is saved.

Low-Power AI Accelerator

State-of-the-art low-power accelerators, such as the Hailo-8 AI Accelerator, can run complex neural networks with remarkable efficiency. Their adaptive compilation process balances precision and performance by selecting between 4-bit, 8-bit, and 16-bit integer weights. This ensures fast inference, low energy use, and high accuracy. In the pipeline, the accelerator refines or validates the camera's detections and can also expand them with more advanced analysis.

Multi-Core ARM CPU

The ARM CPU orchestrates the pipeline, configuring the camera, preparing data for the accelerator, and fusing their outputs. It can also run selective, higher-precision inference (e.g., full 32-bit models at lower FPS or small image crops) or apply post-processing across multiple frames to boost accuracy. By smartly distributing tasks across its cores, the CPU ensures smooth operation while still leaving capacity for storage and data transfer of both results and selected frames.

Experimental Pipeline

Our experimental multi-stage pipeline, built on the aforementioned hardware components, demonstrates effective vehicle detection and automatic license plate recognition, achieving strong results under daylight conditions.

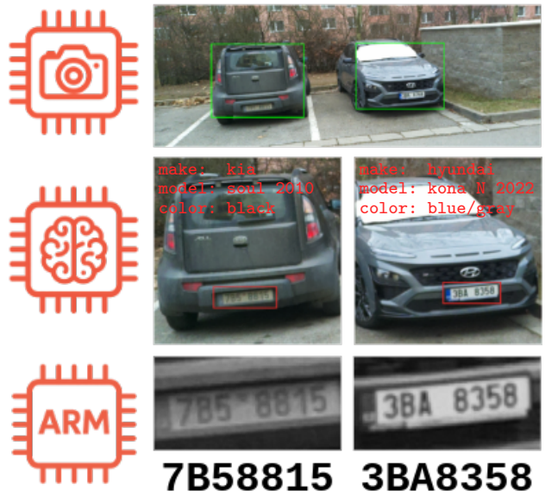
DETECTION PRECISION	DETECTION RECALL	LICENSE PLATE READING ACCURACY
0.983	0.980	0.993

THESIS SMART CAMERA FOR TRACKING OBJECTS OF INTEREST

AUTHOR Ing. David Mihola

SUPERVISOR doc. Dr. Ing. Otto Fučík

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY



Experiments & Results

As a practical demonstration, we developed an application for vehicle detection and automatic license plate recognition, combining the Sony IMX500 AI-enabled sensor, the Hailo-8 AI Accelerator, and a SoC Broadcom BCM2712 as an ARM CPU. The pipeline is structured into three stages: the camera first detects vehicle fronts and rears, the accelerator stage is designed to perform license plate localization along with vehicle make, model, and other classification tasks, and finally, the ARM CPU runs a full 32-bit model to recognize plate numbers. Tracking and confidence-based aggregation algorithms across multiple frames further improve robustness and accuracy.

The system proved highly versatile: it can operate as a handheld device for parked car inspection, as a fixed traffic monitoring camera, or mounted on a vehicle for mobile data collection, while handling diverse lighting conditions. In daylight, it achieved detection precision of 0.983, detection recall of 0.980, and plate reading accuracy of 0.993. At night, performance remained strong with precision of 0.934, recall of 0.989, and accuracy of 0.951.

Impact

Our technology can be applied to many computer vision tasks, from smart city infrastructure to improved security systems. This makes everyday processes power-efficient, reliable, and scalable.

We are working on the realization of the demonstration into a practical commercial product.