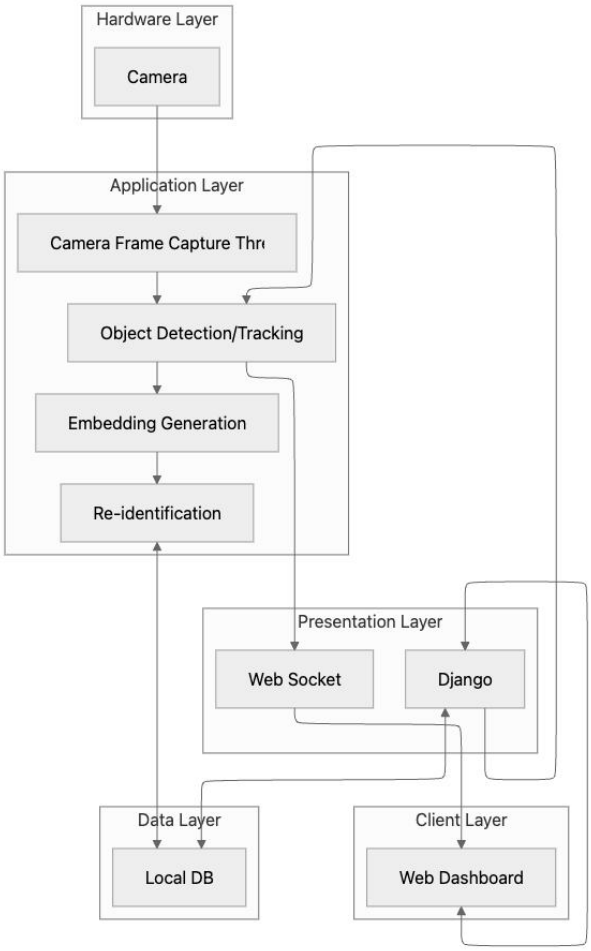
Seungwan, Hong www.hosugator.com

Pic Tag

Development of Lightweight CCTV AI SaaS Optimized for Small Business Owners

System Architecture

A modular AI pipeline design was adopted, separating the logic for Object Detection, Embedding Generation, and Re-identification to ensure system scalability and reusability. This structure was designed to flexibly accommodate new AI models or algorithm changes without causing service disruption.
To enhance system throughput, an asynchronous high-performance processing pipeline was constructed. The entire process, from camera frame capture to the final re-identification result, is structured using Async Threads to handle I/O and computational tasks in parallel.



Core Technology

Re-identification Algorithm Development & Learning Advancement

The generalization performance of the AI model was enhanced by ensuring diversity in the training dataset through data augmentation of a Korean Person Re-identification img dataset. Systematic sampling was applied to filter the dataset, resulting in an improvement of overall training efficiency by over 50%.
Logic was implemented to determine the optimal Threshold by analyzing the similarity difference between the same person and different people, which is then used for ID assignment.

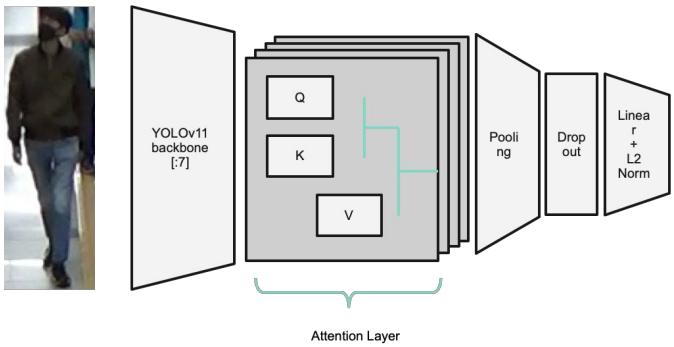
Real-time Object Detection & Dashboard Integration

Streaming data received via the RTSP protocol is processed in the Camera Frame Capture thread, and a Websocket connection with the Django-based Presentation Layer is used to reflect object detection results to the Web Dashboard in real-time.
Buffer-based data passing is used when transferring data to subsequent models in the AI pipeline (Object Detection, Tracking) to manage data bottlenecks and increase system throughput.

Problem Solving

Resolving Data Bottlenecks through Algorithm Optimization

The initial re-identification algorithm design had a high potential for data bottlenecks, necessitating speed improvements.
Following Yolo Backbone Decomposition, Feature Extraction algorithms (Linear $\to$ Pooling $\to$ Attention) were comparatively analyzed for performance enhancement. Ultimately, the Attention algorithm was selected. This algorithm effectively resolved data bottlenecks and improved re-identification accuracy by assigning weights (Attention Score) to all hidden state vectors and generating a new vector based on these weights at every moment. The ID assignment logic is based on the difference in similarity values (Threshold) between the same person and different people.



Retrospective

The project involved implementing various protocols, including RTSP and Websocket-based real-time communication, and finding high-level technical solutions such as Yolo Backbone Decomposition and OpenVINO. This secured a spirit of challenge toward unknown technologies and the capability to design solutions that consider the actual operating environment.
The experience of designing a real-time processing pipeline to optimize performance under strict hardware constraints was particularly instrumental in acquiring solution design competency that is mindful of the real-world usage environment.

RTSP Real-Time Streaming Protocol · Websocket · django · PyTorch · ultralytics · asyncio · OpenVINO · POSTMAN