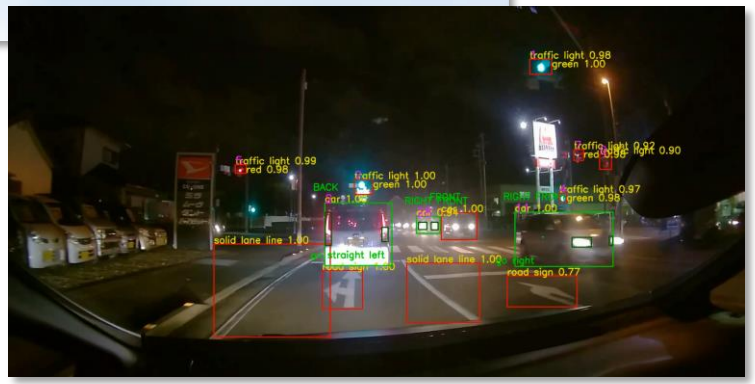
Practical Software-Hardware Co-Design and Implementation Method for AIoT System

Fujitsu Laboratories

Koichiro YAMASHITA

Topic of Today

- AI based Image Analysis System using CCTV and On-vehicle Camera



Technical Market (Japan)

- CCTV Network : Roadside 40,000 Cameras
  Others 5,000,000 Cameras

- Usage : Real-time traffic reports
- Location : Fixed position
- Operating time : 24h x 7d monitoring

- On-vehicle Camera : 23,000,000 Cameras

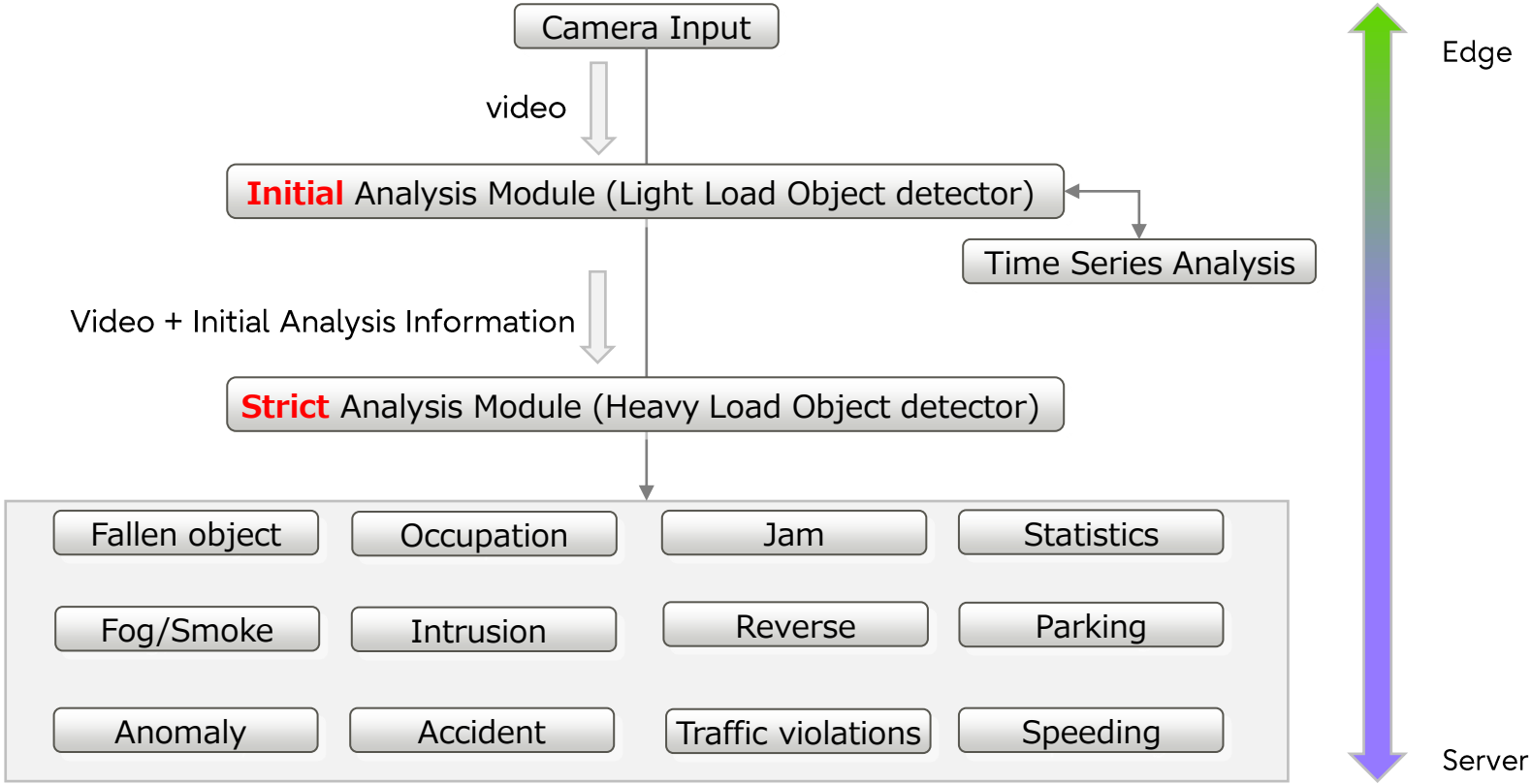
- Usage : Insurance assessment (accident situation analysis), Non-real time
- Location : Anywhere
- Operating time : When an accident occurs
  300,000 cases/year

FUJITSU

Traffic Surveillance System

Recognize events in video surveillance

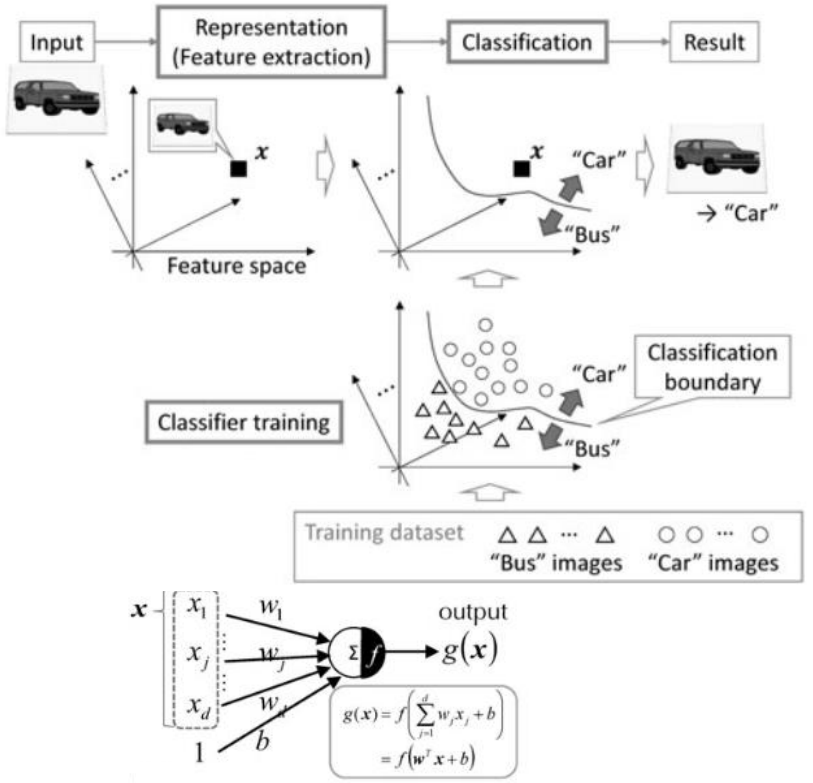
Functions



#1 Traditional Imaging: Light Load

- Usable level of accuracy (Low robustness)
- Easy to estimate calculation time
- Low computational load on CPU

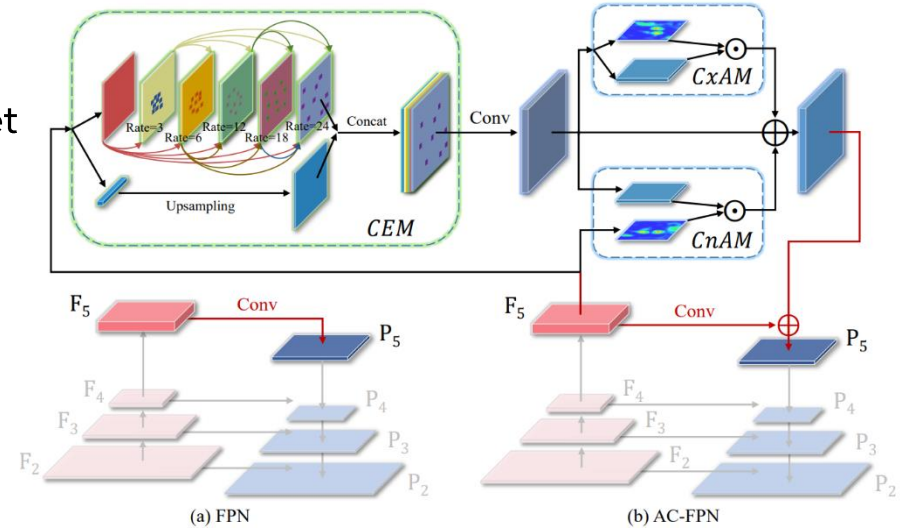
White vehicle in the snow !?



Pattern Recognition and Image Informatics [1]

#2 Deep learning : Heavy load

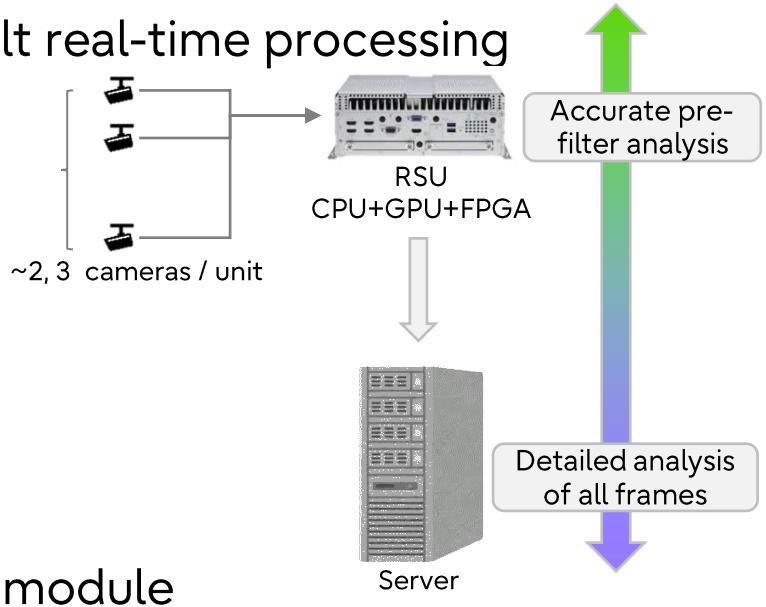
- High accuracy (inference robust to disturbances)
- Difficult to estimate calculation time
- Huge amount of CPU GPU** resources
- Large amount of training data(100K samples / target objects or scene)



A Structure* of Attention-guided Context Feature Pyramid Network (FPN) [2]

Edge – Server Distribution

- Hardwired DL modules lack of flexibility
- Current embedded GPU capabilities, it is difficult real-time processing



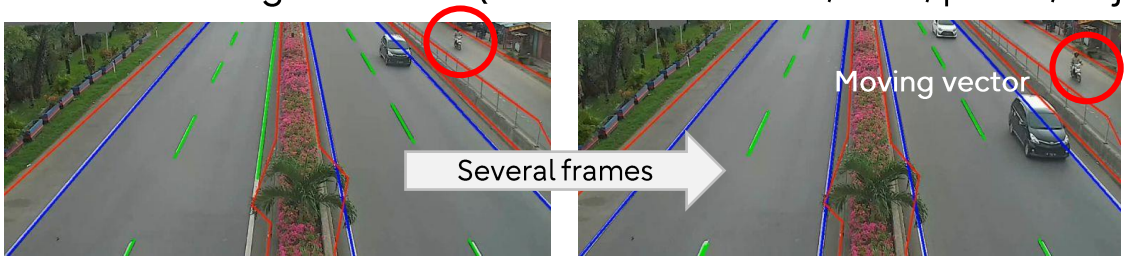
Intermediate solution: RSU (Road Side Unit)

- Weather-resistant, relatively high-performance module
- Used as a hub for V2X communications between connected-cars
- Although it has some effect in reducing server load, it is not cost-effective.

Practical Edge – Server Distribution

- "Key frame determination" on the edge side
  - Traditional Video Compression
    Key frame : Regular Interval Frames and Detected Scene Changes
    Dependent frame : “Moving Vehicle” → Background + Moving vector of moving object
  - Optical flow + Semantic Segmentation (Paint each material; road, plants, objects…, separately)

Traditional video compression



Mobile object identification for reducing server load

Even if the object is small and cannot be detected, you can see that the colored parts are moving.

Moving Vector of Colored part > Threshold : Keyframe

Scheduling Methodology

FUJITSU

+key frame information

multiple video stream (15-20 streams / one server)

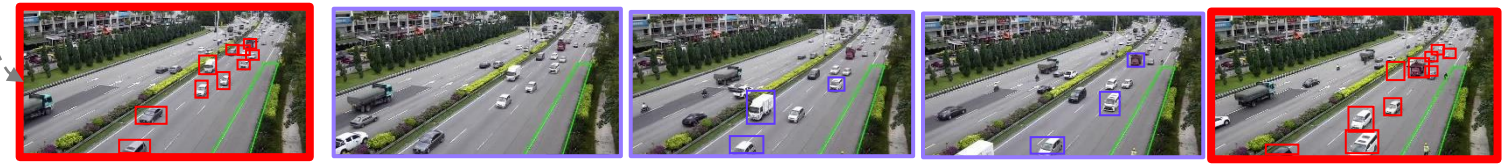
Server

Video decoder

Frame decomposition

CPU GPU scheduler

KFn DFn(1) DFn(2) DFn(k) KFn+1

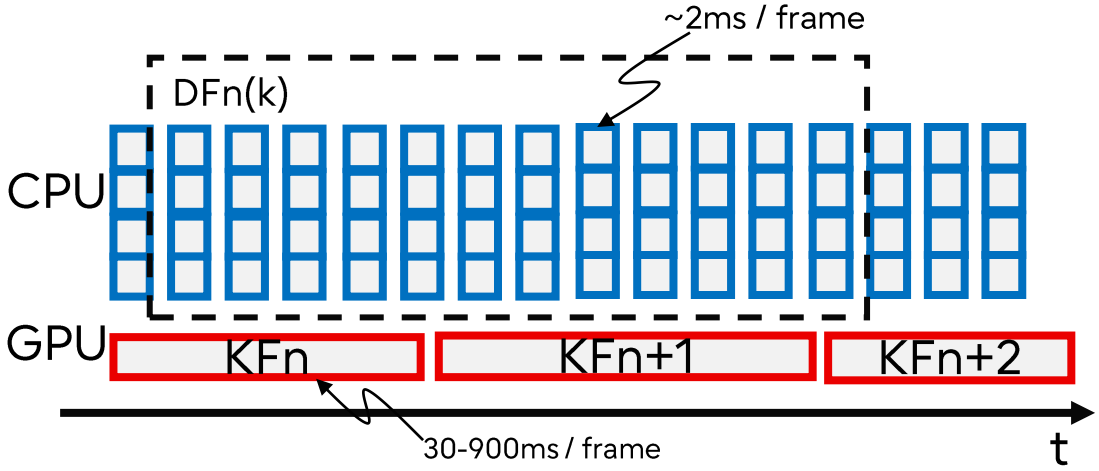


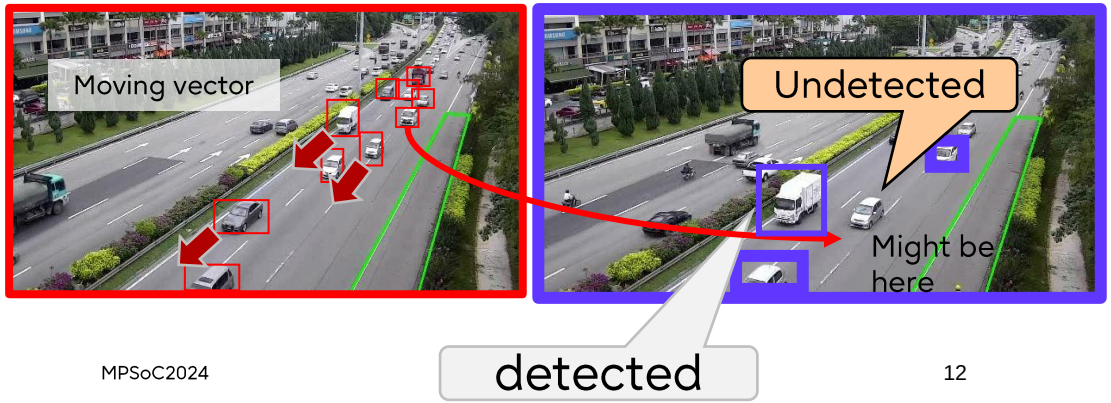
Key frame
The most accurate and slowest DL processing

Dependent frames
Low accuracy but high speed processing (Traditional imaging operation)

Key frame
The most accurate and slowest DL processing

Scheduling Methodology





CPU/GPU Server

- Video decoder
- Frame decomposition
- **CPU GPU scheduler**
- **Undetected object completion**
- Frame reconstruction
- Event Analysis

Effectiveness

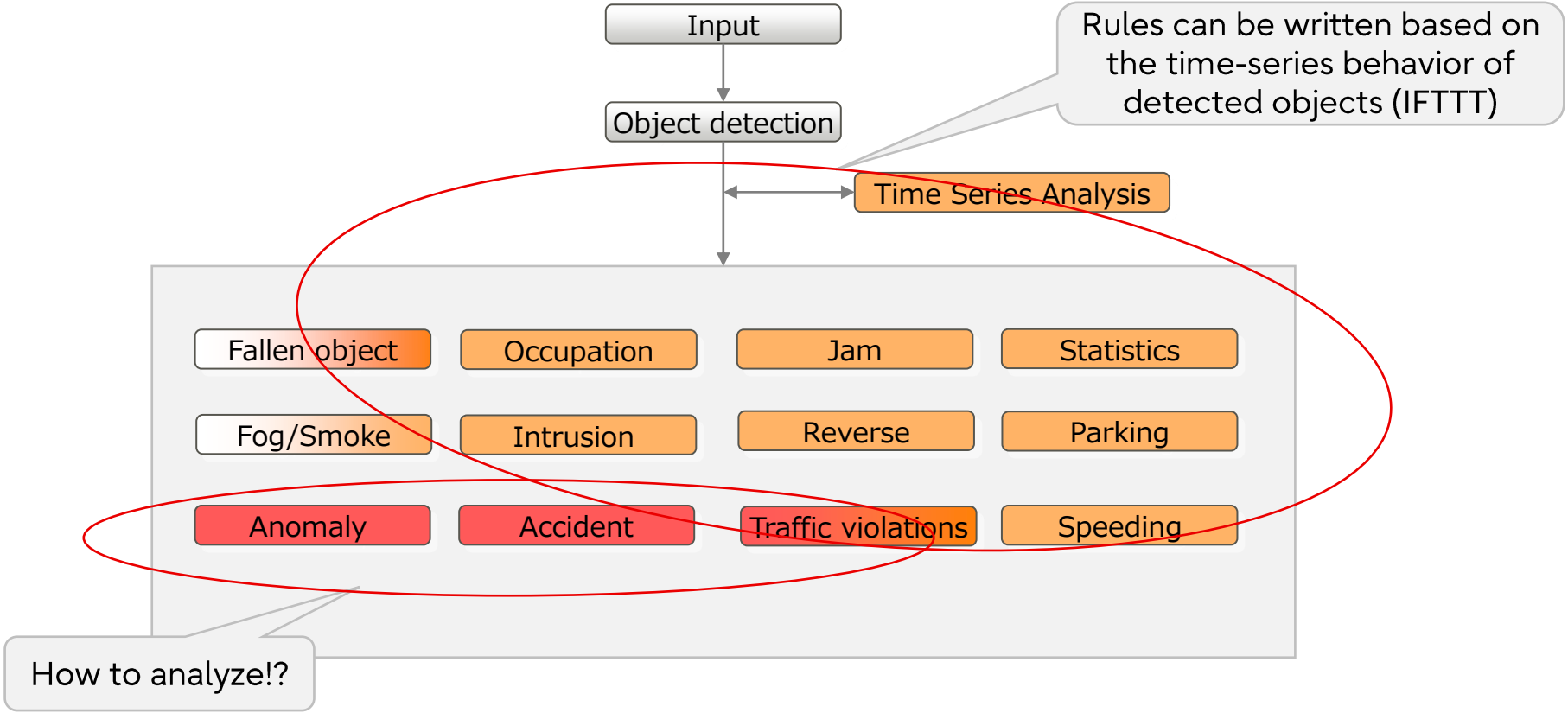
FUJITSU

- When vectors are small, such as in traffic jams, it is difficult to exceed the threshold (the key frame interval becomes longer)
  ⇒In traffic jams, there are many detected objects, DL operation time becomes longer on GPU, during dependent frames on CPU

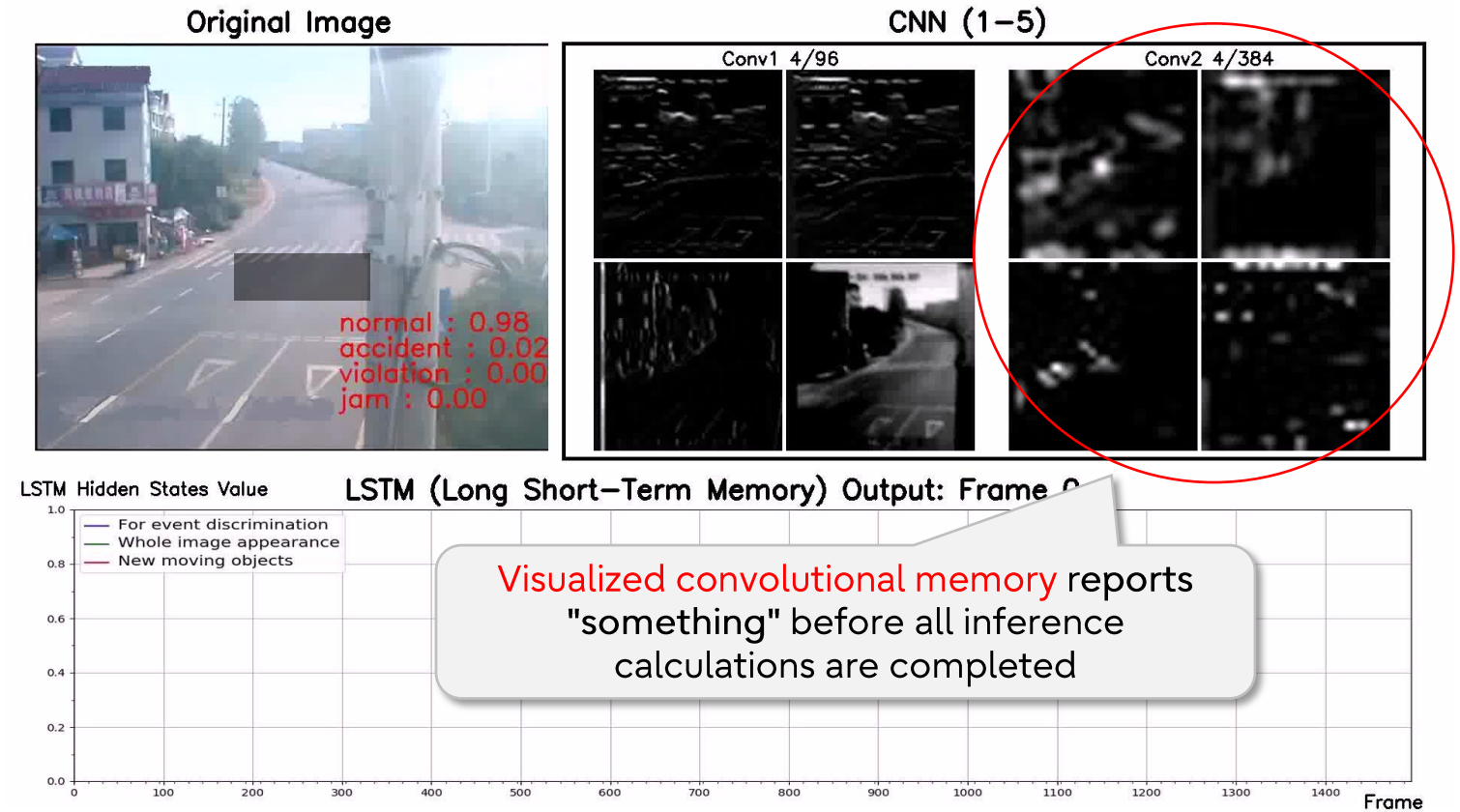
- In a scene with smooth flow, vectors are large and key frames are short-period
  ⇒There are few objects, so analysis can be completed in a short time even with DL

- As a result, the processing time is stable according to the traffic conditions

Functions



Deep Scene Recognition

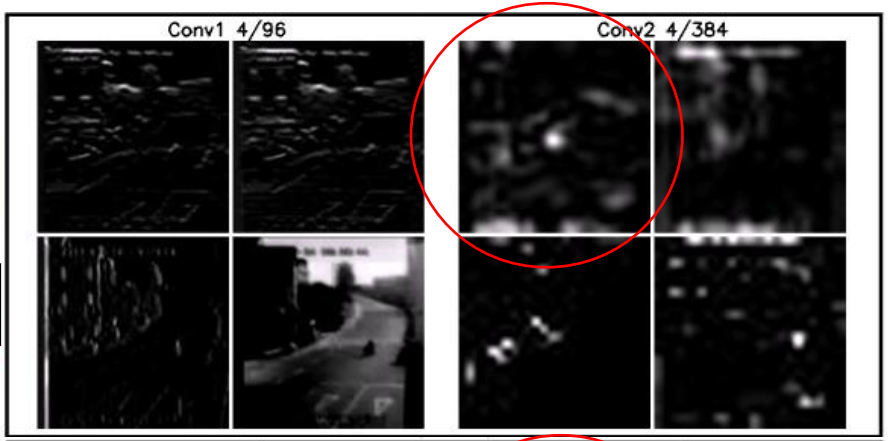


Attention Based Deep Scene Recognition [3]

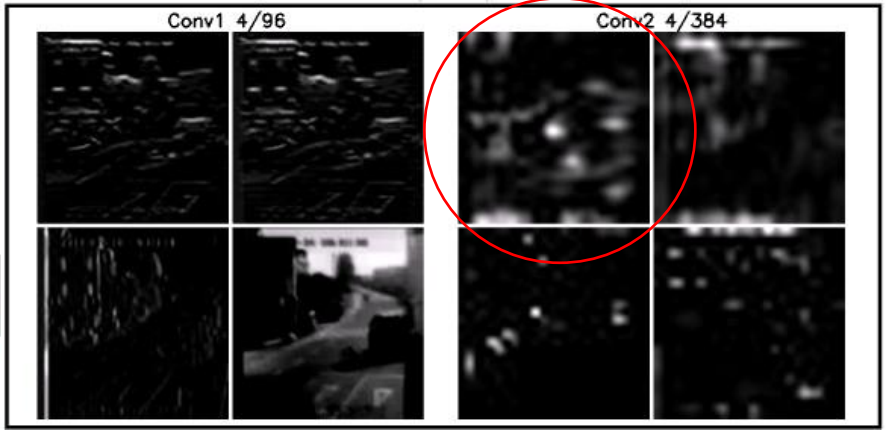
Deep Scene Recognition

FUJITSU

Before the accident



After the accident



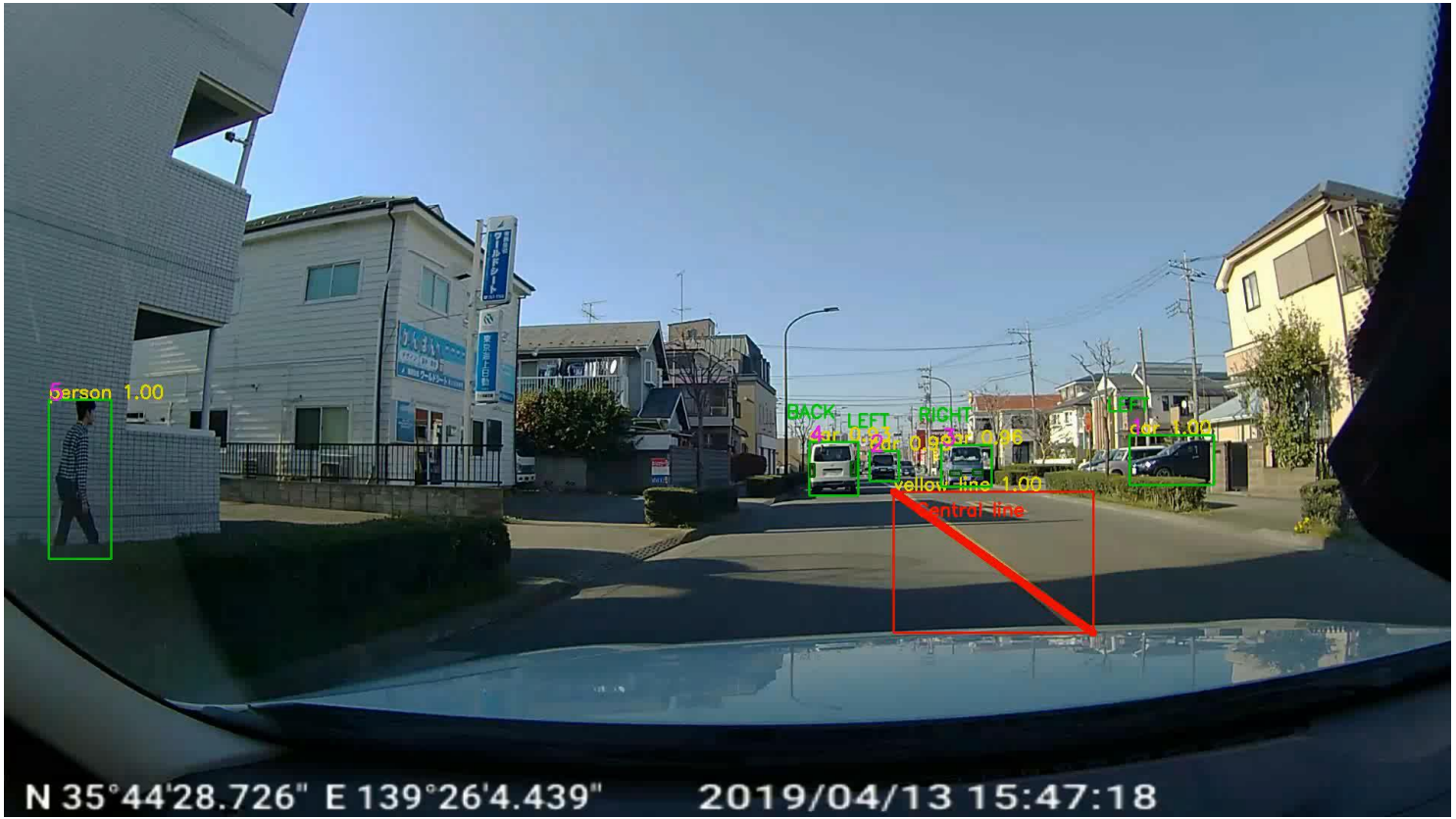
Deep Scene Recognition

FUJITSU

- It is difficult to define what constitutes an "accident" or "some kind of abnormality.“
- Although there is active research such as GPT4V (imaging GPT), it is insufficient in terms of practicality and realistic processing time.
- With traditional DL model, looking at the convolutional memory arranged in chronological order.
- It can be seen some kind of mutation before all the calculations are completed.
- If using this as a trigger, the system recognizes that "something is happening.“
- In reality, it is practical to just raise a flag and have a human judge the situation.

On Vehicle Camera System

Video Sample (Near accident case)



Outline

FUJITSU

- The widespread adoption of autonomous driving is still a long way off
- As long as humans are driving, mistakes will occur, anytime anywhere
- Use of on vehicle camera in insurance arbitration
  - When an accident occurs, it becomes a problem depending on the percentage of responsibility
  - After an accident is reported, insurance companies spend a long time investigating the cause of the accident and preparing documents
  - The evidence prevails (no inferences allowed)

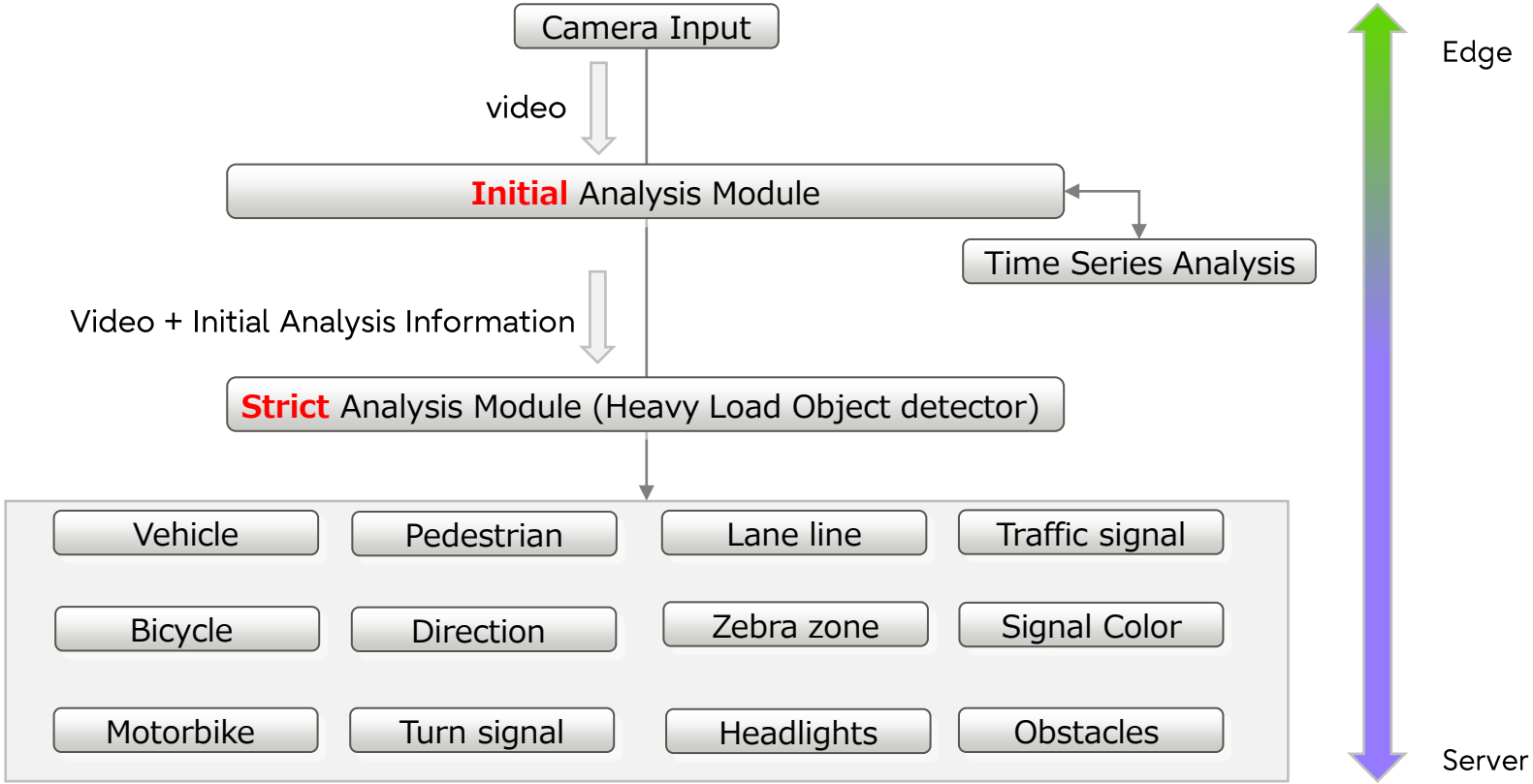
	Roadside Camera	On vehicle Camera (insurance)
Accuracy (Object detection / Scene recognition)	<	
Stability of input/output load	Stable	unstable
Camera position	Fixed (Key – dependent frame operation)	Moving (Full frame DL)
Real-time performance	ms – sec order	None (Min – hour)

Functions

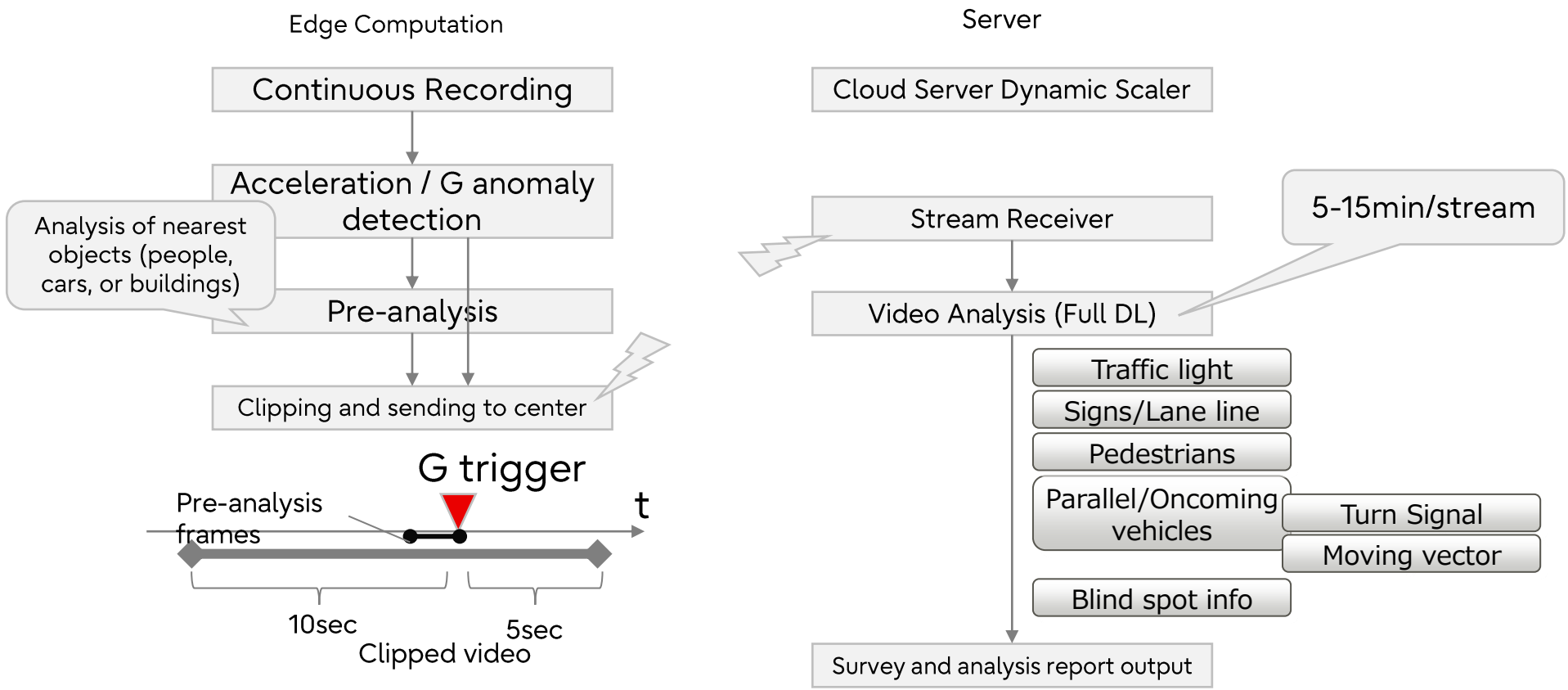
FUJITSU

- Basically, the concept is the same as roadside cameras
- Since there is no RSU on the edge side, it must be completed by the CPU in the camera
- Only the nearest object is identified when an accident occurs (G sensor)
- The first action, police, ambulance, etc., is determined based on whether the initial report is a person, a car, or some other object
- Detailed video analysis is performed when the accident investigation report is created (low real-time performance)

Functions

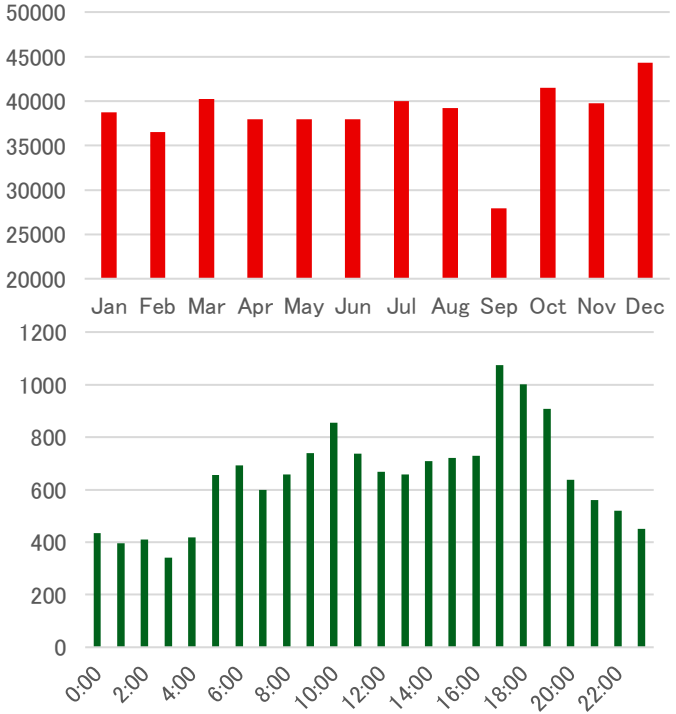


Functions

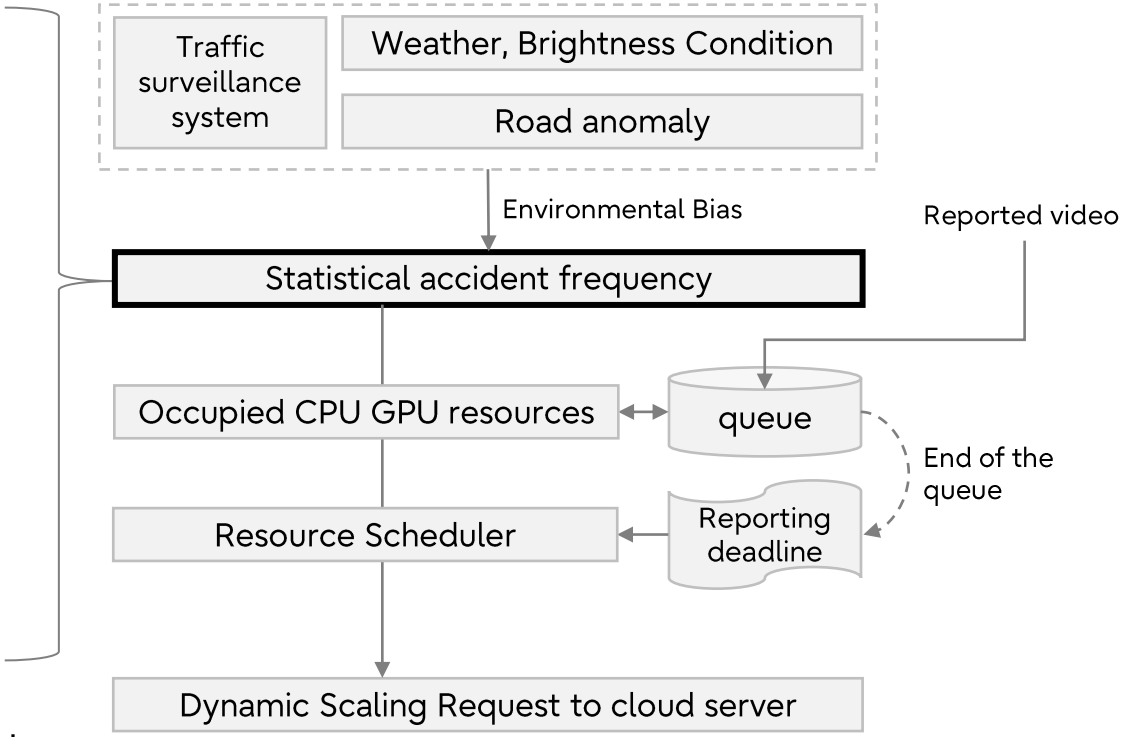


Dynamic Cloud Server Scaler and Scheduler

FUJITSU



Accident frequency varies by time and season(stream arrival load)

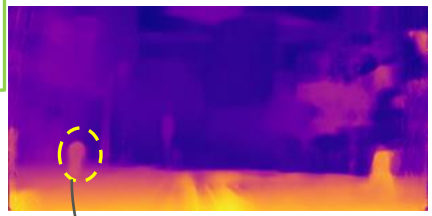


Analysis of the nearest objects

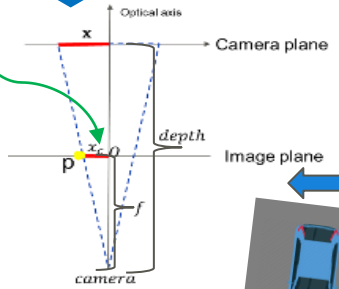
FUJITSU

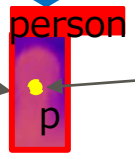
Object detection

- **Bbox;**
- **Category;**
- **Confidence;**



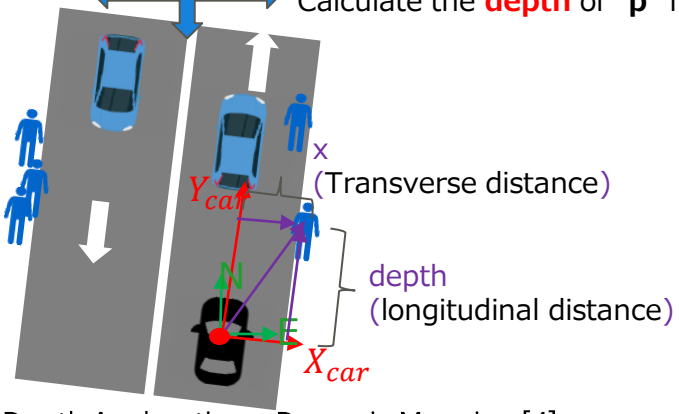
Pseudo depth (distance) analysis





Center point of Bbox

Calculate the **depth** of "**p**" from disparity image;

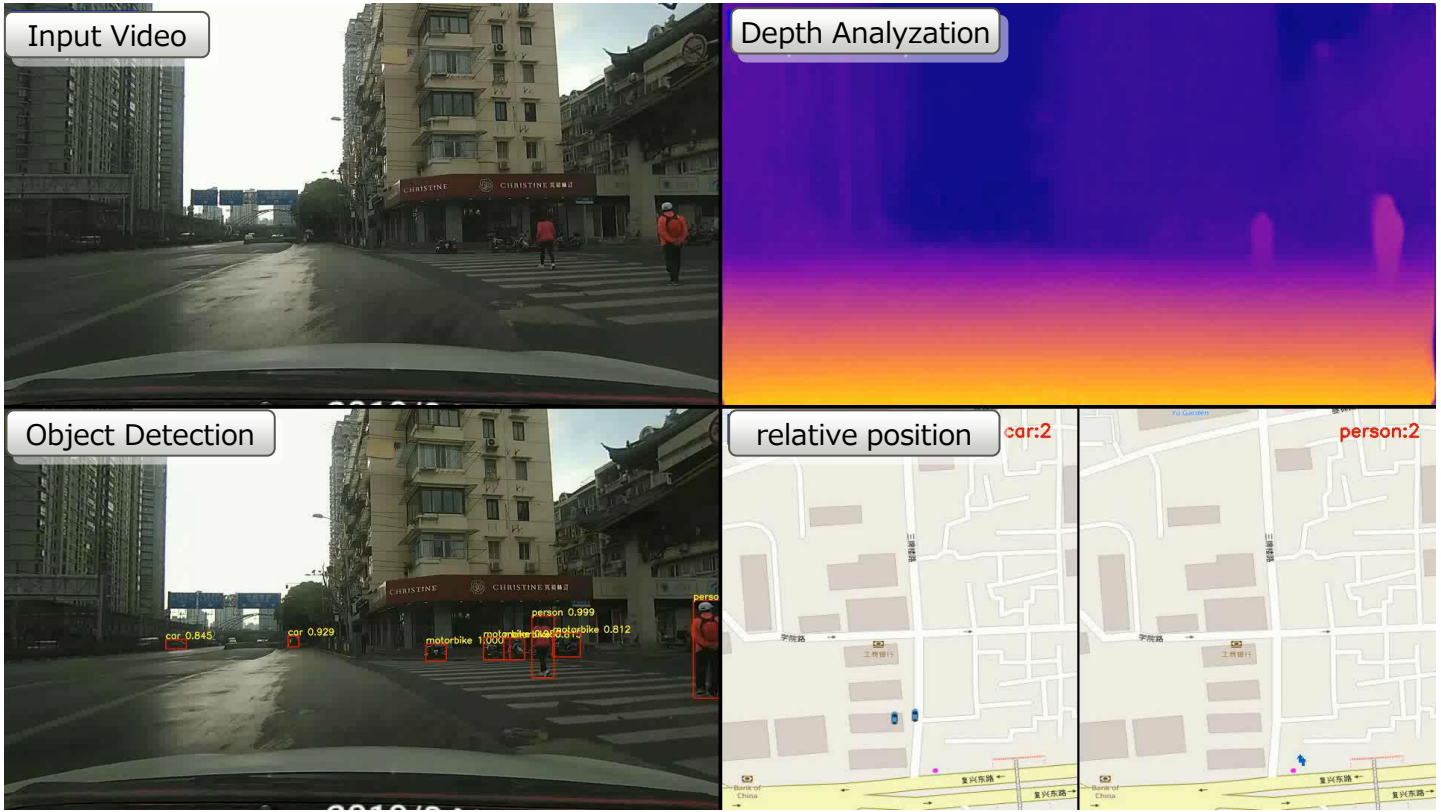


x (Transverse distance)

depth (longitudinal distance)

Object Detection + Depth Analyzation = Dynamic Mapping [4]

Analysis of the nearest objects



Conclusion

FUJITSU

- It is necessary to apply analysis modules according to the purpose of the application.
- Computer-assisted design according to the characteristics of the application.
- Integration of not only video and sensor data but also various related data.

Reference

[1] Seiichi Uchida, “Pattern Recognition and Image Informatics”, 2016
[2] “Ibutuate’s techblog” Attention-guided Context Feature Pyramid Network for Object Detection, 2022 April
[3] Z.Tang, K.Yamashita, et al. ” Deep Scene Recognition with Object Detection”, ASP-DAC 2019
[4] S.Chen, K.Yamashita,et al. “Real-time Object Perception with Accelerated On-vehicle Edge AI”, IPSJ SWoPP 2019

FUJITSU

Thank you