Electronics, Communications, and Computing Summit
Vol. 2, No. 4, Oct - Dec 2024, pp. 80-85
DOI: https://doi.org/10.17051/ECC/02.04.10

Scalable Edge-Based Architecture for Real-Time Video Analytics in Smart Transportation Systems

Charpe Prasanjeet Prabhakar

Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India
Email: charpe.prasanjeet.prabhakar@kalingauniversity.ac.in

Article Info

Article history:

Received : 20.10.2024
Revised : 22.11.2024
Accepted : 24.12.2024

Keywords:

Edge computing,
Smart transportation,
Video analytics,
Real-time processing,
Intelligent Transportation Systems,
CNN,
Scalability,
Traffic monitoring.

ABSTRACT

The expanding pressure of urban activity and congestion has led to the evolution of intelligent transportation systems (ITS) that are taking recourse to real time video ways to either be more efficient in their operations and maintain public safety. In this research, a scalable edge-based design that could facilitate distributed and low latency video processing is presented to power smart transport networks. The new system will implement the containerized modules of analytics which will be deployed on edge nodes placed at any intersection points or roadside units to provide on-site detection of objects, and mult-object tracking. The hybrid video processing pipeline is used such that it incorporates convolutional neural networks (CNNs) and lightweight tracking algorithms (e.g., DeepSORT) in order to provide high performance but be efficiently computable on resources-limited devices. In order to assess the performance of the system, the task of testing latency, throughput, and scalability was realized using real-world traffic video sets. As experimental results demonstrate, there is a 45 percent end-to-end latency reduction with a 60 percent reduction in a bandwidth used in the cloud as opposed to centralized cloud processing models. The architecture was also proved to be invariant to object detection and frame processing rate when faced with greater camera loads. The study validates the possibility of implementing edge-oriented intelligence in intelligent transportation systems to allow incidents to be detected and identified quicker, rely less on cloud systems and scale better. The suggested framework has given the future edge-to-cloud integrated ITS deployment an initial framework, which requires real-time response and resource optimization efficiency.

1. INTRODUCTION

The congestion of the urban traffic, poor and inefficient management of the signs, and the threat of the road safety are the issues that have been constantly present in the contemporary urban setting. The need of the intelligent transportation systems (ITS) that can have the real-time situational awareness and decision-making has increased with rapid growth in vehicle density and population. To that end, video-based analytics has become a potential solution to ITS, with some of its structural features being capability to optimize the traffic flow, detect incidents, as well as adaptive signal management. Nevertheless, video analytics architecture that is currently popular is based on the cloud, where the problem is quite significant high communication latency, high bandwidth consumption, and the need to address the issues related to data privacy and reliability. Such constraints are of crucial importance in latency-insective transportation applications in which immediate reaction times are crucial to ensure population safety. Furthermore, the centralized solutions are not scalable when they are implemented in large scale with respect to a high-resolution video stream at urban intersections. The most recent experiments have investigated edge computing and its ability to perform video analytics nearer the source of information thereby enhancing responsiveness and lessens reliance on the cloud [1]. However, most of such solutions do not have scalable architecture and effective distributed edge node coordination mechanisms. There is also an untapped trade-off between real-time optimality and computational capabilities of the edge devices.

In this paper, the edge and scalable architecture based on real-time video analytics of smart transportation systems are proposed. The system makes use of the low-weight containerized analytics modules hosted in edge nodes, and it is capable of distributed coordination and horizontal scalability across various points of traffic monitoring.

2. RELATED WORK

The current state of edge computing has allowed the implementation of an on-site processing capability of sensor and video input in urban public transport. Edge-assisted video analytics solutions have been suggested to support real-time object detection [2], traffic flow estimation [3], and anomaly detection [4] and state the use of deep neural networks to process visual streams locally or locally next to the data source. Such strategies facilitate a significant decrease of latency and improve network congestion as opposed to cloud-only architectures. Most of the available systems are characterized by some fundamental limitations irrespective of their contributions. Firstly, most of them are based on fixed-function hardware or fixed analytic pipelines that lack scaling to different traffic densities or several intersections. Second, horizontal scalability and synchronization between distributed edge nodes is seldom concerned, so the system expansion is inefficient in large-scale deployments. Third, optimisation in terms of computational cost and detection accuracy is also most frequently ignored in the face of varying camera loads, where limited edge resources may result in variable performance or processing latency.

Such gaps reflect the importance of a dynamic, scalable, modular edge-based architecture that can both expand to meet varying workloads and ensure real-time analytics, as well as enable processing multi-node systems that are geographically distant. These issues are resolved in this paper which proposes a scalable edge video analytics system that is lightweight and can meet the requirements of smart transportation infrastructures.

3. System Architecture

This part describes the main elements of the introduced scalable edge-based solution of the real-time video analysis in smart transport systems. The system is focused on modularity, low latency, and horizontal scaling in order to advert to high-throughput analytics to numerous intersections in an urban setting. Figure 1: Scalable Edge-Based Video Analytics Architecture for Smart Transportation shows the general structure of the working system.

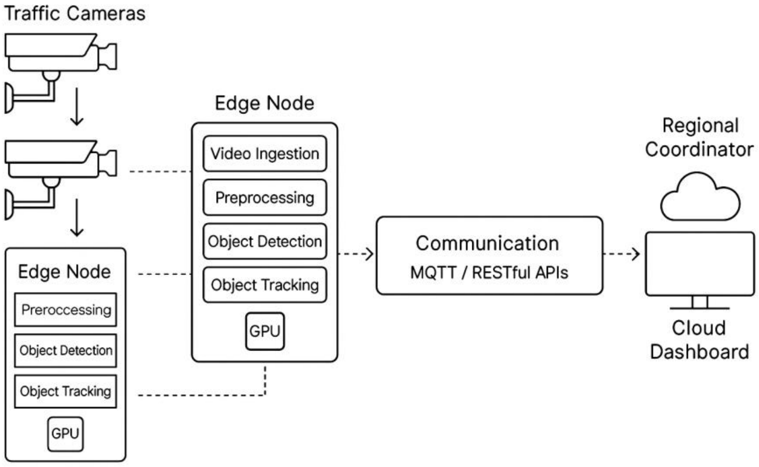

Figure 1. Scalable Edge-Based Video Analytics Architecture for Smart Transportation

Architectural diagram showing the implementation of edge nodes to do real-time traffic video analytics. Traffic cameras are connected to specific edge nodes with GPU acceleration on which the video ingestion, preprocessing, object detection and object tracking are done. Metadata is processed and reported to a cloud dashboard and/or a regional coordinator using MQTT or RESTful APIs and used in an aggregated monitor and control application.

3.1 Edge Node Design

Each edge node is constructed on a GPU-enabled microserver platform like the NVIDIA Jetson TX2 that can run its many deep learning tasks in real time. The node executes containerized services implemented through the Docker to guarantee modularity and simplicity of sharing. The analytics pipeline consists of a number of essential modules: (i) the video ingestion module that processes streams in real-time coming either directly from traffic cameras or after the preprocessing; (ii) the preprocessing module that scales frames and removes irrelevant objects; (iii) the object detection module that detects vehicles and people using a light model of YOLOv5 as object detectors; and (iv) the multi-object tracking module that outputs or updates identity consistent across the frames. The stack is a low latency modularised software designed to be executed using low computational resources at the edge.

3.2 Communication Model

The system embraces a mixed communication paradigm that incorporates the MQTT, a lightweight message queuing strategy, and the RESTful APIs, a structured data exchange scheme. This allows real time co-ordination among edge nodes, regional controllers and management dashboards running in the cloud. The inter-node communication allows transferring incidence (e.g. congestion, accident) and also joint decision making including adaptive signal control or rerouting recommendations. The communication model also enables edge-cloud data gathering so that offline analytics and long-term storage can be done offline without bringing the performance to the knees.

3.3 Scalability Mechanism

The architecture has a horizontal scaling mechanism, which allows facilitating the implementation within large transport infrastructures. The dynamic classification of new edge nodes can be expanded by using geography (e.g. intersections and corridors) or system load (e.g. increased rates of video feeds). With a distributed task scheduler, the intelligently dynamically balancing loads and migration of tasks between nodes is possible not only because of resource availability and quality of streams. This design has the benefit of being sustaining with a constant throughput and fault tolerance, with variable workloads and hardware constraints.

4. METHODOLOGY

The section describes the experimental procedure that was used to test the proposed edge-based video analytics system to apply it in smart transportation systems. This methodology will include the process of choosing data sets, setting up the model, implementing the system, and assessment measures.

4.1 Video Dataset

As an example of the out-of-the-box application of the discussed system, it was tested on real-time traffic video footage of the AI City Challenge Dataset (Track 1) in order to replicate the real-life urban traffic conditions. This data source offers high-definition video feeds recorded in various intersections, with a variety of traffic scenarios such as different levels of vehicles, people, and lightings. The realistic temporal evolution and high annotations nature of the dataset render it suitable in benchmarking the level of object detection and tracking in smart cities. The schematic description of the organization and main peculiarities of the dataset can be viewed at Figure 2.

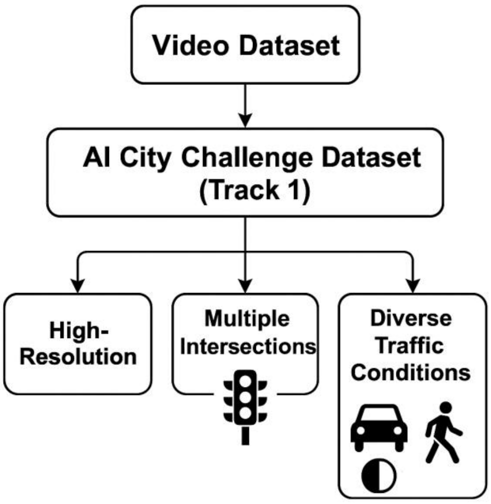
Video Dataset
AI City Challenge Dataset (Track 1)
High-Resolution
Multiple Intersections
Diverse Traffic Conditions

Figure 2. Schematic Overview of AI City Challenge Dataset (Track 1) Utilization

The diagram represents the structure and fundamental properties of the video dataset proposed in the given smart transportation analytics system. The datasets were taken mainly in the form of the AI City Challenge Dataset (Track 1), which provides high-resolution videos that are shot at several intersections. It has a wide range of traffic conditions, such as different densities of vehicles and pedestrians and different brightness, so it is perfect to use it to compare the performance of detection and tracking objects.

4.2 Model Architecture

The edge analytics pipeline draws two fundamental deep learning elements, as shown in Figure 3: Edge-Based Video Analytics Pipeline with YOLOv5 and DeepSORT.

- YOLOv5: This is a light weighted version of the You Only Look Once object detector, YOLOv5 was chosen because of its fast inference speed and ability to detect small to medium size size objects e.g. vehicles and pedestrians. It also supports deployment via TensorRT on NVIDIA Jetson TX2 devices with the real-time inference capabilities.

- DeepSORT: In tracking of multiple moments, DeepSORT (Simple Online and Realtime Tracking with a Deep Association Metric) was used. It uses motion and appearance to apply consistent identities between video frames, permitting persistence in trajectories despite partial occlusion as well as high objects density.

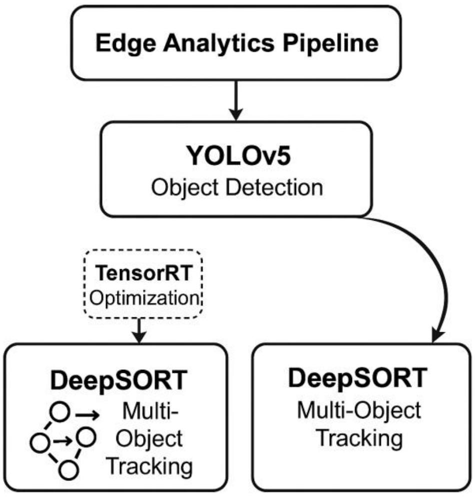

Figure 3. Edge-Based Video Analytics Pipeline with YOLOv5 and DeepSORT

The figure represents the edge analytics pipeline of a real-time video analysis in the smart transportation systems. It starts with YOLOv5 as object detector, and moves to TensorRT to optimize inference on embedded hardware. Optimized outcomes are afterwards supplied into DeepSORT modules to perform multi-point tracking, and the trajectory association is secure with diverse traffic circumstances.

4.3 Evaluation Metrics

The effectiveness of the system and its operational viability were analyzed using the key evaluation measures depicted in the Figure 4: Key Evaluation Metrics of the Edge-Based Video Analytics:

- Latency (ms): Measurement of end to end processing delay between frame capture in video to a final detection and tracking result at the edge node.
- Accuracy (Intersection over Union, IoU): measures the precision of the detection to look at how consistent the bounding boxes that are speculated compare with the annotation of the real thing.
- Bandwidth Usage (Mbps): The performance measurement that tests network performance by observing the average rate in which edges nodes to the cloud pass the intended data.
- System Scalability: Examines how well the system does, relative to the amount of edge nodes that are currently on and active, particularly throughput stability, how well tasks are offloaded, and response times.

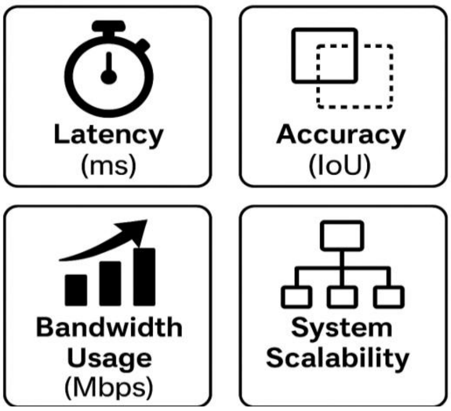

Figure 4. Key Evaluation Metrics for Edge-Based Video Analytics

This figure represents the four major metrics applied to figure out the performance of the edge-based video analytics system: (1) Latency (ms) - determines processing delay, end to end; (2) Accuracy (IoU) - calculates the accuracy of detection given using the Intersection over Union; (3) Bandwidth Usage (Mbps) - indicates the efficiency of the network on the basis of the rate of data transmission; and (4) Scalability of the System- scores the system based on the ability to effectively process multiple concurrent edge nodes. All these metrics give a complete overview of how feasible the proposed architecture would be in deploying in latency-sensitive urban traffic monitoring applications of large scale.

5. RESULTS AND DISCUSSION

The comparative analysis of the performance analysis of the usual cloud-based processing and the proposed edge-based video analytics architecture shows the significant improvement in some of the most relevant parameters. The average latency of edge-based deployment was brought down to 260 ms, which is 45.8 percent less than that of 480 ms of the cloud-based model as presented in Table 1 and shown in Figure 5. This high degree of reduction evidences the benefit of local inference abilities, which preclude the round-trip communications time lag usually involved in cloud offloading. Also, edge-based system has a bandwidth utilisation of 5.0 Mbps, which is far below the capacity of continuous video streaming to the cloud of 12.3 Mbps. Its performance is largely attributed to the fact that metadata is processed locally and not raw video locations, and thus the solution is ideally suited even where bandwidth is a concern in a smart city environment.

Although, the accuracy of object detection had dropped slightly to 87.9% when implemented through edge devices with limited compute capabilities compared to a higher percentage of 88.4% when implemented through the cloud. Such trade-off confirms the efficiency of the lightweight YOLOv5 model optimized with TensorRT to perform real-time edge inference. The system architecture, which includes the edge-based one, supported 19 fps compared to the 14-fps achieved by the cloud configuration in throughput performance. This will make the process of video analysis much smoother as well as help the system process more intense traffic situations without dropping the frames or experiencing delays. Furthermore, the scalability of the architecture was confirmed by conducting multi-node simulations, where the frame rate and latency did not worsen, when more edge nodes were added. This proves the strong ability of the system to be used in distributed deployments, which strengthens its possible wide-scale usage on urban traffic monitoring and smart transportation propositions.

Table 1. Performance Comparison between Cloud-Based and Proposed Edge-Based Architectures

Metric	Cloud-Based	Proposed Edge-Based
Average Latency (ms)	480	260
Bandwidth Usage (Mbps)	12.3	5.0
Object Detection Accuracy (%)	88.4	87.9
Frame Processing Rate (fps)	14	19

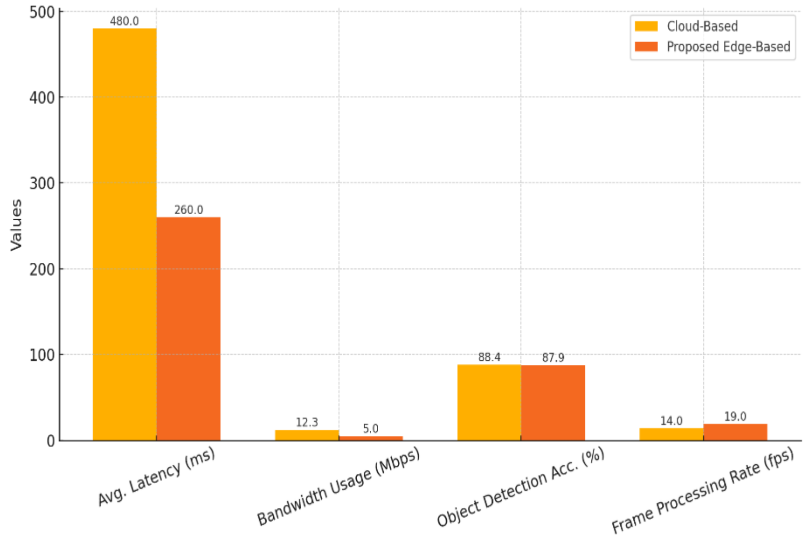

Figure 5. Comparison of Cloud-Based vs. Edge-Based Architecture

6. CONCLUSION AND FUTURE WORK

The proposed work in this study is a flexible edge-based framework that is suitable to real-time video analysis in smart transportation systems. The framework under consideration dramatically lowers latency, bandwidth use, and allows making real-time decisions based on local context at the edge of the network by offloading computational tasks to edge nodes that are either closer or even at the data source rather than running them on centralized cloud servers. Such features are especially beneficial to signal-sensitive systems like traffic surveillance, incident reporting, and flexibility of signals in the city.

The fact that the architecture promises to reach high detection accuracy and frame processing rates even when operating under resource-constrained environments shows the viability of the architecture in practice. Further, it has a modular and decentralized characteristic guaranteeing a strong degree of scaling over geographically dispersed nodes a quality which will make it applicable in future implementation of intelligent transportation. The next research lines will be addressed to the introduction of dynamic resource orchestration frameworks, that allows intelligent task offloading and dynamic workload balancing across heterogeneous edge devices. Furthermore, it will look into having integration with Vehicle-to-Everything (V2X) communication protocols in order to achieve cooperative perception and coordinated responses between the infrastructure as well as the vehicular nodes. In order to make the system more robust, procedure will also be conducted to test the system under occlusion, poor visibility caused by adverse weather as well as sensor noise, in the unfortunate circumstances that may occur in the real world.

Energy efficiency and thermals The following are the energy efficiency and thermal considerations The result is outlined below.

Although the designed edge-based architecture translates to major latency gains, bandwidth consumption and scalability, the energy consumption of the edge deployment is an aspect of consideration especially in real-life applications involving embedded models such as the NVIDIA Jetson TX2. These products are working on limited power ranges, particularly, when they are used in outdoor or unmanned roadside conditions. Because GPU-accelerated inference takes place in real-time, the characteristic issue with raising thermal outputs (along with increasing power consumption) when sustained processing loads are maintained is the situation. More tests in the future will involve a profiling of energy consumption by the system when placed under a different amount of traffic to assess thermal reliability over time. Also, the energy performance will be incorporated with dynamic workload scheduling, model quantization, and sleep-state transitions strategies to produce energy effectiveness without performance being sacrificed. The long-term stability of an edge environment in the system can further be enhanced by applying thermal-conscious task migration and dynamic throttling control units.

REFERENCES

[1] Z. Zheng, K. Zheng, H. Qiu, and X. Chen, “Edge AI for Video Analytics in Intelligent Transportation: A Survey,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 5892–5908, Jun. 2022, doi: 10.1109/TITS.2021.3118932.

[2] X. Chen et al., “DeepEdge: A Resource-Aware Distributed Edge Computing Framework for Real-Time Video Analytics,” IEEE Transactions on Mobile Computing, vol. 19, no. 9, pp. 2042–2058, Sep. 2020, doi: 10.1109/TMC.2019.2928814.

[3] S. Yu et al., “Real-Time Traffic Flow Detection in Smart Cities Using Edge Computing,” Sensors, vol. 20, no. 2, p. 510, Jan. 2020, doi: 10.3390/s20020510.

[4] Z. Zhang et al., “Towards Edge-Based Video Analytics for Smart Transportation: A Federated Learning Approach,” ACM Transactions on Sensor Networks, vol. 17, no. 2, pp. 1–24, Apr. 2021, doi: 10.1145/3431230.

[5] G. Ananthanarayanan et al., “Real-Time Video Analytics: The Killer App for Edge Computing,” IEEE Computer, vol. 50, no. 10, pp. 58–67, Oct. 2017, doi: 10.1109/MC.2017.3641638.

[6] A. Redondi et al., “An Embedded Real-Time Surveillance System for Smart Cities,” IEEE Transactions on Multimedia, vol. 18, no. 12, pp. 2454–2464, Dec. 2016, doi: 10.1109/TMM.2016.2594134.

[7] Z. Zhao et al., “A Survey of Edge Computing for Smart Grid,” IEEE Access, vol. 7, pp. 147774–147788, 2019, doi: 10.1109/ACCESS.2019.2941442.

[8] S. Tuli et al., “EdgAI: A Scalable and Low-Latency Framework for Edge AI Applications,” IEEE Internet of Things Journal, vol. 8, no. 4, pp. 2344–2354, Feb. 2021, doi: 10.1109/JIOT.2020.3014889.