Hierarchical Semantic Gating for Efficient Real-Time Action Recognition in Critical Infrastructure

Julen Beldarrain Portugal[1 [0009−0005−9863−4750]], Javier Calle Armendariz[1,2 [0009−0006−5205−5451]], David Redó[1 [0009−0003−5461−1149]], and Peter Leškovský[1 [0000−0003−4215−051X]]

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua 57, 20009 Donostia-San Sebastián, Spain
{jbeldarrain,jcalle,pleskovsky, dredo}@vicomtech.org
[2] Universidad Politécnica de Madrid, Madrid, Spain
javier.calle.armendariz@alumnos.upm.es

Abstract. Surveillance in Critical Infrastructures (CI) demands real-time understanding of human actions. However, deploying continuous Action Recognition (AR) models on edge devices (such as UGVs or smart CCTVs) presents a severe computational bottleneck. In a target-based approach, the processing cost scales linearly with the number of people in the scene, making standard heavy video classifiers unfeasible for crowded environments or battery-constrained robots.
This position paper proposes a cost-effective perception pipeline designed to operate independently on both fixed cameras and mobile robots. Instead of relying on complex multi-sensor data fusion, we present a hierarchical workflow governed by a "Semantic Gate". This lightweight mechanism utilizes geometric and temporal metadata from the object detector, specifically exploiting the temporal stability of NMS-free end-to-end detectors to suppress routine behaviours. This gating strategy triggers heavy action recognition models only when non-routine dynamics are observed, drastically reducing energy consumption while maintaining responsiveness to critical events like intrusions or fallen person scenarios.

Keywords: Edge Computing · Optimization · Security · Action recognition · UGV · CCTV

1 Introduction

The protection of critical infrastructure (CI) is evolving from passive recording systems to active, autonomous response systems involving collaborative networks of fixed CCTV cameras and unmanned ground vehicles (UGVs). The TESTUDO project aims to secure such facilities by leveraging this hybrid setup. The deployment of fixed CCTVs and UGVs offers a distinct operational advantage through their complementary capabilities. Whilst CCTVs ensure continuous, wide-area

monitoring, UGVs provide the necessary mobility for close-range verification of alerts, mitigation of blind spots, and the maintenance of target tracking beyond the field of view of a single static sensor. Furthermore, UGVs can conduct routine patrols in environments where the installation of fixed infrastructure is unfeasible due to the absence of reliable mounting points or power and network connectivity. By traversing these coverage gaps, UGVs extend the surveillance perimeter, ensuring periodic inspection and an on-demand presence in otherwise unmonitored zones.

A major challenge in deploying AI on the edge is the "Action Recognition Cost". While detecting a person is relatively cheap, understanding their intent (e.g., walking vs. sneaking, sitting vs. falling) requires analysing temporal dynamics using heavy Deep Neural Networks (DNNs) like Video Transformers [1]. In a target-based system, where each detected person is analysed individually, the computational load multiplies by the number of targets. For a UGV operating under strict Size, Weight, and Power (SWaP) constraints, continuously running these heavy models for every person in the Field of View (FoV) is prohibitive.

We argue that the key to efficiency is not simpler models, but smarter activation. We present a hierarchical pipeline where a high-precision "Semantic Gate" filters out routine behaviours (e.g., walking normally) using low-cost geometric data. This design is aligned with broader "conditional computation" ideas (early-exit / dynamic routing), where expensive processing is reserved for harder samples [2], [3]. Modern, temporally stable detectors such as RF-DETR [4] enable this gating mechanism, allowing the system to reserve heavy computation for potential anomalies.

2 Sensor Setup and System Training

To build a robust system suitable for the transition from prototype to operational environments, we define a clear separation of roles between sensors. This avoids the latency and complexity pitfalls of real-time sensor fusion.

2.1 Global Dispatch, Local Inference

We adopt a "Dispatch-and-Verify" paradigm. The system does not fuse video streams; rather, it coordinates physical assets based on independent inference camera detecting an intruder and sending a curved arrow dispatch command to a UGV robot nearby on a grid floor."

- Fixed CCTV (The Spotter): Provides global coverage. Its primary role is to detect a presence in restricted zones or potential anomalies at a distance. Upon detection, it estimates a ground-plane location via planar mapping/homography after camera calibration [5], [6] and sends a high-level dispatch command to the UGV.
- UGV (The Verifier): The robot navigates to the target location. Once visual contact is established, the UGV relies exclusively on its onboard camera

and processing unit to classify the action. This independence ensures that if communication with the central server is cut (a common risk in CI), the robot can still detect a threat and trigger a local alarm.

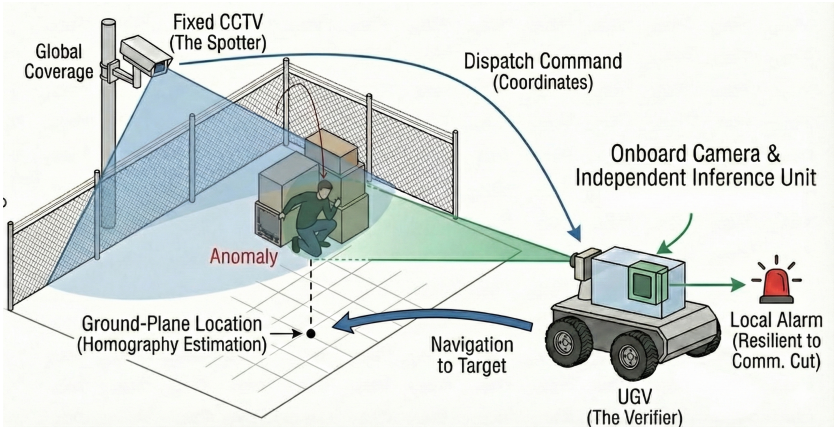

Fig. 1. Operational overview of the "Dispatch-and-Verify" paradigm. The fixed CCTV ("The Spotter") detects a potential anomaly and estimates its geolocation via planar homography. A dispatch command triggers the UGV ("The Verifier") to navigate to the target coordinates for close-range, independent inference, ensuring resilience against communication failures. (Conceptual illustration generated using Google Gemini. Prompt: "A diagram showing a surveillance scene with a fixed CCTV camera detecting an intruder and sending a curved arrow dispatch command to a UGV robot nearby on a grid floor").

2.2 Target-Based Processing

Unlike holistic scene classification, our approach is "Human-Centric" and target-based. The system detects individual humans, tracks them, and analyses their specific actions. While this offers higher precision than analysing the entire scene, it introduces the resources scaling problem mentioned in Section 1. To this end, the Hierarchical Semantic Gating paradigm, described in Section 3, is used to decouple the number of targets from the total computational cost.

2.3 Simulation and Synthetic Data Generation

The primary challenge in adapting general AI tools to an operational environment is the domain shift expected due to the particularities of the end user requirements and final application environment. To circumvent this problem, a data-centric approach focuses on fine-tuning the AI models to operationally relevant, annotated data. In this regard, simulation environments can be leveraged

to augment real-world datasets with high-volume synthetic samples, while effectively alleviating the difficulty of collecting rare anomaly data (e.g., sabotage, falls). Furthermore, simulation resolves privacy issues (GDPR) [7].

A common approach to generate Synthetic data is to use video game engines like GTA V [8], but its limited customization makes it impossible to generate a "Digital Twin". For that reason, we propose utilizing 3D engines such as Unreal [9] or Unity [10] to generate a realistic scenario of the infrastructure and synthetic datasets. This addresses specific challenges of the "Reality Gap" [11]:

- Ego-motion Modelling: Unlike CCTV, UGV cameras vibrate and move. Simulation allows us to generate data with precise camera trajectory ground truth to train the system to distinguish between "camera motion" and "object motion".
- Domain Adaptation: Techniques must be applied to bridge the "Sim2Real" gap, training the DNNs with a mix of synthetic scenarios (smoke, night) and real operational footage [11].

2.4 Hardware Constraints and Sensor Placement

To select suitable UGV hardware, we identify the key constraints that impact edge perception.

- SWaP Constraints: The perception system must share the battery with the robot's locomotion and navigation stack. We define a limited budget for the vision module, targeting embedded platforms like the NVIDIA Jetson Orin NX.
- Optical Considerations: The lens selection significantly distorts the image content. A UGV requires wide-angle lenses for navigation, which introduces barrel distortion affecting the aspect ratio of humans at the image edges. The recognition algorithms must be robust to these geometric deformations.

3 Algorithm deployment

Our criteria for AI-processor selection are compact form factor, low-latency inference support (e.g., TensorRT), and the ability to concurrently process multiple video streams (CCTV and local UGV feed).

3.1 Proposed Architecture: Hierarchical Semantic Gating

Level 1 (Always-on): A lightweight object detector RF-DETR (Nano) [4] and a multi-object tracker such as ByteTrack [12].

Level 2 (The Semantic Gate): Since the gate is the cornerstone of the proposed optimisation, it must achieve high precision and high recall for safety-critical events. Its objective is to suppress routine/normal behaviours with high confidence and to escalate only when target-level evidence suggests non-routine

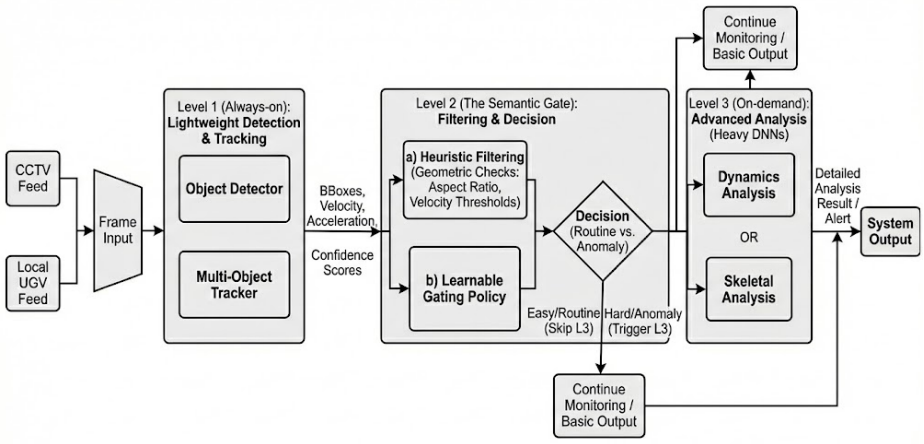

Fig. 2. Conceptual architecture of the Hierarchical Semantic Gating system. Level 1 performs detection; Level 2 filters routine behaviours; Level 3 performs heavy analysis on demand. (Diagram generated using Google Gemini. Prompt: "A hierarchical block diagram showing three levels of processing: Level 1 object detection, Level 2 semantic gating, and Level 3 action recognition").

dynamics. The gate operates exclusively on target-level metadata produced by Level 1 (e.g., trajectories, velocity/acceleration profiles, bounding-box aspect ratio dynamics, confidence margins), ensuring low overhead. We consider two complementary implementations:

a) Heuristic Filtering: A deterministic check of geometric parameters. For instance, an aspect ratio (h/w) dropping suddenly below 1.0 suggests a "fall", while a velocity vector exceeding a normalized threshold implies "running/intrusion."
b) Learnable Gating Policy: To avoid brittle manual thresholds, we propose training a lightweight classifier (e.g., MLP or Random Forest). This is conceptually consistent with conditional computation methods where a learned policy decides whether to invoke expensive processing [2], [3].
   – Inputs: This model receives a vector containing the normalized bounding box coordinates, the acceleration profile derived from the tracker (2nd derivative of position), and the class with the confidence score.
   – Decision: The model is trained to classify the sample as "Easy/Routine" (predictable behaviour) or "Hard/Anomaly" (complex dynamics).
   – Benefit: This allows the system to learn non-linear relationships, for example, distinguishing between a person "sitting on a bench" (routine) versus "collapsing" (anomaly), based on acceleration patterns, without relying on rigid geometric rules.

Gate performance will be evaluated with emphasis on false negatives (missed escalations), since they can make the system blind to subtle anomalies. We will report gate Recall on anomalies, False Escalation Rate (unnecessary Level-3

triggers), and the resulting average energy/latency reduction versus an always-on Level-3 baseline.

Level 3 (On-demand): Only if the Gate detects an anomaly (interaction with object, sudden change in pose), a heavy DNN is triggered, such as MoViNet [13] for dynamics or ViTPose for skeletal analysis [14].

Failure Modes and Safeguards: A high-precision gate introduces a trade-off: subtle, slow-developing anomalies may be misclassified as routine, increasing the risk of false negatives. To mitigate this, we propose (i) periodic forced escalation every n frames/seconds, and (ii) uncertainty-aware escalation, where low confidence margins in Level 1 or ambiguous gate outputs default to Level 3.

An always-on lightweight video model (e.g., MoViNet-A0) remains a strong baseline. However, it requires temporal buffering (multiple frames) and sustained compute, while our approach leverages unavoidable detection and tracking functionalities and triggers temporal models only when target dynamics indicate non-routine behaviour, reducing average compute and reaction latency.

3.2 Preliminary Tests: The Case for RF-DETR

Since the gate relies on subtle target-level dynamics (e.g., aspect ratio and motion derivatives), temporal stability of detections is not optional: noisy boxes can directly translate into false gate activations or missed escalations.

Testing pre-trained generalist DNN models can help us clarify important aspects for the design of the prototype. Specifically, we conducted preliminary observations on the impact of detector temporal stability and generalisation capability on the feasibility of the "Semantic Gating" concept.

We compared the recently proposed RF-DETR [4] against the industry-standard YOLO [15], [16] on video sequences relevant to critical infrastructure.

Key observations:

- Efficiency per Watt: As reported in [4], RF-DETR (Nano) achieves 48.0 AP on COCO [17] compared to YOLOv8 (Nano)'s 35.2 AP, while maintaining comparable latency (2.3ms). This suggests that for the constrained power budget of a UGV, RF-DETR provides significantly richer semantic information.
- Generalisation to "Wild" Data: CI environments differ significantly from standard datasets. RF-DETR demonstrates superior performance on the RF100-VL benchmark [4] , indicating it is less prone to overfitting than YOLO variants when deployed in diverse real-world scenarios.
- NMS-Free Stability: DETR-style set prediction removes the need for Non-Maximum Suppression (NMS) [18], which is commonly used in YOLO pipelines [15]. This is relevant for Level 2 gating: stable, end-to-end predictions increase confidence that aspect ratio or trajectory changes correspond to real physical events rather than post-processing artifacts.

These observations motivate selecting detectors that prioritize temporal stability and generalisation, making end-to-end Transformer-based detectors a compelling option for our hierarchical design [4], [18].

4 Conclusions and future work

In this position paper, we have analysed the main technological factors that the system designers should consider for building a camera-based smart sensing system for CCTV-UGV networks.

We identified that economic and energetic efficiency relies on avoiding redundant computation. We proposed a hierarchical architecture governed by Semantic Gating, where detection and tracking capabilities define an unavoidable baseline cost, and expensive action recognition is selectively triggered only when target-level dynamics suggest non-routine behaviour. This aligns with broader conditional computation principles [2], [3].

Future evaluation should include realistic anomaly benchmarks from surveillance, where weakly supervised or sparse anomaly labels are common, such as real-world CCTV anomaly datasets [19].

5 Acknowledgments

This work was supported by the TESTUDO project (GA 101121258), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

The authors acknowledge the use of Gemini (Google) for proofreading and linguistic refinement of this manuscript.

References

1. Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, New Orleans, LA, USA, Nov. 2022, pp. 10078–10093.
2. S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks," in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Cancun, Mexico, Dec. 2016, pp. 2464–2469.
3. X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "SkipNet: Learning dynamic routing in convolutional networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 440–456.
4. I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, "RF-DETR: Neural architecture search for real-time detection transformers," arXiv preprint arXiv:2511.09554, Nov. 2025.
5. Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
6. R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge, UK: Cambridge Univ. Press, 2004.

7. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, pp. 1–88.
8. M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, and R. Cucchiara, "MOTSynth: How can synthetic data help pedestrian detection and tracking?," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10829–10839. doi: 10.1109/ICCV48922.2021.01067
9. F. Bordes, S. Shekhar, M. Ibrahim, D. Bouchacourt, P. Vincent, and A. S. Morcos, "PUG: Photorealistic and semantically controllable synthetic data for representation learning," arXiv preprint arXiv:2308.03977, 2023.
10. Y.-C. Jhang, A. Palmar, B. Li, S. Dhakad, S. K. Vishwakarma, J. Hogins, A. Crespi, C. Kerr, S. Chockalingam, C. Romero, A. Thaman, and S. Ganguly, "Training a performant object detection ML model on synthetic data using Unity Perception tools," Unity Technologies Blog, Sep. 2020. [Online]. Available: https://blogs.unity3d.com/2020/09/17/training-a-performant-object-detection-ml-model-on-synthetic-data-using-unity-computer-vision-tools/
11. J. Tremblay, T. To, K. Sundaralingam, Y. Luo, E. Dillon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Salt Lake City, UT, USA, Jun. 2018, pp. 969–977.
12. Y. Zhang, C. Sun, Y. Han, Z. Yuan, F. Lei, and Z. Ren, "ByteTrack: Multi-object tracking by associating every detection box," in Proc. Eur. Conf. Comput. Vis. (ECCV), Tel Aviv, Israel, Oct. 2022, pp. 1–21.
13. D. Kondratyuk, L. Liang, Y. Cheng, and S. Christy, "MoViNets: Mobile video networks for efficient video recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, Jun. 2021, pp. 16020–16030.
14. Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, New Orleans, LA, USA, Nov. 2022, pp. 38571–38584.
15. G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," version 8.0.0, Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
16. J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1680–1737, 2023.
17. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, Sep. 2014, pp. 740–755.
18. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, pp. 213–229.
19. W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 6479–6488.