Operational Intelligence through Computer Vision

Evaluating architectures, costs, and deployment strategies for industrial monitoring

A review of the CCTV footage of an assembly line was undertaken to arrive at an ideal approach to build a solution for operational intelligence through computer vision.

Proposed approach

Based on the comparative evaluation of the available options *(covered in the next section)*, an approach comprising YOLO-based object detection, ByteTrack-based object tracking, and a rule-driven event engine - is recommended as the preferred foundation for the proposed solution.

This architecture offers a deterministic and reliable framework for real-time monitoring, structured event generation, and consistent anomaly detection, while maintaining a strong balance between accuracy, operational simplicity, and cost efficiency.

To further enhance system intelligence, a lightweight machine learning model is to be introduced as a secondary layer to identify anomalous patterns that fall outside predefined rule sets. This helps address scenarios where rule-based logic alone may not be sufficient.

Additionally, a Video Language Model (VideoLLM) is to be integrated in an event-driven manner to generate contextual, natural-language descriptions of detected anomalies and operational events. By invoking the model only when required, the solution minimizes computational overhead while reducing the need for manual CCTV footage review.

Overall, the proposed architecture delivers an optimal balance of accuracy, reliability, scalability, cost efficiency, and real-time performance, making it well suited for industrial deployment environments.

June 2026

Executive summary

This report evaluates multiple computer vision and ML-based approaches for real-time industrial monitoring using CCTV video streams. The analysis compares detection, tracking, anomaly identification, and VideoLLM-assisted interpretation frameworks across performance, scalability, implementation complexity, and total cost of ownership.

Research and Analysis

Amish R. Shah
amish@ascentcts.com

Devanandana K P
devanandana@ascentcts.com

Recommended Approach	GPU Cost --- GPU Rating	Operational Cost --- Operational Rating	Total Cost of Ownership	Remarks
YOLO + Tracking + Rule-Based Engine + ML anomaly detection + VideoLLM for anomaly analysis	Medium --- 8 / 10	Medium --- 9 / 10	₹9L - ₹14L	Multiple pipelines running together, complex. Highest sustained GPU load

Evaluation of different approaches

To determine the most suitable implementation strategy, multiple technology approaches were evaluated and compared across key parameters including accuracy, scalability, interpretability, computational requirements, operational cost, and real-time performance.

Approach	Advantages	Disadvantages	GPU Cost --- GPU Rating	Operational Cost --- Operational Rating	Total Cost of Ownership	Remarks
# 1 Classical Computer Vision with Rule-Based Logic	Lightweight, CPU-efficient, real-time processing, easy implementation	Poor identity tracking, sensitive to noise and lighting variations, weak performance in complex environments	Very Low --- 1 / 10	Very Low --- 2 / 10	₹5L - ₹7L	No ML, no Cloud, only fixed rules, so easy to maintain. No GPU needed, CPU-only processing
# 2 YOLO + Tracking + Rule-Based Engine	High accuracy, real-time performance, good balance of speed and interpretability, widely used in industry	Requires manual rule tuning, lacks natural language explanation, limited adaptability to dynamic conditions	Medium --- 6 / 10	Low to Medium --- 5 / 10	₹6L - ₹10L	One ML model to maintain, periodic tuning required. Moderate GPU usage only
# 3 YOLO + Tracking + ML-Based Anomaly Detection	Detects unseen and complex anomalies, learns behavioral patterns from data, more adaptive than rule-based systems	Requires large labeled datasets, higher complexity, difficult to interpret and debug	High --- 5 / 10	Medium to High --- 6 / 10	₹6L - ₹10L	Two ML models used, so more maintainance and more GPU usage

Approach	Advantages	Disadvantages	GPU Cost --- GPU Rating	Operational Cost --- Operational Rating	Total Cost of Ownership	Remarks
# 4 YOLO + Tracking + Rule Engine + VideoLLM (Hybrid)	Strong balance of accuracy and interpretability, provides natural language explanations, supports monitoring and reasoning	Higher latency, increased system complexity, high cost due to LLM integration	High --- 4 / 10	High --- 8 / 10	₹9L - ₹13L	Cloud API usage, storage, monitoring multiple services. Since only YOLO - runs in GPU, LLM handled in cloud, low GPU usage
# 5 End-to-End Video Language Model (No CV Pipeline)	Simplified architecture, strong language understanding, no separate detection / tracking required	Poor accuracy in counting and tracking, inconsistent object identity, unsuitable for real-time industrial systems	Very High --- 6 / 10	Very High --- 9 / 10	₹10L - ₹17L	Continuous API cost, heavy tuning. High compute required if LLM heavily used

Notes:

The ratings for both Operational and GPU cost are assigned on a relative scale from 1 to 10 where 1 / 10 indicates low cost and 10 / 10 indicates high cost.

Factors considered for Operational Cost rating

- Number of AI models in the pipeline (CV, ML, LLM)
- System complexity (single vs multi-stage pipeline)
- Dependency on cloud services (APIs, storage, databases)
- Maintenance effort (monitoring, retraining, updates)
- Integration complexity across components

Factors considered for GPU Cost rating

- Type and number of deep learning models used (e.g., YOLO, anomaly models, VideoLLM)
- Real-time inference load per frame
- Concurrent execution of multiple models on the same edge device
- GPU memory and compute requirements
- Whether computation is local (edge GPU) or offloaded to cloud APIs

Key technologies and concepts

YOLO (You Only Look Once)

YOLO (You Only Look Once) is a state-of-the-art real-time object detection framework designed to identify and localize multiple objects within an image or video stream in a single processing pass. Its ability to deliver high detection accuracy with low latency makes it particularly well suited for industrial monitoring applications that require real-time decision-making.

For the proposed solution, **YOLOv8** has been selected as the primary object detection engine due to its speed, accuracy, and robustness in dynamic operational environments. The model is capable of processing live CCTV feeds and detecting multiple moving objects simultaneously, even under varying lighting and operational conditions.

For each video frame, YOLOv8 generates: (1) Object classification (identifying the type of object), (2) Bounding box coordinates (determining the object's location), (3) Confidence scores (indicating prediction reliability)

These outputs form the foundation for downstream tracking, counting, and event-generation processes.
By converting raw video data into structured, machine-readable information, YOLO enables the system to accurately monitor operational activities and support real-time analytics.

Within the proposed architecture, YOLO facilitates key capabilities such as product detection and counting, object movement analysis, worker presence monitoring, and identification of misplaced or abnormal items. Without an object detection layer, video streams would remain unstructured visual data, limiting the system's ability to derive actionable operational insights.

ByteTrack (Object Tracking)

ByteTrack is a multi-object tracking algorithm designed to maintain the identity of detected objects as they move across consecutive video frames. Working in conjunction with an object detection model such as YOLO, it enables the system to track individual objects throughout their lifecycle within the monitored environment.

In the proposed solution, ByteTrack is used to monitor multiple objects moving simultaneously along the conveyor system. Once an object is detected, the algorithm assigns it a unique identifier and continuously tracks its movement across subsequent frames. This identity is preserved even in challenging scenarios such as temporary occlusions, object overlap, or rapid movement.

ByteTrack achieves this by associating detections from consecutive frames based on positional and motion-based correlations, ensuring continuity and consistency in object tracking.

The integration of ByteTrack provides several operational advantages, including: (1) Accurate object counting without duplication, (2) Continuous tracking of object movement and flow, (3) Consistent identity management across video frames, (4) Improved event generation and anomaly detection accuracy

By transforming isolated detections into continuous object trajectories, ByteTrack enhances the reliability, structure, and overall effectiveness of the monitoring system, making it well suited for industrial conveyor-based operations.

RTSP (Real-Time Streaming Protocol)

RTSP is a network communication protocol used to transmit live video streams from cameras to monitoring or processing systems in real time. Rather than storing footage locally for later review, RTSP enables continuous video streaming with minimal latency. It is widely used in IP cameras and CCTV surveillance systems for live monitoring and video analytics applications.

MOG2 (Mixture of Gaussians 2)

MOG2 is a background subtraction technique commonly used in computer vision to detect motion within video streams. The algorithm continuously learns the static background of a scene and compares incoming video frames against this model. Regions that differ significantly from the learned background are identified as foreground objects, enabling the system to detect moving items within the camera's field of view.

KNN (K-Nearest Neighbours)

KNN is a machine learning-based background subtraction method used to distinguish moving objects from the background in a video stream. The algorithm evaluates previously observed pixel values and determines whether a pixel belongs to the background or foreground based on its similarity to neighbouring observations. This enables effective motion detection in dynamic environments.

PTZ (Pan-Tilt-Zoom) Camera

A PTZ camera is a motorized surveillance camera capable of panning (left and right movement), tilting (up and down movement), and zooming in or out to adjust its field of view. In contrast, a non-PTZ camera remains fixed in position and continuously monitors the same area from a constant viewing angle.

ROI (Region of Interest)

A Region of Interest (ROI) refers to a predefined area within an image or video frame that is selected for focused analysis. By restricting processing to relevant portions of the scene, ROI-based analysis improves computational efficiency and enables more accurate monitoring of specific operational zones.

Disclaimer

This document has been prepared by ASCENT FUTURETECH LLP solely for the use of the intended recipient and is not meant for public circulation, distribution, or reproduction, whether in whole or in part, without prior written consent.

The information contained in this report has been compiled from internal assessments, publicly available sources, industry reports, and other sources believed to be reliable. However, ASCENT FUTURETECH LLP does not represent or warrant the accuracy, completeness, adequacy, or reliability of the information, estimates, opinions, or projections contained herein.

This report is intended purely for informational and discussion purposes and should not be construed as legal, financial, investment, tax, regulatory, or strategic advice. The contents of this document should not be relied upon for making investment, commercial, operational, or policy decisions without undertaking independent evaluation and obtaining appropriate professional advice.

Certain statements contained in this report may include forward-looking observations, projections, estimates, or expectations based on current industry trends, assumptions, and market conditions. Actual outcomes may differ materially due to changes in technology, regulations, geopolitics, market dynamics, infrastructure availability, or other external factors.

ASCENT FUTURETECH LLP, its affiliates, partners, employees, associates, or representatives shall not be held liable for any direct or indirect loss, damage, or consequence arising from the use of, reliance upon, or interpretation of this report or any information contained herein.

The references to companies, technologies, infrastructure projects, countries, or strategic initiatives in this report are for illustrative and analytical purposes only and do not constitute endorsements, recommendations, or affiliations unless explicitly stated otherwise.

Readers are advised to conduct their own independent analysis and verification before acting upon any information contained in this document.

ASCENT FUTURETECH LLP
LLPIN AAF-3423

150 Parivartan, # 4120, Vallabh Baug Lane, Ghatkopar East, Mumbai 400075, Maharashtra, India
ascentft.com • research@ascentft.com