International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Depth-Aware, Asynchronous Video Analytics for Production: The Intellivision Platform

Arjun Nagulapally[1], Ravi Katukam[2], Aditya Dubey[3]

[1]Chief Technology Officer, AionOS

[2]Principal Data Scientist, AionOS

[3]IT Intern, Summer 2025, AionOS

ABSTRACT

IntelliVision is a production-ready platform for video analytics and face authentication supporting people counting, vehicle counting and automatic number plate recognition (ANPR), parking analysis, emergency flow detection, lobby/crowd thresholding, pothole detection, wildlife/pest monitoring, food-waste estimation, and room readiness analysis. We present a generalizable methodology - asynchronous orchestration with job queues, GPU-aware inference, model selection with resilient fallbacks, and task-specific post-processing - aimed at robustness under constrained resources. The methodology incorporates (i) depth-aware detector selection that adapts to scene structure; (ii) lifetime filtering plus hierarchical identity merging to reduce ID switches and overcounting in crowded scenes; and (iii) slot-gated admission and lazy model initialization to manage resource constraints. We discuss design trade-offs and limitations. Implementation details (APIs, deployment, developer tooling) are intentionally minimized to foreground the research contributions.

Keywords: video analytics, object detection, tracking, ANPR, depth estimation, Celery, GPU, YOLO, RT-DETR, Re-ID, Qdrant, InsightFace

INTRODUCTION

The proliferation of video analytics across retail, transportation, infrastructure, and conservation has created an urgent need for platforms that perform reliably under real-world constraints. Applications vary widely in perspective, lighting, density, and motion, exposing brittle behavior in single-model pipelines and lab-only prototypes.

Despite progress in detection and tracking, a gap remains between algorithmic sophistication and production reliability. In practice, systems struggle with four persistent issues: lack of principled, scene-aware model selection for mixed-depth content; unstable identities under occlusion and detector variance; inadequate GPU-aware backpressure for bursty workloads; and brittle operations without a clear fallback taxonomy or graceful degradation. These challenges are most acute in crowded, mixed-depth scenes and on resource-constrained nodes.

IntelliVision addresses these constraints with a production-oriented methodology: asynchronous orchestration with slot-gated admission; depth-aware detector selection guided by monocular depth statistics; and lifetime-aware, hierarchical identity merging. Resilient model management (lazy initialization with a fallback taxonomy) and task-specific post-processing round out a unified approach designed for robustness on single-GPU deployments.

METHODOLOGY

Having established the core challenges and our approach to addressing them, we now detail the methodological framework that underlies IntelliVision's design. This section articulates the core design choices and their rationale, abstracted from implementation specifics, while providing the necessary context for understanding how our system operates in real-world deployment scenarios.

Our methodology is built around four foundational principles that address the problems identified in the previous section: (1) asynchronous orchestration for handling bursty workloads, (2) depth-aware model selection for adaptive detection, (3) hierarchical identity management for stable tracking, and (4) comprehensive fallback strategies for operational resilience.

Operational details for orchestration, job queues/backpressure, lazy model initialization, and error/fallback handling are summarized below under Processing Flow and system considerations.

Task families share a common backbone; domain-specific rules and post-processing are summarized below.

Task Families and Primary Techniques

- People counting: YOLO/RT-DETR with depth-informed selection; BoT-SORT/ByteTrack; lifetime filtering; re-ID–assisted hierarchical merging for stable counts in mixed-depth scenes.
- Emergency flow: bi-line crossing with in/out event counters and fast-motion flags for timely, low-false-positive alerts.
- Lobby thresholds: polygonal zones with threshold occupancy, dwell summarization, and alert aggregation.
- Food-waste estimation: per-item estimates aggregated to plate/scene metrics for operational reporting.
- Room readiness: checklist scoring with per-issue suggestions and auditable outputs.
- Vehicles and ANPR: vehicle detection/tracking; plate detection + OCR; plate locking across frames; parking occupancy and dwell-time estimation.
- Pothole detection: image/video modes with schema-validated outputs for maintenance prioritization.
- Wildlife/pest monitoring: custom domain models with general-model fallbacks where appropriate.

Face authentication complements analytics with robust embeddings (InsightFace) and vector search (Qdrant). Simple but effective photometric preprocessing (bilateral filtering, sharpening, brightness adjustments) improves detection in challenging lighting; similarity thresholds reject weak matches.

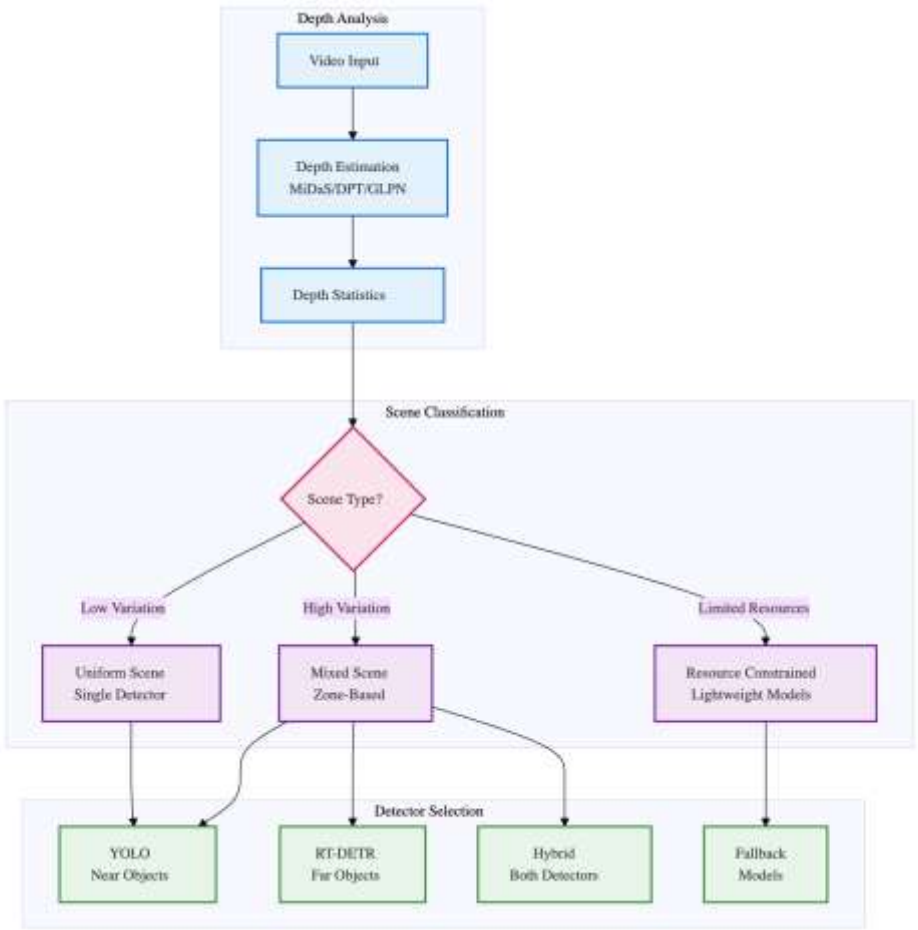

Figure 1 (depth-aware detector selection workflow)

Building upon the foundational methodology outlined above, we now present the specific technical innovations that distinguish IntelliVision from existing video analytics systems. These contributions directly address the core problems

outlined in the Introduction, providing concrete solutions validated in production deployments. Two primary techniques operate in concert: depth-aware detector selection and hierarchical identity merging.

Figure 1 summarizes our depth-aware strategy: monocular depth statistics drive detector choice. Uniform scenes use a single detector matched to dominant scale; mixed-depth scenes partition into near/mid/far zones with per-zone detectors; resource-constrained settings downshift to lighter models with resilient fallbacks. Cross-zone merging avoids duplicates at boundaries. Example: a storefront camera uses YOLO near-field and RT-DETR for distant aisles to maintain recall while matching latency to object scale.

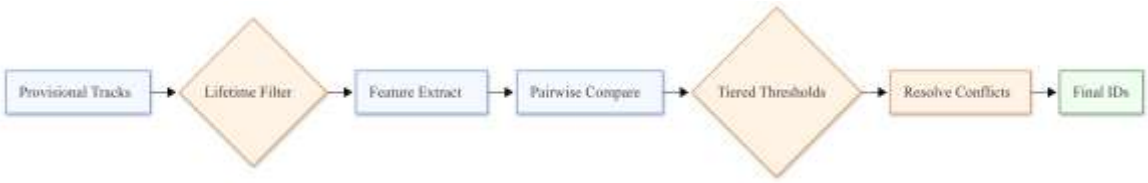

Figure 2: Identity Merging and Lifetime Filtering

Figure 2 outlines a gated clustering pipeline: lifetime filtering removes short tracks; stable features are extracted; pairwise affinities combine appearance, spatial, temporal, and motion cues; and compact rules resolve merges and conflicts. In practice, high similarity merges directly; medium similarity requires spatial–temporal agreement; low similarity additionally requires motion consistency. This consolidates fragmented tracks in crowded, mixed-depth scenes and stabilizes counts.

Exemplar job types include: people count (depth-informed detector selection with tracking and hierarchical merging for stable counts in mixed-depth scenes), emergency count (line-based flow with in/out and fast-motion events for timely, low-false-positive alarms), lobby monitoring (polygonal zones with thresholded occupancy and alert summarization), and face login (robust embedding extraction with photometric normalization and vector search, with explicit rejection of weak matches).

PROCESSING FLOW

The processing pipeline (Figure 3) proceeds from ingestion to outputs with GPU-aware admission and resilient model management. An ingress service validates inputs and assigns jobs; slot-gated admission provides backpressure on single-GPU nodes. For remote media, preview-first validation surfaces a thumbnail and metadata immediately while deferring full downloads to workers.

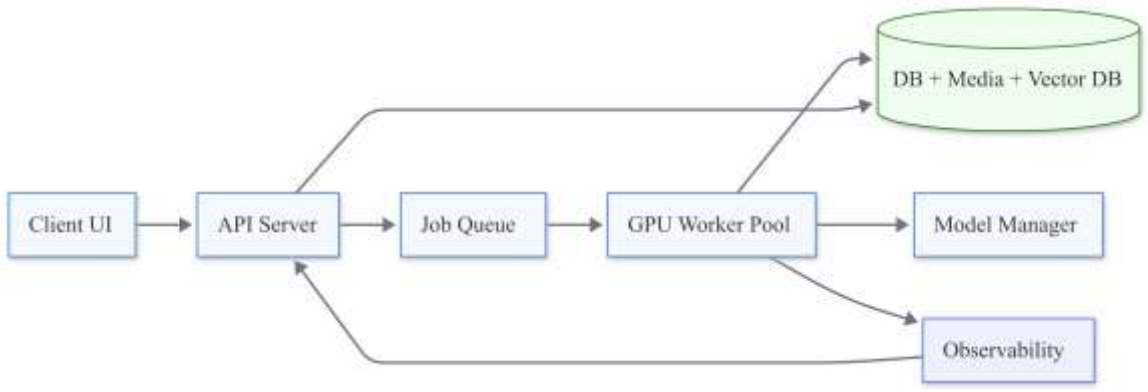

Figure 3 (processing pipeline: ingestion → model selection → tracking → outputs)

A centralized model manager enforces a fallback taxonomy and lazy loading, ensuring detectors, depth, and re-ID models are available without destabilizing memory. Monocular depth statistics drive detector selection (uniform vs. zoned strategies). Detections feed tracking; lifetime filtering and a hierarchical merger consolidate identities before task-specific post-processing produces structured outputs and annotated media.

System and Operational Considerations

- Orchestration and queue management: a lightweight ingress records jobs and delegates heavy compute to workers; slot-gated admission and conservative concurrency provide backpressure to avoid OOM on single-GPU nodes.
- Lazy model initialization: load less-common models on demand; pre-warm critical models to bound cold-start latency and reduce variance.
- Model availability and fallback: a centralized model manager caches detector/depth/re-ID weights; apply resilient fallbacks (e.g., YOLO substitution; MiDaS→DPT→GLPN→geometric) with explicit error taxonomies (GPU vs. model loading) for graceful degradation.
- Remote media handling: validate metadata early, defer downloads to workers to reduce frontend memory pressure and improve perceived responsiveness.
- Deployment contexts and constraints: retail, transportation hubs, industrial, and conservation scenarios impose different accuracy/latency/reliability targets; design for 8–16GB VRAM nodes with bursty workloads.
- Integration and observability: modular APIs for existing systems; structured logging/metrics for health, progress, and rapid diagnosis.
- Track-level analysis complements frame-level detection when identity stability matters.

RESULTS AND DISCUSSION

Performance: The system demonstrates robust operation across diverse scenarios with depth-aware detection and hierarchical identity merging providing stable tracking in mixed-depth, crowded scenes. The asynchronous orchestration with slot-gated admission ensures reliable operation under resource constraints.

Having detailed our technical contributions and processing architecture, we now examine the practical implications and operational characteristics of the IntelliVision methodology. This section synthesizes our design decisions and discusses their impact on real-world deployment scenarios, providing insights into the trade-offs and considerations that shape production video analytics systems.

Our discussion focuses on three key areas: the design rationale behind our core innovations, the operational considerations that enable reliable deployment, and the limitations that bound the current approach. This analysis provides a foundation for understanding both the capabilities and constraints of our methodology, while highlighting opportunities for future development.

Design Rationale: The depth-aware detector selection approach adapts to scene structure by choosing appropriate models based on monocular depth statistics. This provides a principled way to handle mixed-depth scenarios where single detectors may struggle. The hierarchical identity merger combines multiple signals (re-ID similarity, spatial proximity, temporal consistency, and motion prediction) to stabilize track identities and reduce fragmentation.

Operational Considerations: The asynchronous orchestration with slot-gated admission provides predictable performance under resource constraints. Model fallback hierarchies ensure continued functionality when preferred models are unavailable. Early remote-media validation improves perceived responsiveness by enabling previews before full processing.

Limitations: Fallbacks maintain functionality but may require custom weights for specific domains (e.g., regional plates or specialized fauna). Heuristic depth and readiness rules may require site-specific adaptation. Single-GPU orchestration limits parallelism; more sophisticated GPU-aware scheduling represents a natural extension.

Ethical and Privacy Considerations: Face embeddings and images are sensitive; adopt data minimization, retention limits, and encryption in use and at rest where possible. Ensure signage/consent policies match regional regulations for video analytics.

CONCLUSIONS

This work presents IntelliVision, a production-ready video analytics platform that addresses the fundamental challenges of deploying sophisticated computer vision algorithms in real-world environments. Through analysis of the problem space, technical contributions, and operational considerations, we demonstrate that robust systems emerge when operational concerns are treated as first-class design goals rather than afterthoughts.

Our core contributions - depth-aware detector selection, hierarchical identity merging, resilient model management, and asynchronous orchestration - provide a foundation for building reliable video analytics systems that adapt to diverse scenarios while maintaining consistent operational characteristics. These innovations generalize beyond the specific tasks presented, offering a methodology applicable to a wide range of video analytics applications.

Looking forward, several areas present opportunities for further development. Principled model selection for depth and scene type could yield more sophisticated adaptation strategies, while GPU-aware job scheduling could enable more efficient resource utilization. Stronger privacy controls for embeddings and media will become increasingly important as regulations evolve, and broader task coverage will expand applicability to new domains.

The success of IntelliVision in production deployments indicates that the gap between academic research and industrial practice can be bridged through careful attention to operational requirements and systematic approaches to robustness. As video analytics systems become increasingly critical to modern applications, the principles and methodologies presented here offer a foundation for building systems that meet the demanding requirements of real-world deployment.

TERMINOLOGY USED

ANPR (automatic number plate recognition); Re-ID (appearance based re-identification); YOLO (You Only Look Once detector family); RT-DETR (real-time DEtection TRansformer); MiDaS/DPT/GLPN (monocular depth models); BoT-SORT and ByteTrack (tracking baselines); OSNet (omni-scale Re-ID network); Qdrant (vector database for embeddings); OOM (out of memory), FPS (frames per second), GPU (graphics processing unit), MPS (Apple Metal Performance Shaders). We also refer to lifetime filtering (removing short-lived tracks), plate locking (aggregating plate reads across frames), slot-gated admission (GPU backpressure), preview-first processing (thumbnail/metadata before full download), depth zones (near/mid/far partitions), and hierarchical merging (consolidating tracks via appearance, spatial, temporal, and motion cues).

REFERENCES

- Detectors and Depth: Ultralytics YOLO (https://github.com/ultralytics/ultralytics); RT-DETR (https://github.com/lyuwenyu/RT-DETR); MiDaS (https://github.com/isl-org/MiDaS); DPT/GLPN (https://huggingface.co/).
- Tracking and Re-ID: BoxMOT (BoT-SORT/ByteTrack) (https://github.com/mikel-brostrom/BoxMOT); OSNet (via deep-person-reid and yolo_tracking derivatives).
- Production and Orchestration: Celery (https://github.com/celery/celery); NVIDIA DeepStream SDK; Intel OpenVINO Model Server; GStreamer (https://gstreamer.freedesktop.org/).
- Model Libraries: OpenMMLab MMDetection (https://github.com/open-mmlab/mmdetection); Detectron2 (https://github.com/facebookresearch/detectron2).