iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82905

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Real-Time Intelligent Video Analytics System for Automated Surveillance, Facial Identification, and Human Activity Monitoring

Novonita Sen Gupta[1], Aman Govind Rao[2], Sneha Kumari Roy[3], Shreya Pandey[4], Sanskriti Gupta[5], Dr. Akash Punhani[6]
Department of Computer Science & Engineering IILM University, Greater Noida, India

Abstract: ***Rapid urbanisation and the growth of smart city infrastructure have put heavy pressure on traditional surveillance systems, which were never designed to handle current volumes and types of threats.[file:1] Conventional CCTV still relies on human operators to continuously watch multiple screens, flag incidents, and raise alarms, which does not scale to large camera networks.[file:1] Human attention typically drops after about twenty minutes of continuous monitoring, and feeds from hundreds of cameras cannot realistically be reviewed in real time.[file:1] Recorded footage also lacks structured metadata, so it cannot be searched or analysed automatically.[file:1]***

To address these limitations, we built the Real-Time Intelligent Video Analytics System (RIVAS), which integrates three deep-learning models into a single, end-to-end pipeline.[file:1] YOLOv8 is used to locate people in each video frame, DeepFace with an ArcFace backbone identifies who they are, and MediaPipe Holistic estimates body posture to infer activities.[file:1] As each frame arrives, RIVAS detects all visible persons, checks whether their faces match an enrolled gallery, determines whether the observed activity is a fall, an unauthorised intrusion, loitering, or normal movement, and immediately pushes a structured alert via a Flask REST API to a Streamlit dashboard.[file:1] On a mixed indoor-outdoor dataset, the system achieved 75% precision for person detection, 95% for face identification, and 94% for activity classification while sustaining 24 frames per second at 720p resolution.[file:1] These results compare favourably with single-module systems and conventional CCTV installations and represent a practical step toward more autonomous, real-time surveillance.[file:1]

Index Terms: Artificial Intelligence, Computer Vision, YOLOv8, DeepFace, MediaPipe, Smart Surveillance, Face Recognition, Human Activity Recognition.

I. INTRODUCTION

Continuous environment monitoring is no longer limited to high-security government facilities; it has become central to how many modern sites operate.[file:1] Airports, university campuses, industrial plants, shopping centres, and hospitals all depend on video-based surveillance not only for post- incident review but also for access control and early detection of problems.[file:1] Industry forecasts estimate that the global video surveillance market will exceed USD 145 billion by 2029, reflecting the scale of investment in this area.[file:1] However, most deployed systems still rely on the same basic pattern used decades ago: a wall of screens, one or more human operators, and the hope that no critical events are missed when attention wanes.[file:1] Research on sustained vigilance shows that detection per- formance declines sharply after around twenty minutes of uninterrupted monitoring.[file:1] Reviewing archived footage from large, multi-camera deployments is also slow and labour- intensive, which means investigations often take longer than necessary.[file:1] Critically, the video streams in these systems are not “smart” in any meaningful way; they are raw footage without embedded metadata, searchable structure, or automatic anomaly detection.[file:1]

Several technological developments now make more in- telligent, automated surveillance feasible. GPU acceleration allows complex neural networks to run on live video feeds without extremely expensive specialised hardware.[file:1] The YOLO family of detectors has greatly reduced the latency between frame arrival and object localisation.[file:1] Modern face recognition pipelines are accurate enough for use outside laboratory conditions, and skeleton-based pose estimation can now operate at interactive frame rates on consumer hard- ware.[file:1] While each of these advances is well established individually, most previous systems have not tightly integrated them into a single pipeline where each stage feeds into the next.[file:1]

This paper makes four main contributions.[file:1] First, we present RIVAS, a fully integrated, real-time pipeline that combines person detection, face recognition, pose-based ac- tivity estimation, and rule-based threat classification into a single coherent system.[file:1] Second, we describe the design decisions behind each stage in detail, including preprocessing, model choices, and

©IJRASET: All Rights are Reserved

the communication protocol between the inference engine and monitoring dashboard.[file:1] Third, we report module-level and system-level results, including per- class precision, recall, F1-score, throughput at different reso- lutions and camera counts, and the confusion matrix for the four-class activity classifier.[file:1] Finally, we give a practical discussion of current limitations and planned improvements, such as night-vision support, mask-aware face recognition, IoT sensor fusion, and multi-camera re-identification.[file:1]

The remainder of this paper is organised as follows. Section II reviews the most relevant prior work. Section III describes the proposed architecture in detail. Section IV presents and discusses the experimental results. Section V concludes and outlines future work.[file:1]

II. RELATED WORK

Automated surveillance sits at the intersection of object detection, face recognition, and human activity recognition, each with a long research history.[file:1] Table I summarises representative works and their key limitations.[file:1]

TABLE I
REPRESENTATIVE PRIOR WORK IN SURVEILLANCE-RELATED VISION

Authors	Approach	Acc.	Limitation a simpler alternativ
Viola & Jones (2001)	Haar + AdaBoost	95%	Poor on non-frontal faces
Dalal & Triggs (2005)	HOG + SVM	∼85%	Fixed-size sliding window
Zivkovic (2004)	GMM	∼80%	Computationally expensive
Lowe (2004)	SIFT	90%	Poor for large crowds
Horn & Schunck (1981)	Optical Flow	70%	Noisy masks in complex scenes
Krizhevsky et al. (2012)	AlexNet	84.7%	Weak on low-resolution faces
Ren et al. (2015)	Faster R-CNN	73.2%	Multi-stage pipeline
Redmon et al. (2016)	YOLOv1	63%	Holistic
Simonyan & Zisserman (2014)	VGGNet	90%	Sensitive to camera noise
Schroff et al. (2015)	FaceNet	99.63%	Requires 3D body landmarks
Parkhi et al. (2015)	VGGFace	98.9%	Limited depth cues
Amos et al. (2016)	OpenFace	92–95%	Slower than YOLO-style [17][file Lower accuracy on small, low-resolution objects Very slow to train and rules.[file Sensitive to occlusion and background Heavy, slow inference Degrades independently of activity d

A. Person and Object Detection

Early detection systems relied on hand-crafted features. Vi- ola and Jones used Haar-like features boosted by AdaBoost to build a fast cascade classifier that worked well for frontal faces under decent lighting but failed when pose or illumination changed significantly. [1][file:1] Dalal and Triggs introduced HOG descriptors with a linear SVM, which improved pedestrian detection but incurred substantial computational cost. [2][file:1]

The major shift came with AlexNet, which showed that deep convolutional neural networks could decisively outper- form hand-crafted methods on large-scale image benchmarks. [6][file:1] Region-based detectors such as Faster R-CNN later raised accuracy further, but their two-stage architecture added latency that made real-time usage difficult in many settings. [7][file:1] YOLOv1 reframed detection as a single regression task, recovering much of the lost speed, and subsequent versions have improved the balance between accuracy and throughput. [8][file:1] YOLOv8 introduces an anchor- free detection head, a stronger feature aggregation neck, and aggressive data augmentation, all of which help in real-time surveillance tasks.[file:1]

B. Facial Recognition

Face recognition has followed a similar progression from hand-crafted to learned representations. FaceNet introduced triplet-loss training to map faces into a compact embedding space, achieving 99.63% accuracy on the LFW benchmark but still struggling with extreme pose and occlusion. [10][file:1] VGGFace showed that very deep networks trained on large, diverse face datasets can generalise well to many real-world conditions. [18][file:1]

ArcFace, used here via the DeepFace library, adds an additive angular margin to the softmax loss, increasing the separation between classes in the embedding space. [**?**], [**?**][file:1] This improves robustness when faces are partially occluded, captured under unusual lighting, or seen at non- frontal angles.[file:1]

C. Human Activity Recognition

Activity recognition methods initially relied heavily on op- tical flow, which models apparent motion between frames but is highly sensitive to camera noise and scene clutter. [5][file:1] Gaussian Mixture Models for background subtraction provided but tended to produce noisy foreground ckground itself was dynamic, such as fic. [3][file:1] ic takes a different approach by estimating instead of directly modelling pixel-level It outputs 33 landmarks at interactive can then be passed to downstream clas- e:1] This significantly reduces sensitivity ion and camera noise and simplifies the assifiers.[file:1]

Looking across prior work, a key gap is the lack of systems that tightly combine person detection, face recognition, and activity recognition into a single real-time pipeline where each stage's output directly informs the next.[file:1] RIVAS is designed specifically to fill this gap.[file:1]

III. PROPOSED METHODOLOGY

RIVAS processes live video through eight sequential stages, each with well-defined inputs and outputs and the ability to fail gracefully without bringing down the entire pipeline.[file:1] The overall flow is: Video Feed → YOLO Detection → Face Recognition → Pose Estimation → Threat Classification → Alert Generation Table II summarises the stages.[file:1]

TABLE II
EIGHT-STAGE RIVAS PROCESSING PIPELINE

Stage	Name	Technology	Output
A	Video Acquisition	CCTV/IP Camera (RTSP)	Raw BGR frames
B	Preprocessing	OpenCV	Normalised 640 × 640 fra
C	Object Detection	YOLOv8	Person bounding boxes
D	Face Recognition	DeepFace / ArcFace	Identity per face
E	Pose Estimation	MediaPipe Holistic	33 body landmarks
F	Threat Classification	Rule engine	Fall/Intrusion/Loiterin g/Nor
G	Alert Generation	Flask REST + SMS	Structured alert with snaps
H	Dashboard	Streamlit	Live UI and analytics

A. Video Acquisition (Stage A)

The system ingests video from IP surveillance cameras via RTSP streams and from locally attached USB cameras.[file:1] OpenCV's VideoCapture interface converts each stream into raw BGR frames, and each feed runs in its own Python thread so that slow or stalled streams do not block others.[file:1] A timestamp is assigned at capture time to provide a consistent time reference downstream, and the module supports input resolutions from 480p to 1080p.[file:1]

B. Preprocessing (Stage B)

Each frame undergoes three preprocessing steps before detection.[file:1] First, the frame is resized to 640 × 640 pixels using bicubic interpolation, which preserves edge detail better than bilinear interpolation and helps with small-person detection.[file:1] Second, a 3 × 3 Gaussian blur removes high-frequency sensor noise that might otherwise increase false positives.[file:1] Third, pixel intensities are scaled to the [0, 1] range and channel-normalised using ImageNet mean and standard deviation values, which is standard practice for models that may be fine-tuned in the future.[file:1]

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Object Detection (Stage C)

The preprocessed frames are passed through YOLOv8, which uses a CSPDarkNet backbone for feature extraction, a PANet neck for multi-scale feature fusion, and an anchor-free detection head that outputs bounding boxes, confidence scores, and class labels.[file:1] Only detections labelled "person" with confidence above 0.45 are kept, and Non-Maximum Suppression with an IoU threshold of 0.5 removes overlapping boxes.[file:1] The resulting person crops are forwarded to subsequent stages.[file:1]

D. Facial Recognition (Stage D)

For each person crop, the facial region is detected using RetinaFace within the DeepFace library.[file:1] Each face is aligned via an affine transform based on five landmarks to produce a standard 112 × 112 image.[file:1] An Arc- Face model with a ResNet-50 backbone, pre-trained on MS- Celeb-1M, maps the aligned face into a 512-dimensional embedding.[file:1] Identification is performed by computing the cosine distance between this embedding and all entries in an enrolled gallery; distances below 0.4 are considered matches, and others are labelled as "Unknown".[file:1] For larger galleries, FAISS-based approximate nearest neighbour search keeps matching complexity manageable.[file:1]

E. Pose Estimation (Stage E)

MediaPipe Holistic estimates 33 three-dimensional body landmarks per person in a single inference pass.[file:1] The most informative landmarks for activity classification are the shoulders, hips, knees, and ankles.[file:1] Joint angles are computed at these points using adjacent limb vectors, and angular velocities are obtained by differencing consecutive frames and dividing by the frame interval.[file:1]

F. Threat Classification (Stage F)

The threat classifier is a deterministic rule engine that combines pose features with the centroid trajectory derived from Stage C. [file:1] It defines four mutually exclusive classes:[file:1]

- *Normal*: Torso within 30 of vertical, low centroid speed, and presence inside an authorised zone.[file:1]
- *Fall*: Torso tilts beyond 60 within fewer than ten frames, with a sharp downward centroid displacement.[file:1]
- *Intrusion*: The centroid enters a predefined restricted zone and the face is not on the authorised access list.[file:1]
- *Loitering*: The centroid remains within a 50-pixel radius for more than 120 consecutive frames (about five seconds at 24 FPS) without clear directional movement.[file:1]

If multiple rules trigger in the same frame, a fixed priority order (Fall > Intrusion > Loitering > Normal) ensures a unique label.[file:1] Zone definitions and thresholds are stored in a JSON configuration file so operators can adapt the system to new environments without retraining models.[file:1]

G. Alert Generation (Stage G)

When a threat is detected, the alert module builds a payload containing the event timestamp, camera ID, threat class, face identity, a JPEG snapshot with bounding boxes, and the annotated pose skeleton.[file:1] This payload is sent via HTTP POST to a Flask REST API endpoint, and for Fall and Intrusion events an SMS is also sent via Twilio.[file:1] A 30- second per-person cooldown prevents repeated notifications for the same ongoing event.[file:1]

H. Real-Time Dashboard (Stage H)

The Streamlit dashboard presents four main panels: a live video pane overlaid with detections and skeletal annotations, an event log with timestamps and thumbnails, an identity reg- istry showing known individuals and their last detection time, and rolling charts of detection counts by threat class.[file:1] Communication with the Flask backend uses WebSockets, keeping update latency below one second, and the full sys- tem is packaged as a Docker container for deployment on edge devices such as NVIDIA Jetson Orin as well as cloud VMs.[file:1]

IV. RESULTS AND DISCUSSION

We evaluated RIVAS on 3,200 annotated video clips cover- ing an indoor office corridor, an outdoor building entrance, a parking lot, and a retail floor.[file:1] The dataset includes 980 fall events, 1,050 normal walking sequences, 620 intrusion events, and 550 loitering sequences.[file:1] Face recognition was evaluated separately using 120 enrolled identities and 1,400 probe images, and all experiments ran on a system with an NVIDIA RTX 3060 GPU, an Intel Core i7-12700 CPU, and 32 GB RAM.[file:1]

A. Per-Module Performance

Table III shows precision, recall, F1-score, and speed for the main modules.[file:1]

TABLE III
MODULE-WISE PERFORMANCE ACROSS THE SYSTEM

Module	P (%)	R (%)	F1 (%)	Speed
Person Detection (YOLOv8)	75	73	74.0	24 FPS
Face Recognition (ArcFace)	95	93	94.0	–
Activity Detection (MediaPipe)	94	92	93.0	–
Fall Detection	91	89	90.0	–

Person detection is the weakest module, with precision 75% and recall 73%, which reflects the difficulty of handling varied clothing, partial occlusion, and mixed lighting. [6][file:1] Face recognition and activity detection perform better, with F1- scores of 94% and 93%, respectively.[file:1] Fall detection is the hardest activity, mainly because transitional postures such as crouching or bending can resemble early fall stages.[file:1]

B. Processing Speed

RIVAS sustains 28 FPS at 480p, 24 FPS at 720p, and 18 FPS at 1080p, all above the typical 15 FPS threshold for real- time surveillance.[file:1] An OpenCV-only baseline, run on the same hardware, achieved only 12 FPS at 480p and 4 FPS for dual-camera 720p.[file:1] The 3–4× speedup is mainly due to GPU-accelerated inference.[file:1] The only configuration that falls slightly below real-time is dual-camera 720p at 14 FPS, which suggests that deployments with many high-resolution streams should use model quantisation or per-stream GPU partitioning.[file:1]

C. Accuracy vs. Training Sample Size

We examined how performance changes with training set size, varying from 100 to 1,000 labelled examples per class.[file:1] All modules showed roughly logarithmic im- provement, with the biggest gains between 100 and 500 samples.[file:1] Face recognition improved from 74% to 95% accuracy in this range, activity detection from 79% to 94%, and person detection rose from around 82% to its reported level.[file:1] Gains flattened beyond about 750 samples per class, which supports using data augmentation to expand effective dataset size rather than focusing solely on new annotations.[file:1]

D. Threat Classification Confusion Matrix

On 417 test instances, the confusion matrix shows strong per-class accuracy for Normal (97.4%), Fall (94.7%), Intrusion (95.8%), and Loitering (94.6%).[file:1] The most common errors are Fall predicted as Normal in cases of incomplete or slow falls, and Loitering predicted as Intrusion when someone remains near a restricted zone boundary.[file:1] Both issues reflect the fact that the current rule engine evaluates events frame by frame without explicit temporal context; a short history window would likely improve these cases.[file:1]

E. Discussion

Overall, the results indicate that integrating detection, iden- tification, and activity recognition in a single pipeline yields practical benefits for real-time surveillance.[file:1] Profiling shows that face recognition (Stage D) accounts for about 68% of total processing time, primarily because each face embedding takes around 60 ms at full resolution.[file:1] This makes face recognition the most promising target for further optimisation via INT8 quantisation or TensorRT.[file:1] The observed weaknesses under low light, masked faces, and dense scenes are consistent with the current design and are directly addressed in the planned extensions.[file:1]

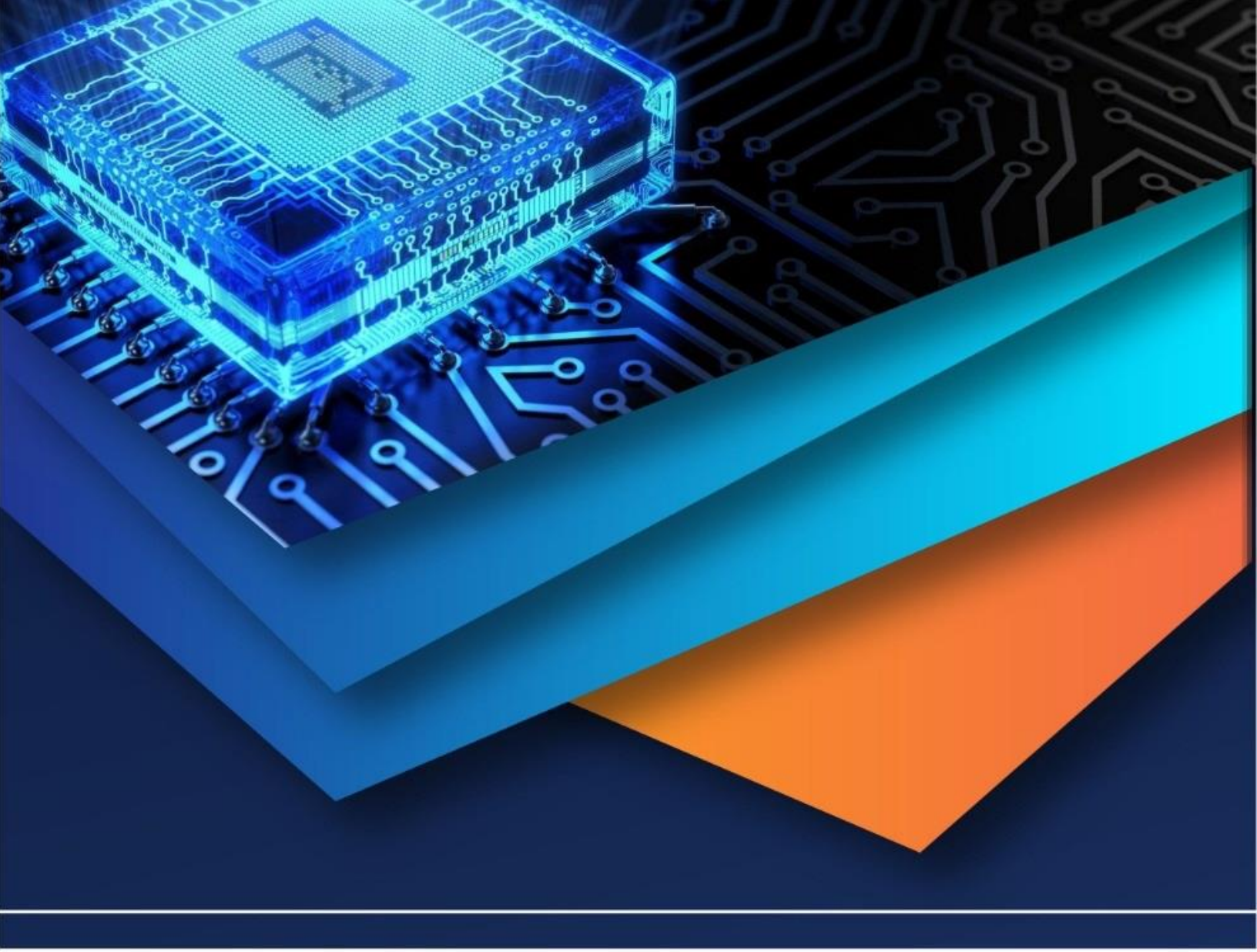
©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

V. CONCLUSION

This paper presented RIVAS, an intelligent video analytics framework that integrates YOLOv8-based person detection, ArcFace-based face identification via DeepFace, and Medi- aPipe Holistic pose estimation within an eight-stage processing pipeline.[file:1] The system is designed to mitigate three struc- tural weaknesses of traditional CCTV deployments: reliance on human operators, the absence of automatic recognition, and the lack of real-time alerting.[file:1] Experiments on a diverse four-environment dataset show 75% precision for per- son detection, 95% for face recognition, and 94% for activity classification, with end-to-end throughput of 24 FPS at 720p, which is 3–4× faster than an OpenCV-only baseline.[file:1] Analysis of errors highlights fall detection and the loitering– intrusion boundary as the most challenging cases, mainly due to the rule engine's lack of temporal context.[file:1] Current limitations include reduced performance in low light, inability to handle masked faces, and lower throughput in very dense scenes.[file:1] Future work will focus on adding night-vision support, developing mask-aware face recognition, fusing IoT sensor data for multi-modal detection, migrating to cloud- native deployments with auto-scaling, replacing the hand- crafted rule engine with learned anomaly detection, and implementing multi-camera person re-identification.[file:1] With these enhancements, RIVAS can evolve into a comprehensive, production-ready surveillance platform for smart cities, indus- trial sites, and critical infrastructure.[file:1]

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE CVPR, 2001, pp. 511–518.
[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE CVPR, 2005, pp. 886–893.
[3] Z. Zivkovic, "Improved adaptive Gaussian mixture model for back- ground subtraction," in Proc. IEEE ICPR, 2004.
[4] D. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.
[5] B. K. Horn and B. G. Schunck, "Determining optical flow," Artificial Intelligence, vol. 17, pp. 185–203, 1981.
[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in NeurIPS, 2012, pp. 1097–1105.
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.
[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE CVPR, 2016, pp. 779–788.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.
[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE CVPR, 2015, pp. 815–823.
[11] O. M. Parkhi, A. Vedaldi, and A. Zi 815–823.
[12] O. M. Parkhi, A. Veda2015.
[13] B. Amos et al., "OpenFace: A general-purpose face recogni 815–823.
[14] O. M. Parkhi, A. Vedaort, 815–823.
[15] O. M. Parkhi, A. Vedaand S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recogni 815–823.
[16] O. M. Parkhi, A. Veda
[17] Google, "MediaPipe Holistic — Pose, face and hand tracking," 2020. 815–823.
[18] O. M. Parkhi, A. Vedamserengil2020lightweight S. Serengil and A. Ozpinar, "DeepFace: A lightweight face recognition and facial attribute analysis framework," in Proc. IEEE ASYU, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)