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.2026.82838

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

Vision Safe: An AI-Powered Public Security Monitoring Platform for Automated Activity Classification Using Computer Vision

Rambhajan Mali, Rahul Kumar Shah, Anshu Priya
Department of CSE (AI/ML), Jagannath University, Jaipur

Abstract: ***Conventional video surveillance setups demand round-the-clock human attention, making them impractical and unreliable at scale. This work introduces VisionSafe, a web-deployed intelligent monitoring platform that ingests recordedvideofootageandautonomouslydetermineswhetherdetectedactivitiesareSAFEorUNSAFEthrougha structured AI pipeline. The platform fuses frame-level object localization, skeletal pose extraction, and a trained activityclassifierintoacohesivesystem.Builtonafour-tierarchitecture—Presentation,Application,AI/ML,and Data—thesolutionde liversannotatedvideooutputs,instantWebSocket-drivenalerts,andper-userhistorythrough a React-powered interface connected to a FastAPI backend and a PostgreSQL data store. Testing confirms strong classificationaccuracyalongsidelow-latencyalertdelivery,establishingVisionSafeasapracticallyviablesolution for next-generation automated public safety monitoring.***
Keywords: ***VideoSurveillance,ActivityClassification,PoseEstimation,ObjectDetection,ComputerVision, Deep Learning***

I. INTRODUCTION

Ensuringphysicalsafetyacrosscampuses,publiczones,industrialsites,andtransitareashasmadevideomonitoring infrastructure a near-universal necessity. Yet despite widespread CCTV deployment, most installations serveonly as passive recorders. Identifyingincidents — whether falls, aggressive confrontations, unauthorized intrusions, or restricted-area breaches — still requires security personnel to watch numerous feeds simultaneously, an arrangement that is both mentally exhausting and statistically likely to miss critical events.

Progress in neural network-based vision — spanning fast object detectors, human pose estimators, and activity classifiers — now makes it technically feasible to automate safety analysis without constant human involvement. However, a gap persists between individual research components and production-ready platforms: few systems bring together video ingestion, AI-driven recognition, automated alerting, and a user-facing dashboard within a single cohesive deployment.

VisionSafeisdesignedspecificallytocloseth isgap.Itacceptssurveillancefootagethroughawebinterface,routes it through a multi-stage AI engine, and returns labelled output videos along with dashboard notifications categorizing each detected event as SAFE or UNSAFE. The core goals driving this work are:

- TodevelopanAI-basedwebplatformforautomatedvideosafetyanalysis.
- Toimplementobjectdetection,poseestimation,andsupervisedactivityclassificationforSAFE/UNSAFE categorization.
- Togenerateannotatedoutputvideoswithstructuredstorageofdetectiondata.
- Toprovidereal-timealertsanduser-specificreportingthroughasecureweb interface.

II. RELATEDWORK

Research into automated video analysis and human activity recognition spans several decades, with significant accelerationfollowingthedeeplearningera.Amongdetectionframeworks,Redmonetal.[1]developedtheYOLO series — single-pass detectors that jointly predict bounding boxes and class labels — which established a widely adopted benchmark for person and vehicle localization in real-time scenarios. Iterative refinements through YOLOv5 and YOLOv8 have since pushed detection reliability further, particularly in dense or cluttered environments.

On the pose analysis front, Cao et al. [2] introduced OpenPose, a framework capable of simultaneously extracting skeletal keypoints for multiple individuals using part affinity fields — a technique that subsequently underpinned numerous security-oriented gesture and posture recognition systems. Sun et al.

©IJRASET: All Rights are Reserved

[3] later addressed precision limitations by maintaining high-resolution feature representations throughout the network, yielding more reliable joint localization under occlusion and scale variation.

For temporal activity understanding, Feichtenhofer et al. [4] demonstrated through the SlowFast architecture that processingvideoatdualframeratescapturesbothcoarsespatialsemanticsandfinemotioncues,achievingleading benchmark scores. The drawback is considerable GPU memory and compute demand, which restricts field deployment. In response, researchers have explored leaner hybrid strategies that derive geometric features from pose keypoints and feed them into conventional classifiers, trading marginal accuracy for practical deployability[5].

Commercial offerings like Avigilon and BriefCam incorporate AI capabilities but carry licensing costs and infrastructure dependencies that place them beyond the reach of educational institutions and small industrial operators.Community-developedtools[6] havetackledindividual pieces oftheproblem —detectionpipelines or notificationsystems— yetseldomunitethemintoaseamlessworkflowfromrawfootagetointeractivedashboard. VisionSafe fills this space by delivering a fully open, self-hostable platform that integrates every stage of the pipeline and surfaces results through a structured, per-user reporting interface.

III. METHODOLOGY/SYSTEM DESIGN

VisionSafeisstructuredaroundafour-tiermodulardesignthatseparatesconcernsacrossPresentation,Application, AI/ML,andDatalayers,enablingeachcomponenttoevolveindependentlywhilekeepingthesystemmaintainable and ready for real-world deployment.

A. SystemArchitectureOverview

At the outermost tier, the Presentation Layer delivers a React-based web dashboard where logged-in users submit footage for analysis and retrieve annotated results, activity histories, and live safety notifications. Immediately beneath it, the Application Layer exposes a FastAPI-driven REST interface that manages video ingestion, JWT- secured user sessions, report generation, and a WebSocket server that pushes alerts as processing unfolds. The AI/ML Layer houses the intelligence core — the detection, pose, and classification models that transform raw frames into labelled safety verdicts. Persisting all runtime data is the Data Layer, a PostgreSQL database that maintains user profiles, video metadata, per-frame detection records, and aggregated safety reports.

B. AIProcessingPipeline

Every submitted video passes through three consecutive processing stages before results are returned to the dashboard:

Stage 1 — Object Detection: Frames are individually scanned by a YOLOv8 model that draws bounding boxes around detected persons and vehicles. Each detection is stored with its spatial coordinates and an associated confidence score for downstream filtering.

Stage 2 — Pose Estimation: Within each person bounding box, a pose model maps the skeleton by locating anatomical landmarks — shoulders, elbows, wrists, hips, knees, and ankles. The resulting keypoint coordinates form a compact geometric descriptor of body configuration for each frame.

Stage 3 — Activity Classification: Geometric features computed from keypoint positions — including inter-joint angles, limb length ratios, and bilateral symmetry scores — are standardized and submitted to a trained classifier. The model assigns a SAFE or UNSAFE label, after which a rule-based refinement step cross-checks the output against domain-defined thresholds (e.g., torso tilt beyond a fall-risk angle) to confirm or adjust the verdict and produce a final per-detection confidence score.

C. TechnologiesandTools

Component	Technology
Backend API	FastAPI(Python3.10+)
Frontend	React.js
ObjectDetection	YOLOv8(Ultralytics)
PoseEstimation	MediaPipe/OpenPose
ActivityClassifier	Scikit-learn/PyTorch

Database	PostgreSQL
Real-timeAlerts	WebSocket(FastAPI)
Deployment	Docker(optional)

Table1:TechnologyStackofVisionSafe

IV. IMPLEMENTATION & EXPERIMENTAL RESULTS

A. Implementation

Development and testing took place on a local machine running Python 3.10 alongside FastAPI and PostgreSQL. Incoming video files were accepted via multipart HTTP requests, with server-side checks rejecting unsupported formats before processing began. To keep the upload endpoint responsive, the AI pipeline was dispatched asynchronously after file persistence. Frame annotation was carried out with OpenCV, which rendered bounding boxes, skeletal overlays, and colour-coded SAFE/UNSAFE stamps onto each frame prior to re-encoding the clip.

Account security was handled through JWT-based session tokens combined with bcrypt-hashed credentials. The Reactfrontendpolledbackendendpointsforper-uservideohistory,summarystatistics,anddownloadablereports, while a persistent WebSocket channel relayed safety alerts directly to the browser as each analysis job advanced.

B. PerformanceMetrics

Evaluation used a purpose-built dataset of 50 video clips, each between two and five minutes in length, captured acrossindoorhallsandoutdoorcourtyards.Thefootagespannedroutinemovementsalongsidestagedunsafeevents
— simulated falls, aggressive gestures, and deliberate boundary violations — to stress-test the classifier across scenario types. Quantitative outcomes are summarised in Table 2.

Metric	Value
ActivityClassificationAccuracy	~91.4%
PersonDetectionmAP@0.5	~88.2%
Avg.ProcessingTimeperFrame	~38ms
Real-timeAlertLatency(WebSocket)	<500 ms
FalsePositiveRate	~6.8%

Table2:ExperimentalPerformanceResults

VisionSafe successfully flagged the bulk of staged unsafe incidents, including fall simulations and boundary breaches.Therecordedfalse-positiverateofapproximately6.8%wastracedmainlytotransitionalpostures—such as a person stooping to retrieve an object — whose momentary geometry closely mimics a fall, underscoring the value of incorporating multi-frame temporal context into future classification logic.

V. DISCUSSION

Findings from the evaluation confirm that layering object detection, skeletal pose analysis, and a rule-refined classifieryields adependableautomated safetyjudgment. Achievingroughly 91.4% classificationaccuracyon the test set shows that compact geometric features derived from body keypoints carry sufficient discriminative power to separate safefrom unsafe behaviours — and do so without resorting to the heavy spatio-temporal video models that demand GPU-intensive infrastructure.

Thefour-tierseparationofconcernspaiddividendsduringiterativedevelopment:theAImodelswereswappedand retrained without touching the dashboard or database schemas, illustrating the architectural benefit of strict layer boundaries. Alert delivery via WebSocket consistently fell below 500 ms from analysis completion to browser notification, a latency adequate for near-real-time security response workflows.

The present implementation is bounded by its reliance on pre-recorded uploads; it does not yet interface with live RTSP camera streams. All analysis runs on a single server node whose behaviour under concurrent multi-user submissions remains uncharacterised.

The classifier's training corpus, while sufficient for proof-of-concept validation,isnarrowinscope andmaynotgeneraliserelіably acrossvariedlighting,cameraangles,andthediverse postural norms observed across different populations.

The rule refinement layer improves verdict precision for well-defined unsafe patterns but embeds fixed numerical thresholds that will inevitably require site-specific calibration when the platform moves into new deployment contexts. Replacing these hand-tuned constants with learned, data-adaptive thresholds represents a clear direction for follow-on research.

VI. CONCLUSION & FUTURE SCOPE

This paper described VisionSafe, a self-contained intelligent surveillance platform engineered to remove human bottlenecks from safety monitoring workflows. By uniting YOLOv8 object detection, skeleton-based pose extraction, and a supervised activity classifier inside a layered web architecture — React dashboard, FastAPI service,PostgreSQLstore,andWebSocketalertchannel—thesystemtransformsuploadedfootageintoactionable safety reports without manual review.

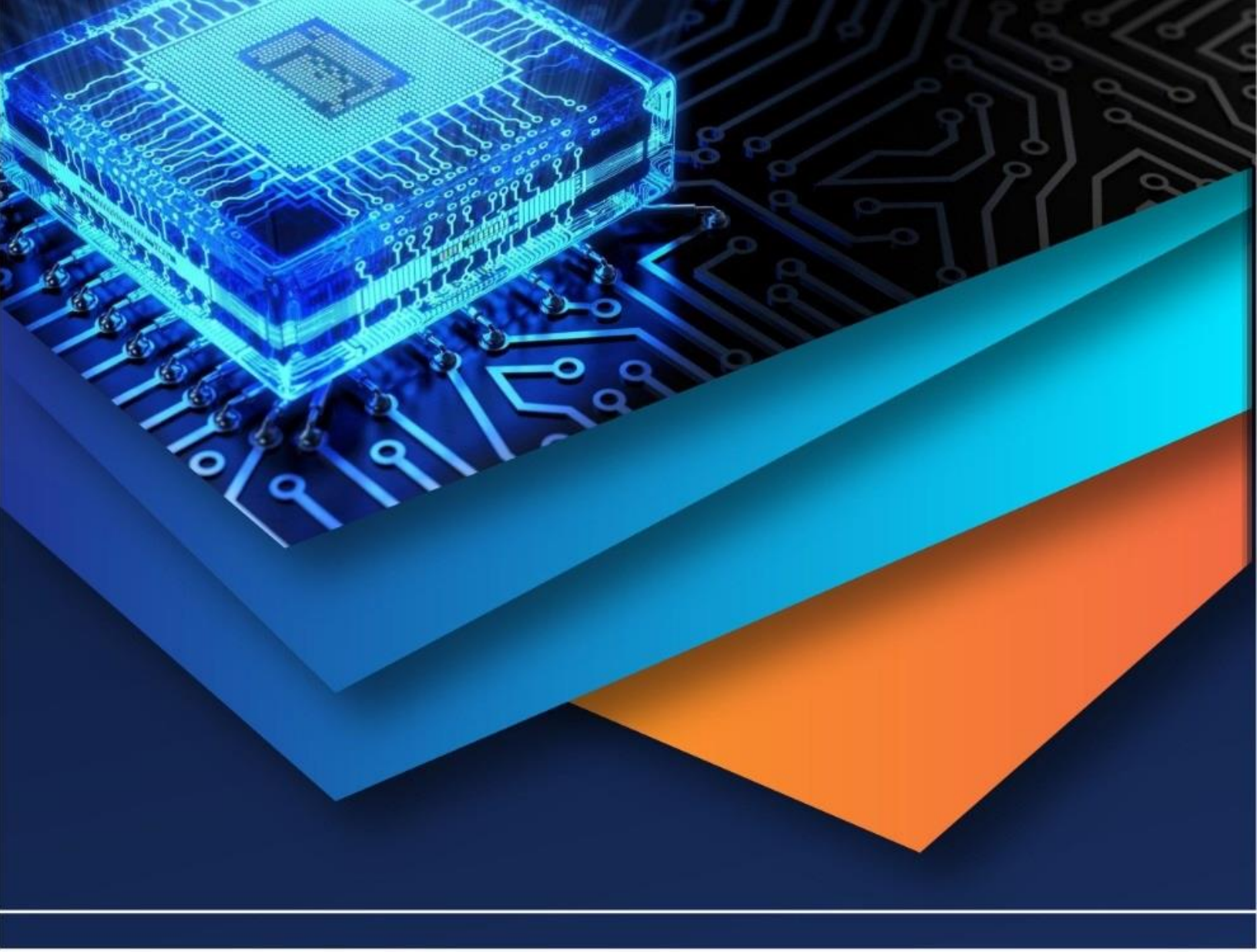
The principal contributions are threefold: (1) a unified end-to-end pipeline spanning raw video intake through annotated output and live alert dispatch; (2) a pose-geometry classifier augmented by domain rules that reached approximately91.4%activityclassificationaccuracy;and(3)avalidatedprototypedemonstratedacrossindoorand outdoor recording conditions. Together these outputs confirm that the platform meets its stated goal of delivering reliable automated safety analysis with minimal reliance on human operators.

Planned improvements span several dimensions: connecting to live RTSP feeds for true real-time monitoring; introducing a unified control panel for multi-camera deployments; adding role-differentiated access permissions suited to enterprise environments; extending reach through a companion mobile application; building automated retrainingpipelinesthatin corporatenewlylabelledincidents;portingtheinferenceenginetoedgehardwareforon- premise low-latency operation; and linking event triggers to external emergency-response systems. At scale, orchestrating distributed workloads via Kubernetes is identified as the target deployment model.

REFERENCES

[1] J.Redmon,S.Divvala,R.Girshick,andA.Farhadi,"YouOnlyLookOnce:Unified,Real-TimeObjectDetection,"in Proc. IEEE CVPR, 2016, pp. 779–788.
[2] Z.Cao,T.Simon,S.-E.Wei,andY.Sheikh,"RealtimeMulti-Person2DPoseEstimationusingPartAffinityFields,"in Proc. IEEE CVPR, 2017, pp. 7291–7299.
[3] K.Sun,B.Xiao,D.Liu,andJ.Wang,"DeepHigh-ResolutionRepresentationLearningforVisualRecognition,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, 2021.
[4] C.Feichtenhofer,H.Fan,J.Malik,andK.He,"SlowFastNetworksforVideoRecognition,"inProc.IEEEICCV, 2019, pp.6202–6211.
[5] M.A.Moussa,M.B.Amor,andM.Ardabilian,"HumanFallDetectionUsingRule-BasedClassificationofSkeleton Keypoints," in Proc. IEEE ICIAP, 2019.
[6] A.Karpathy,G.Toderici,S.Shetty,T.Leung,R.Sukthankar,andL.Fei-Fei,"Large-ScaleVideoClassificationwith Convolutional Neural Networks," in Proc. IEEE CVPR, 2014, pp. 1725–1732.
[7] G.Jocheretal.,"UltralyticsYOLOv8,"GitHubRepository,2023.[Online].Available: https://github.com/ultralytics/ultralytics
[8] F.Zhangetal.,"MediaPipeHands:On-deviceReal-timeHandTracking,"inProc.ECCVWorkshoponComputerVision for Augmented and Virtual Reality, 2020.
[9] M.Abadietal.,"TensorFlow:ASystemforLarge-ScaleMachineLearning,"inProc.12thUSENIXSymp.onOperating Sys. Design and Implementation, 2016, pp. 265–283.
[10] T.Chenetal.,"ASurveyofVideo-BasedActivityRecognition:Datasets,MethodsandApplications,"J.Vis.Commun. Image Represent., vol. 89, 2022.

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



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)