EchoLens: Empowering Video Intelligence Through AI-Driven Surveillance

A. Dawood
Faculty of Computer Science and Engineering
Alamein International University
Alamein, Egypt

A. Ginidy
Faculty of Computer Science and Engineering
Alamein International University
Alamein, Egypt

M. Elslmawy
Faculty of Computer Science and Engineering
Alamein International University
Alamein, Egypt

G. Nashaat
Faculty of Computer Science and Engineering
Alamein International University
Alamein, Egypt

M.A. Abdou
Informatics Research Institute
City for Scientific Research and
Technological Applications
(SRTA-City)
New Borg El-Arab, Egypt

A. Khaled
Faculty of Computer Science and Engineering
Alamein International University
Alamein, Egypt

***Abstract*—In the realm of AI-driven surveillance, automatic video summarization facilitates rapid threat detection, anomaly identification, and operational efficiency, revolutionizing security systems. EchoLens is an innovative AI-driven video surveillance system designed to automate the detection, classification, and summarization of events from video footage. Addressing the inefficiencies of traditional manual monitoring, EchoLens leverages advanced deep learning models, including YOLOv12 for precise object detection, a fine-tuned I3D model for accurate event classification, and Google's Gemini API for generating concise textual narratives. The system supports both pre-recorded and live video feeds, delivering real-time or near-real-time analysis through an intuitive Flask-based web interface. Key features include multi-event detection (e.g., arrest, explosion, fight), confidence scoring, keyframe extraction, and comprehensive PDF reports. By seamlessly integrating computer vision and natural language processing, EchoLens enhances security, operational efficiency, and situational awareness, providing a scalable, innovative solution for modern surveillance needs.**

***Index Terms*—AI Surveillance, Event Detection, Video Summarization, Action Recognition, Deep Learning, YOLOv12, I3D, Natural Language Generation.**

I. Introduction

Traditional video surveillance systems rely heavily on manual monitoring, where human operators review footage to identify critical events, a process that is both time-consuming and prone to errors due to fatigue and limited attention spans [1]. These classical methods often employ basic motion detection algorithms or simple rule-based systems that trigger alerts based on predefined thresholds, such as pixel changes or object size, lacking the ability to interpret complex scenarios [2]. Such approaches are inefficient for managing the vast amounts of data generated by modern CCTV systems, particularly in high-stakes environments like airports or public spaces, where delayed or missed detections can have significant consequences [3].

In the realm of AI-driven surveillance, EchoLens revolutionizes video processing by automating event detection, classification, and summarization, facilitating rapid threat detection, anomaly identification, and operational efficiency [4]. Leveraging advanced deep learning models, EchoLens employs YOLOv12 [5] for precise object detection, a fine-tuned I3D model [6] for accurate event classification across categories like arrest, explosion, and fight, and Google's Gemini API for generating human-readable textual narratives, as detailed in our system architecture (Fig. 1). By integrating computer vision and natural language processing, the system supports both pre-recorded and live video feeds, delivering real-time or near-real-time analysis through a user-friendly Flask-based web interface, complete with multi-event detection, confidence scoring, keyframe extraction, and comprehensive PDF reports, thus providing a scalable solution for modern surveillance needs [7]. The system processes a diverse data set, as shown in TABLE I, to ensure robust performance across various scenarios [8].

A. Objectives

The proliferation of CCTV cameras generates enormous data, often reviewed only post-incident due to human limitations. EchoLens aims to transform surveillance by automating event detection, reducing response times, and providing contextual narratives, thereby enhancing safety and resource efficiency across diverse environments such as airports, streets, and malls. The primary objectives of EchoLens are:

- Develop an AI-driven system for real-time detection and classification of security-relevant events.

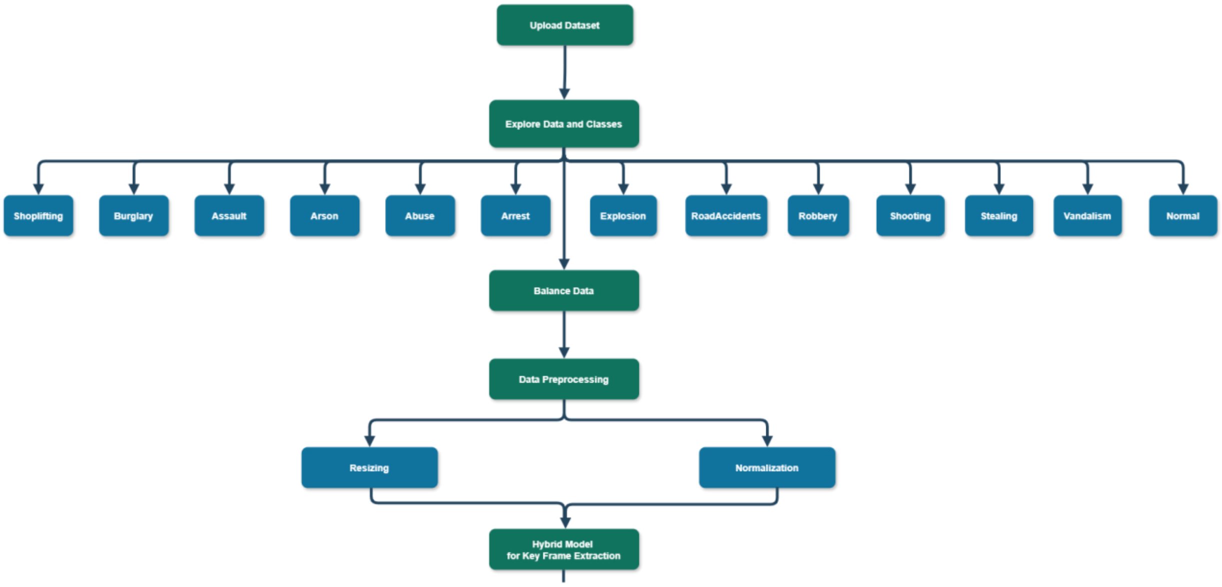

Fig. 1. EchoLens System Architecture.

- Generate human-readable narratives from video data for enhanced situational awareness.
- Provide a user-friendly interface for non-technical users, ensuring accessibility and usability.
- Ensure ethical considerations, including privacy and transparency, are integrated into the system design.

II. Related Work

Automated video analysis has seen remarkable progress with the advent of deep learning, enabling sophisticated surveillance systems like EchoLens. The following subsections explore key methodologies in object detection and tracking, action recognition, and video captioning and summarization, providing a comprehensive foundation for EchoLens's approach.

A. Object Detection and Tracking

Object detection forms the backbone of modern video surveillance, with models like YOLO [9], Faster R-CNN [10], and SSD [11] widely adopted for identifying objects in video frames. YOLO's single-stage architecture prioritizes speed, making it ideal for real-time applications, while Faster R-CNN offers high accuracy through region proposal networks, albeit at a computational cost. Recent advancements, such as YOLOv10 [9] and its successor YOLOv12 [5], enhance detection precision and efficiency. Object tracking complements detection by maintaining object identities across frames. Algorithms like SORT [12] and Deep SORT [13] are popular, and advanced variants like Bot-SORT [14] leverage motion and appearance cues to handle occlusions effectively. EchoLens employs YOLOv12 [5] and Bot-SORT [14] for robust detection and tracking.

B. Action Recognition

Action recognition has transitioned from traditional handcrafted features to deep learning-based approaches. Models like I3D [6] introduced 3D Convolutional Neural Networks to model temporal relationships effectively. I3D, pre-trained on large-scale datasets like Kinetics [15], excels at classifying complex actions. Alternative architectures, such as Temporal Segment Networks (TSN) [16] and SlowFast Networks [17], offer complementary approaches. Fine-tuning on datasets like UCF101 [18] and HMDB51 [19] enhances model performance for specific action categories. EchoLens utilizes a fine-tuned I3D [6] model, trained on UCF101 [18], to classify security-relevant events.

C. Video Captioning and Summarization

Video captioning and summarization bridge visual and textual domains. Early work, such as VideoBERT [20], combined visual features from CNNs with language models. Recent advancements, including models like CLIP-VIT [21] and Transformer-based architectures [22], leverage large-scale vision-language pre-training. Google's Gemini API, utilized by EchoLens, builds on these advancements to generate detailed frame descriptions and concise event summaries, enhancing situational awareness and cross-validating classifications to correct potential errors.

III. Methodology

EchoLens integrates advanced computer vision and natural language processing to deliver automated, real-time video surveillance with actionable insights. By combining state-of-the-art deep learning models and a user-centric interface, the system transforms raw video data into meaningful event detections and narratives, enhancing security and operational efficiency, as illustrated in Fig. 1.

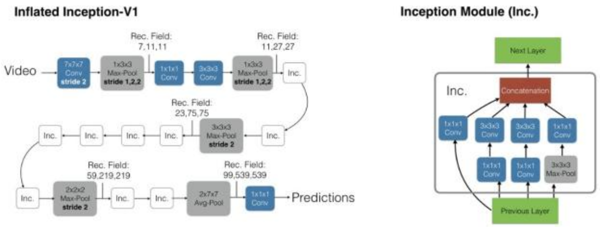

Fig. 2. I3D model Architecture.

A. Data Preprocessing

The system processes a diverse dataset of 1,900 video clips from Kaggle's "Real-Time Anomaly Detection in CCTV Surveillance" [23], consolidated into eight event categories. Frames are resized to 224x224 pixels and normalized. Keyframe extraction identifies significant frames based on motion and object activity.

B. Event Detection and Classification

The pipeline involves:

1) *Keyframe Extraction*: YOLOv12 [5] detects objects, while BoT-SORT [14] tracks them. Motion detection, color change analysis, and optical flow identify significant frames.
2) *Event Classification*: A fine-tuned I3D model (Fig. 2), trained on UCF101 [18], classifies keyframes into eight categories with confidence scores.
3) *Event Logging*: Detected events are logged with precise start and end times.

C. Narrative Generation

Google's Gemini API is used to generate human-readable narratives. Keyframes are analyzed to create concise frame descriptions and a comprehensive event summary. The API also cross-validates initial classifications, enhancing reliability.

D. User Interface

A proposed web interface (Fig. 3) offers seamless interaction, supporting video uploads, live stream analysis, and downloadable PDF reports. Accessibility features, such as multi-language support and dark mode, ensure broad usability.

TABLE I
VIDEO DATASET STATISTICS

Category	Videos	Selected	Frames	Keyframes
Normal	900	100	50,000	7,995
Fight	150	100	180,000	7,992
Explosion	100	100	520,000	10,035
Stealing	400	100	100,000	7,983
Road Accident	150	100	100,000	7,915
Vandalism	50	50	150,000	7,992
Arrest	100	50	300,000	8,288
Shooting	50	50	50,000	7,934

TABLE II
KEYFRAME EXTRACTION COMPARISON

Video Name	Total Frames	Proposed Keyframes
RoadAccidents002x264.mp4	347	9
Fighting005x264.mp4	1784	32
Stealing015x264.mp4	1800	52

IV. RESULTS AND DISCUSSION

In evaluating EchoLens's performance, we adopted a multi-faceted strategy focusing on key metrics such as keyframe efficiency, classification accuracy, and narrative quality. Experiments were conducted using Python-based frameworks like PyTorch, OpenCV, and Flask, deployed on a server with an NVIDIA RTX 3090 GPU.

A. Performance Analysis

Keyframe Efficiency: The proposed keyframe extraction method significantly reduces the number of frames from the raw video (e.g., from 1784 to 32 for a fight video), as shown in TABLE II.

Classification Accuracy: The I3D model, fine-tuned on the UCF101 dataset, demonstrates a clear learning trajectory. As seen in Fig. 4, training loss steadily decreases while training accuracy improves sharply. However, validation accuracy begins to plateau and diverge from training accuracy after the seventh epoch, suggesting overfitting. To combat this, we employed early stopping at epoch 7, integrated dropout layers, and applied data augmentation techniques. These interventions successfully reduced overfitting and enhanced the model's generalization ability.

Narrative Quality: Gemini-generated summaries provide clear, actionable insights, improving situational awareness.

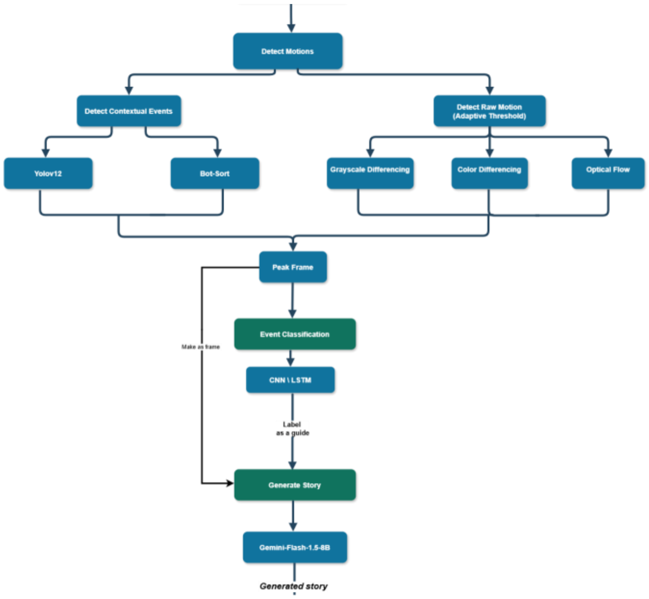

Fig. 3. EchoLens User Interface.

TABLE III compares the generated story with the ground truth from the UCA (UCF Crime Annotation) Dataset [24].

B. Ethical Considerations

EchoLens is designed with ethical considerations at its core to ensure responsible deployment. Key measures include:

- *Privacy*: All processed video data is automatically cleaned from the system after analysis and report generation to prevent unauthorized access or misuse of personal information.
- *Transparency*: The system provides clear confidence scores for each detected event, allowing operators to understand the model's certainty. The generated narratives further offer a transparent, human-readable log of events.
- *Fairness*: By leveraging narrative-based cross-validation with the Gemini API, the system can identify and flag potential biases or misclassifications from the vision model, promoting a fairer and more accurate interpretation of events.

V. Conclusion & Future Work

EchoLens represents a significant advancement in AI-driven surveillance by integrating YOLOv12 [5], I3D [6], and the Gemini API to deliver real-time event detection and narrative generation. The system enhances operational efficiency, reduces human workload, and provides actionable insights, with a user-friendly interface that ensures accessibility. The proposed model achieved a training accuracy of approximately 69% and a validation accuracy of around 55% before early stopping was applied to prevent overfitting.

Future enhancements will focus on several key areas to expand the system's capabilities. We plan to introduce synchronized multi-camera analysis for broader situational awareness and support for user-defined anomaly types to increase flexibility. Additionally, we will integrate real-time alerts via SMS or email for critical events and develop edge-optimized models to reduce latency for on-site deployments.

Acknowledgment

We acknowledge the use of Google's Gemini API for generating human-readable narratives in the EchoLens system. We also thank the Faculty of Computer Science and Engineering at Alamein International University for providing the computational resources necessary for this research.

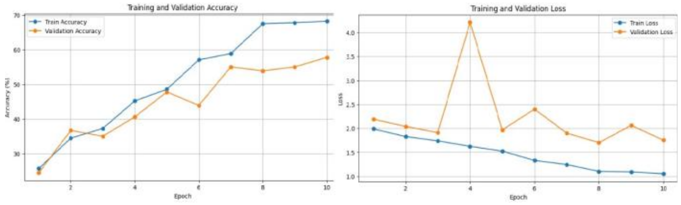

Fig. 4. Training and Validation Metrics.

References

[1] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering psychology and human performance*, 4th ed. Routledge, 2015.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the 1999 IEEE computer society conference on computer vision and pattern recognition*, vol. 2, 1999, pp. 246–252.

[3] B. Elias, *Airport and aviation security: US policy and strategy in the age of global terrorism*. CRC press, 2010.

[4] T. Bouwmans, S. Javed, W. Sultani, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Networks*, vol. 117, pp. 8–66, 2019.

[5] Y. Tian and Q. Ye, "YOLOv12: Attention-Centric Real-Time Object Detectors," 2025.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[7] Y. Yang, H. Chen, and Z. Zhang, "Multimodal video analysis with vision-language models," 2021.

[8] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[9] e. a. Wang, Ao, "YOLOv10: Real-Time End-to-End Object Detection," in *Advances in Neural Information Processing Systems*, 2024.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[12] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[13] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[14] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," 2022.

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, , and A. Zisserman, "The kinetics human action video dataset," 2017.

[16] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[17] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[18] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in The Wild," 2012.

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[20] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] Kaggle, "Real-Time Anomaly Detection in CCTV Surveillance," https://www.kaggle.com/datasets/webadvisor/real-time-anomaly-detection-in-cctv-surveillance, 2023, accessed: 2024-09-10.

[24] Vigneshwar, "UCA (UCF Crime Annotation) Dataset," https://www.kaggle.com/datasets/vigneshwar472/ucaucf-crime-annotation-dataset, 2023, accessed: 2024-09-10.

TABLE III
COMPARISON OF GROUND TRUTH AND GENERATED STORY

Video Name	Ground Truth	Story Generated
RoadAccidents002_x264	A white bus hit the sidewalk, many cars are driving on the road, many people walking on the sidewalk.	A series of surveillance video frames depict a traffic incident in a busy city street where multiple vehicles (a car and a bus) struck pedestrians, resulting in injuries and causing people to scatter.
Fighting005_x264	Two men in black clothes and masks ran from the door to the right side of the screen, Two men dragged a box out the door, A bald man on the left ran out to stop him.	The surveillance footage depicts a robbery in progress at a store. Multiple scenarios are presented, showing perpetrators robbing the store, sometimes assaulting or threatening an employee while stealing items.
Arrest015_x264	A police car chased a white truck, The white truck hit a police car. A policeman got out of a hit police car... The white truck fled, and two policemen chased it.	A series of surveillance camera images depict a traffic incident, causing a significant traffic jam. A police car subsequently stops a white van in a nearby roundabout, suggesting an apprehension related to the incident.