Mrs. Vibhavari Jawale 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Live Surveillance with Actionable Intelligence: A Review

[1]Mrs. Vibhavari Jawale, [2] Mrs. Deepali Hajare, [3] Arhant Sahuji, [4] Tanay Shinde, [5]Ananya Vaishnav, [6]Ritesh Kadam

Abstract- The rapid advancement of computer vision and natural language processing has paved the way for new forms of intelligent video surveillance. Traditional closed-circuit television (CCTV) and motion-based monitoring systems are limited in their ability to understand contextual information, often resulting in false alarms and requiring extensive human intervention. To address this gap, recent research explores the integration of vision-language models (VLMs) and sentiment analysis for context-aware surveillance. This review focuses on emerging methodologies where image captioning models such as Salesforce BLIP are used to describe real-time video frames in natural language, followed by sentiment-driven analysis to assess the nature of the detected activity. The combination of visual understanding, language-based context generation, and sentiment inference enables systems to differentiate between benign and suspicious behavior, thereby reducing false positives and providing actionable insights. Key applications include public safety in smart cities, security in high-risk environments like airports and banks, and monitoring sensitive areas such as hospitals and military zones. The core contribution of this review is the evaluation of how VLM-based context awareness augments conventional object detection pipelines, shifting surveillance toward more explainable and human-like alerting mechanisms. Furthermore, we discuss computational challenges, accuracy limitations, and privacy concerns while highlighting the societal implications of deploying such systems, including alignment with Sustainable Development Goals (SDGs) such as fostering safe cities and reducing crime. Future directions include multimodal fusion, real-time optimization, and ethical frameworks for responsible deployment.

Keywords- Vision-Language Model, Intelligent Surveillance, Anomaly Detection, Object Tracking, Real-Time

I. INTRODUCTION

Surveillance technologies have evolved as a cornerstone of modern security practices, finding application across diverse domains such as public safety, healthcare, defence, and critical infrastructure. Traditional systems, however, remain largely limited in functionality—primarily focused on motion detection, video feed recording, or simple anomaly detection using rule-based algorithms. These methods, while functional, fail to provide a deeper semantic understanding of activities captured by cameras. As a result, they place a heavy burden on human operators to interpret events, often leading to delayed response times, frequent false alarms, and a lack of contextual explanations for alerts. With increasing risks related to urban security, terrorism, and workplace safety, the need for context-aware, intelligent surveillance systems is greater than ever. Recent advances in artificial intelligence
(AI), particularly Vision-Language Models (VLMs), have brought substantial improvements in bridging the gap between visual input and human-like interpretation. Models such as Salesforce BLIP (Bootstrapping Language-Image Pretraining) can generate meaningful captions from still images, while transformer-based natural language

processing (NLP) tools, including sentiment analysis, allow textual outputs to be evaluated for intent and threat severity. This convergence of computer vision and natural language understanding represents a major shift in the way surveillance systems can be built transitioning from passive monitoring systems into proactive, context-aware agents capable of describing activities and determining situational risks in real time. The purpose of this review is to synthesize the current state of research at the intersection of computer vision, captioning, and sentiment analysis, and to highlight the potential impact of integrating these methods into intelligent surveillance systems. Specifically, the review seeks:

- To highlight advances that enable human-like interpretability of surveillance feeds.
- To identify key methodologies and architectures, such as BLIP for captioning and transformer-based sentiment models for evaluating textual descriptions.
- To discuss real-world applications, including airports, hospitals, banking, defence, and smart cities, where safety and efficiency demand intelligent monitoring.
- To assess challenges and limitations, such as computational requirements for real-time processing, risks of misclassification in complex scenes, and privacy considerations in large-scale deployments.
- To outline future research directions where emerging trends like multimodal learning, edge computing, and ethical AI design will enhance system performance and societal usefulness.

The contributions of this work are threefold. First, it emphasizes the importance of moving beyond simple motion detection toward fully interpretable monitoring systems. Second, it outlines a novel context-aware methodology based on the integration of vision-language modelling and sentiment-based interpretation for intelligent alerting. Finally, it situates this technological direction within the broader sustainability and social impact framework, particularly aligning with United Nations Sustainable Development Goals (SDGs), such as SDG 11 (Sustainable Cities and Communities) and SDG 16 (Peace, Justice, and Strong Institutions), demonstrating how intelligent surveillance can contribute to safer and more resilient societies.

In terms of organization, this paper is structured as follows. Section 1 introduces the domain and motivation for context-aware surveillance. Section 2 provides a comprehensive review of existing literature, including recent advancements in vision-language models, video analytics, and NLP-based sentiment analysis. Section 3 explores the proposed integrative framework that combines video captioning with sentiment classification for descriptive, real-time system alerts. Section 4 discusses key applications across industries, while Section 5 highlights open challenges, ethical considerations, and areas for future research. Finally, Section 6 concludes the review with reflections on the social impact, policy implications, and technological pathways toward sustainable urban safety. By presenting a consolidated view of advances and challenges in context-aware surveillance, this review aims to provide both academic researchers and industry practitioners with insights into the capabilities, limitations, and promising directions of these emerging systems.

II. LITERATURE REVIEW

Table -1: Literature Review

Reference No.	Title of Paper	Key Features/ Key Findings	Models / Algorithms Used	Evaluation Parameters Used	Research Gaps / Limitations
Array 27 100471	Vision transformer embedded video anomaly detection using attention driven recurrence	Attention recurrence boosts detection	Vision Transformer + attention recurrence	AUC, precision, recall, runtime	Dataset diversity, false alarms
arXiv:2504.05299v1	SmolVLM:Redefining small and efficient multimodal models	Lightweight models retain accuracy	Compact VLMs, distillation, lightweight transformers	Retrieval accuracy, latency, memory usage	Limited long-video modelling
EAAI110787	Cross-modal Target Retrieval for Tracking by Natural Language	Joint VLM features improve tracking	Joint feature extraction + alignment with VLM embeddings	Tracking accuracy, robustness tests	Integration complexity, limited benchmarks
arXiv:2405.17247v1	An Introduction to VisionLanguage Modeling	Multimodal fusion enhances understanding	CLIP, encoderdecoder VLMs, adapters	Retrieval accuracy, BLEU, VQA score	Bias, lack standard benchmarks
IEEE ICCV, pp. 6836–6846.	ViViT: A Video Vision Transformer	Pure transformer improves video	Pure Transformer, spatio-temporal tokenization	Top-1 accuracy, ablations, efficiency	High compute, limited realtime
IEEE10205347	Joint Visual Grounding and Tracking with Natural Language Specification	Combines grounding and tracking via language	Joint grounding–tracking framework	mIoU, Success Rate, Precision	Struggles with ambiguity and occlusion
IEEE10285487	Visual Grounding With Joint Multimodal Representation and Interaction	Joint multimodal vision–language representation	Multimodal fusion + interaction modules	Accuracy, IoU, Grounding Recall	High cost, limited realtime evaluation

EAAI10698 8	PRAT: Accurate object tracking based on progressive attention	Progressive attention improves tracking accuracy	PRAT (Progressive Attention Network)	Precision, Success Rate, FPS	No language integration, multi-object challenges
IEEE9857151	Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms	Language guides accurate tracking	Cross-modal retrieval, NLguided tracking	Retrieval accurac, tracking precision	Generalization to unseen language queries

III. METHODOLOGY

In order to ensure a systematic and credible synthesis of the existing body of literature, this review follows a structured methodology grounded in established practices for academic review papers. The approach integrates transparent literature selection, thematic categorization of prior work, and the elaboration of a proposed framework that combines vision-language modelling with sentiment analysis for intelligent surveillance. This section highlights the steps undertaken in the review process, provides an overview of system architecture relevant to context-aware video analysis, and outlines the comparative evaluation of methodologies The literature surveyed for this review was sourced from established academic databases and digital libraries to ensure reliability and scholarly credibility. The primary sources included:

- Scopus – for multi-disciplinary coverage with citations tracking.
- IEEE Xplore – for works related to computer vision, video analytics, and applied AI in security systems.
- Web of Science – for extracting highly cited foundational studies and review articles.
- ScienceDirect (Elsevier) – for applied research papers in machine learning, multimedia, and intelligent systems.
- PubMed – for healthcare-related surveillance applications, where monitoring overlaps with patient safety and medical imaging.
- ACM Digital Library – for AI and human-computer interaction aspects underpinning surveillance technologies.

IV. Literature Selection Strategy:

The timeframe chosen for this analysis spanned 2013–2025, which represents the period of accelerated advancement in deep learning, transformer architectures, and multimodal systems. Earlier works prior to 2013—dominated by rule-based systems or handcrafted computer vision techniques—were selectively included only if they represented seminal contributions to surveillance research.

V. Inclusion and Exclusion Criteria

The review was constrained by the following criteria:

Inclusion:

o Peer-reviewed journal and conference papers.
o Publications in English.
o Recent studies (last 10–12 years) focusing on real-world surveillance applications.
o Work involving computer vision, video activity recognition, image captioning, or sentiment analysis as applied to intelligent monitoring.

Exclusion:

o Non-peer-reviewed sources (blogs, unverified online reports).
o Highly domain-specific medical imaging works unrelated to general security and surveillance.
o Studies without experimental validation or proper benchmarking.

This ensured that only methodologically sound and practically relevant works were included in the review.

VI. CATEGORIZATION AND ANALYSIS APPROACH

The selected works were categorized based on a thematic approach, which enabled the identification of recurring patterns and technological trajectories:

- Traditional Surveillance Techniques motion-detection, anomaly detection, and rule-based monitoring.
- Deep Learning-Based Video Analytics CNNs, RNNs, 3D ConvNets, and action recognition methods.
- Vision-Language Models for Captioning BLIP, CLIP, and related multimodal approaches that generate textual descriptions.
- Sentiment Analysis and NLP Techniques transformer-based sentiment classification and contextual decision-making.
- Integrated and Hybrid Security Systems emerging systems combining visual intelligence with contextual awareness.
- The analysis approach thus combines chronological trends (early handcrafted vs. data-driven deep learning systems) with methodological themes (single-modality vs. multimodal).

VII. SYSTEM FLOW AND ARCHITECTURE

The proposed framework for context-aware surveillance follows a multi-stage system flow:

1. Video Input Capture – real-time acquisition of frames from CCTV feeds.
2. Frame Pre-Processing – noise reduction, resolution standardization (using OpenCV).
3. Vision-Language Modelling – BLIP model generates textual captions of activities detected in the frame.
4. Sentiment Analysis – generated captions are processed through a transformer-based sentiment model to determine whether the described action is normal, suspicious, or hostile.
5. Alert Generation – If negative or aggressive sentiment is detected, the system produces a descriptive, actionable alert.
6. Dashboard/Interface – alerts are displayed in real time on a user interface designed for security personnel.

This sequence forms a pipeline architecture, which can be implemented either on cloud infrastructure or as an edge-based solution with GPU support for low latency

VIII. Tabular Comparison of Methods

Below is a comparative summary of approaches commonly used in intelligent surveillance.

Table - 2: Comparison of Methods Method /Model	Approach	Strengths	Limitations	Performance Trends
Motion Detection & Rule-based	Threshold-based activity monitoring	Lightweight, real-time, low computing power	High false positives, no contextual reasoning	Effective only in controlled environments
CNN/RNN/3D ConvNets	Deep learning video analytics	Robust recognition, adaptable to new	Requires large datasets, GPU-intensive	Improved action recognition (80–90%

		datasets		accuracy in studies)

IX. Proposed Method

The proposed model integrates Salesforce BLIP for image captioning with a transformer-based sentiment analysis classifier to enable real-time, context-aware monitoring. By leveraging BLIP, the system extracts *semantic descriptions* directly from video frames, which reflect human-understandable activities rather than raw motion patterns. These captions are then processed through a sentiment classifier (e.g., BERT or RoBERTa fine-tuned for aggression/threat detection), enabling a higher-level decision layer where security alerts are issued only when descriptions indicate negative or suspicious intent.

Key features of this integrative approach include:

- Human-readable explanations: Each alert carries a natural language description of the detected event.
- Reduced false positives: Alerts are sentiment-driven rather than motion-triggered, making them more relevant.
- Contextual intelligence: The system differentiates between benign activities (e.g., walking, sitting) and harmful scenarios (fighting, breaking objects).
- Scalability and application: Deployment across critical environments such as airports, banks, and hospitals where descriptive and interpretable alerts improve operational response.

The novelty of this method lies in bridging computer vision and natural language analysis for actionable outcomes, transforming surveillance from passive monitoring to active situational intelligence

X. THEMATIC/STRUCTURED BODY

Thematic Synthesis of Context-Aware Surveillance Systems

This section organizes the review along major research themes that have shaped the development of surveillance technologies from traditional to state-of-the-art context-aware intelligent systems. The thematic approach highlights progress, challenges, and open questions across four core dimensions: (1) traditional motion-based surveillance, (2) deep learning-driven video analytics, (3) vision-language models for descriptive captioning, and (4) natural language processing for sentiment-aware security monitoring. Each theme summarizes foundational knowledge, key findings, ongoing debates, and identifies gaps that justify the integration proposed in this review.

Traditional Motion-Based Surveillance:

Early and widely deployed surveillance systems predominantly utilized motion detection and simple rule-based algorithms to trigger alerts. These methods monitor pixel changes or track object movements within a frame sequence, flagging unusual activity primarily based on velocity or spatial anomalies. Research in this domain demonstrated the feasibility of automated monitoring but also revealed significant limitations:

- High false alarm rates caused by innocuous movements such as animals, weather effects (rain, shadows), or minor environmental changes.
- Lack of semantic context, preventing understanding why an alert triggered or what a suspicious behaviour entailed.
- Dependence on preset thresholds and manual rule adjustment, limiting adaptability to changing environments.

Although computationally efficient and suitable for legacy systems, these techniques became inadequate for complex settings requiring nuanced interpretations of dynamic human activities. The field thus experienced a paradigm shift with the rise of deep learning approaches.

Deep Learning for Video Analytics:

The advent of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) enabled researchers to move beyond pixel-level motion detection toward feature-based learning of human actions and anomaly detection. Techniques such as 3D CNNs, Long Short-Term Memory (LSTM)

networks, and 3D ConvNets modelled spatial-temporal patterns within video feeds. These approaches demonstrated:

- Improved recognition accuracy for specific activities such as fighting, running, or loitering.
- Capability to generalize across different scenes and lighting conditions through transfer learning.
- End-to-end learning of actions without manual feature engineering.

Despite these advances, deep learning models often operate as black boxes, providing classification outputs without intuitive explanations. Their performance is also constrained by the requirement of large annotated datasets, high computational cost, and challenges handling complex multi-person interactions. These limitations motivated integration with multimodal systems capable of generating human-readable interpretations.

Vision-Language Models for Surveillance Captioning:

The integration of computer vision and natural language processing led to the development of Vision-Language Models (VLMs) designed to generate natural language descriptions (captions) from images or video frames. Models such as Salesforce BLIP and OpenAI's CLIP employ large-scale pretraining on paired image-text datasets, enabling them to:

- Produce detailed descriptions of objects, scenes, and activities detected within images.
- Offer contextual information that static action classification misses, such as "A man running quickly inside the building" rather than just "running."
- Facilitate interoperability between visual data and language-based reasoning systems.

However, challenges persist in real-time implementation due to the models' computational overhead, and the accuracy of captions can degrade in visually complex or low-quality feeds. These factors highlight the need for optimized architectures and the combination with downstream text analysis.

Sentiment Analysis and Context-Aware Alerting:

Sentiment analysis applies transformer-based NLP techniques such as BERT, RoBERTa, and their variants to classify the polarity or emotional tone of textual inputs. By analysing captions generated from video frames, these models can:

- Discern whether described actions signal aggressive, suspicious, or benign behaviour.
- Assign interpretative labels that transform raw descriptions into actionable security insights.
- Reduce false alarms compared to motion-only triggers by focusing alerts on negative or threatening sentiment.

The main constraint is the potential for misinterpretation in complex scenarios where context is subtle or ambiguous—for example, distinguishing a playful push from an aggressive shove. Research is ongoing to improve context sensitivity through multimodal fusion and temporal analysis of sequential frames.

Key Findings, Limitations, and Gaps:

Across these themes, several common insights and challenges emerge:

- There is a clear evolution from low-level motion detection to high-level semantic understanding, driven by advances in AI and model architectures.
- Combining vision-language models with sentiment analysis addresses critical gaps in explainability and alert relevance but remains an emerging area requiring further research into robustness and efficiency.
- Most existing studies focus on single-frame analysis or offline datasets, with limited exploration of real-time deployment in noisy and resource-constrained environments.
- Privacy and ethical concerns related to surveillance persist as under-addressed issues in technical literature, necessitating interdisciplinary approaches involving policy and social sciences.
- This structured synthesis highlights the trajectory of surveillance technology and the rationale for the proposed context-aware framework, aiming to bridge existing gaps by delivering human-like, interpretable activity

descriptions combined with sentiment-driven alerting for enhanced security monitoring.

XII. CRITICAL ANALYSIS & SYNTHESIS

Critical Analysis and Synthesis

The landscape of surveillance technology demonstrates a dynamic evolution marked by shifting paradigms from simple motion detection to sophisticated context-aware systems that leverage advances in artificial intelligence, particularly in vision-language modelling and sentiment analysis. This section critically analyses and synthesizes the key findings from existing studies, highlighting significant trends, contrasting approaches, and areas meriting further exploration to advance the field.

Comparing and Contrasting Existing Studies

Traditional systems, largely reliant on pixel-level motion detection, provided foundational tools for surveillance yet suffered from high false alarm rates and a lack of semantic insight . In contrast, deep learning-based video analytics introduced more nuanced understanding by analysing spatial-temporal features through architectures like 3D CNNs and LSTMs . These approaches significantly improved detection accuracy for predefined actions but often remained opaque 'black-box' classifiers, leaving end-users without explanations for alerts.

Vision-language models such as Salesforce BLIP and OpenAI's CLIP mark a paradigm shift by generating explicit, natural language captions that describe what is visible in camera feeds . These models have demonstrated superior performance in bridging visual content with textual understanding, enabling more interpretable surveillance systems. Sentiment analysis models further enhance interpretability by categorizing descriptions for threat levels, delivering decision-making support .

Nevertheless, there is a marked divergence in approaches regarding real-time applicability. While traditional and some deep learning systems prioritize speed and efficiency, VLMs integrated with sentiment classifiers often grapple with computational demands that limit deployment on edge devices or resource-constrained environments. Studies differ in their strategies to balance latency against descriptive richness, a critical trade off yet unresolved.

Trends and Patterns:

A consistent trend is the movement from detection to interpretation. Earlier works focused on event detection as an end goal, but recent studies emphasize the need for contextual descriptions that add valuable semantic meaning. Multimodal approaches combining image captioning and NLP represent an emerging pattern that aims to produce actionable insights rather than mere flags.

Another observable pattern is the increasing adoption of transformer-based architectures for both vision and language tasks, which has led to enhanced accuracy and versatility. Pretrained models fine-tuned on domain-specific datasets form a growing practice to mitigate the challenges posed by limited labelled data in security contexts.

Furthermore, integration with cloud and edge computing provides a pathway for scalable, real-time surveillance, illustrating a trend toward distributed processing architectures that blend high-performance GPU capabilities with on-site decision nodes.

Contradictions and Controversies:

The literature reveals some contradictions, particularly concerning the evaluation of system effectiveness. While quantitative metrics such as accuracy, precision and recall dominate, there is less consensus on qualitative measures like user trust, alert interpretability, and cognitive load on security personnel. Few studies systematically evaluate how descriptive alerts impact operational decision-making or reduce response times in practice.

Another area of debate is the balance between surveillance capability and privacy. Ethical considerations and regulatory compliance often receive insufficient attention in technical papers, though recent discourse increasingly calls for embedding privacy-preserving mechanisms and ensuring system transparency .

Additionally, the relative merits of different sentiment analysis models in handling the ambiguity of natural scenes remain contested. Some research argues transformer models excel, while others point

to the need for multimodal sentiment analysis that incorporates audio and contextual metadata.

Methodological Strengths and Weaknesses

Many studies' strengths lie in their robust experimental design, using benchmark datasets such as UCF101 or MS-COCO for image captioning and sentiment benchmarks like SST (Stanford Sentiment Treebank). However, limitations include the frequent use of static single-frame analysis rather than leveraging temporal continuity available in video data, which may omit dynamic context critical for threat evaluation.

Another methodological weakness is the lack of unified datasets encompassing both visual and textual annotations reflecting security scenarios, hindering direct comparison across studies. Few works integrate human-in-the-loop evaluation or simulate real-world operational conditions, reducing ecological validity.

On the positive side, methodological innovations such as fine-tuning VLMs for specific surveillance domains and the application of transfer learning have enabled moderate generalization despite the small training sets typical of this domain.

Underexplored Areas:

There remain important gaps in context-aware surveillance research. Most notably:

- Temporal context integration: Very few systems fully exploit sequential frame analysis to refine sentiment evaluation or disambiguate transient actions.
- Multimodal fusion: Combining visual data with audio, sensor metadata, or contextual location information is underutilized but crucial for holistic understanding.
- Adaptability and personalization: Customizing alert thresholds and behaviour profiles for specific environments or security needs remains an open challenge.
- Explainability and transparency metrics: Developing standardized ways to assess and communicate model interpretability and user trust needs deeper investigation.
- Ethical, legal, and social implications: Privacy-preserving AI models and compliance mechanisms tailored for surveillance contexts are emerging but require accelerated exploration.

XII. FUTURE DIRECTIONS / RESEARCH GAPS

Future Directions and Research Gaps

As context-aware surveillance systems integrating vision-language models and sentiment analysis continue to evolve, several promising avenues for future research and key gaps remain to be addressed. This section outlines critical areas where advancing methodology, technology, and interdisciplinary collaboration can substantially improve system performance, reliability, and societal impact. These recommendations aim to guide researchers in closing current voids and exploring innovative frontiers in intelligent security monitoring.

Enhancing Real-Time, Scalable Performance

Achieving real-time processing remains a substantial challenge due to the high computational demands of large-scale vision-language models and transformer-based sentiment classifiers. Future work should focus on:

- Model optimization and compression techniques: Research on neural network pruning, quantization, and knowledge distillation could enable more efficient models that retain high accuracy while operating on edge devices or resource-constrained environments.
- Hardware-Software Co-design: Tailoring architectures that leverage GPU, FPGA, or specialized AI accelerators designed for multimodal inference can improve latency and throughput.
- Distributed and edge-cloud hybrid computing: Designing frameworks that intelligently partition computation between local devices and cloud services depending on network and workload conditions will enhance scalability and responsiveness.

Integrating Temporal and Multimodal Context

Current approaches mostly focus on single-frame captioning and sentiment evaluation, limiting

understanding of evolving scenarios or subtle behavioural cues. Future research can address this by:

- Temporal sequence modelling: Developing architectures that incorporate frame-to-frame temporal relationships using video transformers, 3D convolutional networks, or recurrent structures to capture dynamic event context.
- Multimodal fusion: Incorporating additional sensory data such as audio streams, depth sensors, thermal imaging, or contextual metadata (e.g., location, time of day) can enrich semantic interpretation and reduce ambiguity in detection.
- Multimodal sentiment and intent estimation: Holistic analysis of combined modalities could lead to more robust threat assessment beyond visual cues alone.

XIII. MPROVING INTERPRETABILITY AND USER INTERACTION

To ensure surveillance systems translate technical outputs into actionable insights, more research is needed around:

- Explainable AI (XAI) techniques: Methods that elucidate why particular captions or alert classifications are generated will increase trust and facilitate human operator decision-making.
- Adaptive alerting systems: Investigating personalization frameworks that learn user preferences or dynamically adjust threat thresholds based on operational context can minimize alert fatigue.
- Human-in-the-loop designs: Integrating user feedback to iteratively improve model predictions and interface usability remains an open challenge.

XIV. ADDRESSING PRIVACY, ETHICS, AND TRUST

As surveillance technologies intensify scrutiny over individual activities, balancing security with privacy and ethical accountability is critical:

- Privacy-preserving AI: Research on federated learning, differential privacy, and secure multi-party computation could provide avenues for training and deploying models without exposing sensitive data.
- Regulatory compliance frameworks: Developing transparent mechanisms aligned with international legal standards and ethical principles will facilitate safer, more acceptable deployments.
- Bias mitigation: Ensuring datasets and models are representative and free from demographic or situational biases requires standardized audit protocols and fairness-enhancing algorithms.

XV. CREATING COMPREHENSIVE BENCHMARK DATASETS

There is a pronounced scarcity of large-scale datasets explicitly tailored for context-aware surveillance with paired video and caption/sentiment annotations:

▪ Domain-specific multimodal datasets: Curating datasets that combine high-quality video feeds with rich linguistic descriptions and sentiment labels in diverse scenarios (public spaces, hospitals, critical infrastructure) will spur method development.

▪ Real-world and adversarial scenarios: Including noisy, ambiguous, and adversarial behaviour types to test model robustness reflects real operational challenges absent in clean academic datasets.

XVI.EXPLORING SOCIETAL IMPACT AND SUSTAINABILITY

Finally, the broader implications of deploying intelligent surveillance warrant parallel investigation:

- Social acceptability studies: Understanding public attitudes toward automated surveillance and human-like interpretation systems will inform design and deployment strategies.
- Alignment with Sustainable Development Goals (SDGs): Research articulating how context-aware surveillance contributes to SDG 11 (Sustainable Cities) and SDG 16 (Peace, Justice, and Strong Institutions) can anchor technological advances within global human development frameworks.
- Interdisciplinary collaboration: Combining expertise from computer science, ethics, social

sciences, and policy can produce more holistic solutions that balance innovation with responsibility.

XVII. Conclusion

The advancement of context-aware surveillance systems marks a pivotal shift in security technology, emphasizing interpretability and actionable intelligence rather than mere detection. This review underscores the transformative potential of integrating vision-language models with sentiment analysis to deliver descriptive, human-like alerts that align surveillance outputs with the cognitive needs of security personnel. By bridging the visual and linguistic domains, these systems enhance situational awareness, reduce false alarms, and facilitate faster, more informed responses.

The critical examination of existing methodologies highlights both the promise and limitations of current approaches, clarifying the trade-offs between computational efficiency, accuracy, and privacy. This synthesis provides a cohesive framework that situates these technologies within broader social and ethical contexts, emphasizing the necessity of multidisciplinary integration to address real-world complexities. Moreover, the identification of research gaps and future directions shapes a progressive research agenda to guide continued innovation.

Ultimately, this review contributes by structuring disparate research strands into a coherent narrative, fostering deeper understanding, and promoting the development of intelligent surveillance systems that are not only technologically advanced but also ethically responsible and socially impactful. Such systems have the potential to profoundly enhance public safety and urban resilience in alignment with sustainable development goals.

Acknowledgements

The authors sincerely thank their project guide and faculty members for their guidance, support, and valuable feedback throughout the development of this review. Their insights and suggestions were instrumental in enhancing the clarity and quality of the work. The authors also express their gratitude to the department for providing a supportive academic environment and the necessary resources to complete this study. Furthermore, the authors acknowledge the researchers and scholars whose published work has been reviewed and cited, as their contributions significantly enriched this paper.

REFERENCES

1. Wang, X., Li, C., Yang, R., Zhang, T., Tang, J., Luo, B. (2018). Describe and Attend to Track: Learning Natural Language Guided Structural Representation and Visual Attention for Object Tracking. arXiv:1811.10014.
2. Zhao, M., Okada, K., Inaba, M. (2021). TRTR: Visual Tracking with Transformer. arXiv:2105.03817.
3. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y. (2022). CoCa: Contrastive Captioners Are Image-Text Foundation Models. arXiv:2205.01917.
4. Kim, W., Son, B., Kim, I. (2021). ViLT: Vision-andLanguage Transformer without Convolution or Region Supervision. arXiv:2102.03334.
5. Zeng, Y., Zeng, B., Hu, H., Zhang, H. (2023). PRAT: Accurate Object Tracking Based on Progressive Attention. Engineering Applications of Artificial Intelligence, 126, 106988.
6. Li, Y., Yu, J., Cai, Z., Pan, Y. (2022). Cross-Modal Target Retrieval for Tracking by Natural Language. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
7. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C. (2021). ViViT: A Video Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6836–6846.
8. Shu, H., Lu, Q., Xue, L., Xue, M., Yuan, G., Zhong, B. (2023). Visual Grounding with Joint Multimodal Representation and Interaction. IEEE Transactions on Instrumentation and Measurement, 72, 5031811.
9. Zhou, L., Zhou, Z., Mao, K., He, Z. (2023). Joint Visual Grounding and Tracking with Natural Language Specification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

10. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F. (2021). Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13763–13773.

Author's details

1. Vibhavari Jawale, Ass. Professor, Dept. of AI & DS, Dr. D. Y. Patil Institute of Engineering, Management and Research, Pune, India, vibhavarijawale175@gmail.com
2. Deepali Hajare, Ass. Professor, Dept. of AI & DS, Dr. D. Y. Patil Institute of Engineering, Management and Research, Pune, India, deepali.hajare@dypiemr.ac.in
3. Arhant Sahuji, UG Scholar, Dept. of AI & DS, Dr. D. Y. Patil Institute of Engineering, Management and Research, Pune, India, arhantsahuji134@gmail.com
4. Tanay Shinde, UG Scholar, Dept. of AI & DS, Dr. D. Y. Patil Institute of Engineering, Management and Research, Pune, India, tanayshinde1820@gmail.com
5. Ananya Vaishnav, UG Scholar, Dept. of AI & DS, Dr. D. Y. Patil Institute of Engineering, Management and Research, Pune, India, ananyavaishnav05@gmail.com
6. Ritesh Kadam, UG Scholar, Dept. of AI & DS, Dr. D. Y. Patil Institute of Engineering, Management and Research, Pune, India, riteshbkadam@gmail.com