Development of Intelligent CCTV System with Zero-Shot Learning in Closed Network Environment for Nuclear Security Facilities

Minjong Kim[a], Dongju Kim[a], Jaeyong Lee [b], Yonggyun Yu [a*]
[a]*Applied Artificial Intelligence Section, Korea Atomic Energy Research Institute, Daedeok-daero 989beon-gil*
[b]*Physical Protection Team, Korea Atomic Energy Research Institute, Daedeok-daero 989beon-gil*
[*]*Corresponding author: ygyu@kaeri.re.kr*

***Keywords :** intelligent CCTV, zero-shot learning, closed network environment, security management

1. Introduction

With increasing reports of physical attacks on nuclear facilities worldwide, there is a growing emphasis on round-the-clock surveillance and rapid response. In this context, intelligent CCTV systems have emerged as an effective means to reduce security personnel fatigue and identify abnormal situations in real time, enhancing both the efficiency and accuracy of security monitoring [1–3].

Given the stringent security standards at national facilities, especially those operating within closed networks that are completely isolated from external internet access, selecting the appropriate CCTV architecture becomes crucial. Common deployments cloud-based, on-device, on-premises, and hybrid each exhibit trade-offs in security, flexibility, and scalability. To minimize the risk of video data leakage while enabling advanced analytics, the on-premises approach is often favored for critical installations. In order to address the demanding security requirements of nuclear facilities, this paper presents a prototype implemented within the closed-network environment of KAERI and discusses the broader potential for applying this approach across other nuclear installations.

This paper proposes an intelligent CCTV system fully implemented within the closed-network environment of KAERI, in collaboration with the physical protection team. The system integrates Zero-Shot Learning (ZSL) [4], an advanced machine learning technique that differs fundamentally from traditional approaches. While conventional CCTV systems can only recognize objects or behaviors they were explicitly trained on, ZSL can identify novel entities described through text prompts without requiring additional training data. This capability is particularly valuable in nuclear facilities where security threats evolve rapidly and access to external data for model retraining is limited by network isolation requirements.

The remainder of this paper is structured as follows: Section 2 introduces the proposed methodology, including system architecture, user interface design, and specific steps to integrate ZSL. Potential application scenarios such as detecting unauthorized personnel, tracking the movement of sensitive equipment, and identifying unusual behaviors are also discussed. Section 3 presents the conclusion and future directions, emphasizing how this system can enhance nuclear facility security and accommodate further extensions, including guided data training methods

2. Methodology

This chapter presents the overall structure and prototype of an intelligent CCTV system specifically designed for the high-security environment at the KAERI. By offering a comprehensive description of the system architecture and a detailed implementation plan, the proposed approach meets the stringent security requirements of KAERI under a closed-network infrastructure.

2.1 AI-Based Video Transmission System and User Interface Development

The network environment of KAERI is structured for security, with the external and internal networks completely isolated, and the internal network operating in a closed manner. In order to establish an effective CCTV surveillance system within this closed network environment, it is crucial to adopt an approach that remains compatible with existing infrastructure while meeting stringent security requirements. We propose a system architecture that can securely handle video without modifying the existing CCTV infrastructure. This approach directly captures the video signal and delivers it to an analysis system within the closed network environment, enabling implementation without additional changes to the infrastructure. In the setup we propose, the capture card is connected to the closed network environment inside KAERI, thereby eliminating any possibility of video leakage to the outside. In order to effectively utilize the video acquired through the capture card, a user interface (UI) is necessary to monitor CCTV footage and intuitively display analysis results. The prototype interface provides real-time monitoring, analysis, and AI capabilities, and is organized in a tab format for user convenience. When anomalies are detected, an immediate alert can be sent to control center personnel, and further improvements are being made to maximize operational efficiency.

2.2 Application of Zero-Shot Learning

Fig. 1. This is a prototype image of an intelligent CCTV system using zero-shot learning (ZSL) to detect an individual committing arson. Due to security concerns that prohibit the use of internal images, the image displayed in the UI is sourced from the AI Hub dataset.

ZSL is the core AI technology leveraged in this study. This advanced machine learning paradigm can identify objects and situations without explicit prior training examples. Unlike conventional AI systems that require large labeled datasets for each detection category, ZSL can generalize to entirely new categories based solely on semantic descriptions. In our intelligent CCTV system, ZSL operates by linking visual information from CCTV footage with textual descriptions of security scenarios in a shared embedding space.

Concretely, the system extracts frames from the CCTV footage and processes them via a specialized neural network known as an image encoder to convert each frame into a high-dimensional feature vector. Meanwhile, security personnel can define potential threats in natural language as text prompts, which are transformed into vector representations by a language model serving as a text encoder. Both sets of vectors are mapped to the same embedding space, allowing direct comparison of their semantic similarity. When the similarity between an encoded image and a text prompt surpasses a predefined threshold, the system identifies a match and raises an appropriate alert.

Traditional intelligent CCTV systems typically recognize only objects or scenarios for which they have been trained in advance, limiting their responsiveness to newly emerging or unforeseen abnormalities. In contrast, ZSL leverages accumulated knowledge to detect and classify threats not previously encountered, enabling rapid detection and response to unexpected security challenges. For instance, the system can detect unusual human behaviors or objects that do not normally appear in a secure area, issuing alerts upon identifying such anomalies. Figure 1 illustrates a prototype of our intelligent CCTV system employing ZSL to detect an individual committing arson. One widely used approach for implementing ZSL in tasks such as object detection, Visual Question Answering (VQA), and image captioning involves multimodal learning techniques [5]. A simplified representation of this process is shown in Equation (1).

$$(1)\quad sim = \frac{f_i(I) \cdot f_t(T)}{\| f_i(I) \| \| f_t(T) \|}$$

In this formulation, I denotes the input image, T represents the text prompt, and f_{i} and f_t refer to the respective image and text encoders. The $sim(\mathrm{I},\mathrm{T})$ indicate the cosine similarity between these two embeddings. When $sim(\mathrm{I},\mathrm{T})$ surpasses a predefined threshold or achieves the highest similarity score among multiple candidates, the system identifies the corresponding object or action. This recognition is possible even in scenarios where the system is never explicitly trained. This prototype application provides real-time frame acquisition and allows users to control detection tasks, including pausing, resuming, and capturing. It also enables dynamic loading of images from specified folders. Additionally, it offers a user-friendly multi-tab GUI for diverse analytical functions. Users can easily enable or disable the ZSL feature via a checkbox and switch between live camera feeds and offline image folders. Moreover, it supports both scene-based visual question answering (VQA) and image captioning, further expanding its analytical capabilities. All processing is carried out within the closed network environment of KAERI, ensuring strict adherence to security guidelines while leveraging AI technology to maintain a robust security environment.

2.3 Potential Application Scenarios

ZSL enables security personnel to define diverse surveillance scenarios using simple text-based prompts. This flexibility is particularly valuable in high-risk settings such as nuclear facilities. By converting textual prompts directly into visual detection tasks, the system can adapt to novel or evolving threats without extensive model retraining. A prime example of this adaptability is early fire detection. Comprehensive prompts can describe various indicators of fire risk, including sudden smoke accumulation, unusual heat signatures near electrical equipment, unauthorized use of flammable materials, or abnormal electrical system behaviors. The ZSL model maps these textual descriptions to visual features in real time, creating a proactive mechanism to alert personnel before fire-related incidents escalate. Intrusion detection can also be customized through targeted prompts, such as identifying individuals who violate dress code policies, carry suspicious objects, display irregular movement patterns, or attempt to access restricted zones without proper authorization. In addition, these text-based prompts can be extended to address specialized needs in nuclear environments, such as tracking the movement of radiation-containment equipment or verifying compliance with protective gear protocols, thereby enabling the system to handle a broader range of complex security scenarios. Moreover, the ZSL based intelligent CCTV system can detect changes in the posture of persons using only a simple text prompt such as a person lying on the ground. Because it can discern the difference between a typical standing posture and one that indicates the individual is lying down or has collapsed, based on preexisting learned information, it can promptly identify emergency situations and transmit alerts without requiring additional training. By integrating such capabilities, nuclear facilities can swiftly respond to potential safety incidents and thereby reduce risks associated with delayed interventions.

3. Conclusion

The prototype we are developing in this study demonstrates the effectiveness of an intelligent CCTV solution suitable for critical facilities requiring a high level of security. This is achieved through the implementation of advanced AI technologies, especially flexible analytical methods like ZSL. In particular, they can be applied in a closed network environment without altering existing CCTV infrastructure. Moreover, the various extension scenarios proposed in Section 2.3 highlight the breadth of potential applications in real security operations. System features are refined to meet on-site requirements in collaboration with physical protection teams. Currently, this research remains at the prototype stage, and efforts are underway to evaluate ZSL performance in real-world operating environments and expand its functionality. In future work, we plan to integrate vision Large Language Models (LLMs) to generate task-specific prompts for personnel movement tracking, equipment monitoring, and other security applications. By combining these vision LLM-driven prompts with anomaly detection pipelines, our system will be able to more accurately identify unforeseen or context-specific threats in closed-network environments.

ACKNOWLEDGEMENT

This research was supported by a grant from Korea Atomic Energy Research Institute (KAERI) R&D Program (No. KAERI-524540-25). This paper used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

REFERENCES

[1] De Cauwer, H., Barten, D. G., Tin, D., Mortelmans, L. J., Ciottone, G. R., & Somville, F. (2023). 50 years of terrorism against the nuclear industry: a review of 91 incidents in the Global Terrorism Database. Prehospital and disaster medicine, 38(2), 199-206.

[2] Duja, K. U., Khan, I. A., & Alsuhaibani, M. (2024). Video Surveillance Anomaly Detection: A Review on Deep Learning Benchmarks. IEEE Access.

[3] Direkoglu, C. (2020). Abnormal crowd behavior detection using motion information images and convolutional neural networks. IEEE Access, 8, 80408-80416.

[4] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2024, September). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision (pp. 38-55). Cham: Springer Nature Switzerland.

[5] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning (pp. 12888-12900). PMLR.