Deep Learning-based Integrated 2D-3D Video Analysis for Hazard Recognition in Surveillance Applications

Dongchil Kim
Information Media Research Center
Korea Electronics Technology Institute
Seongnam, Korea
dckim@keti.re.kr

Kyeongeun Seo
Information Media Research Center
Korea Electronics Technology Institute
Seongnam, Korea
ke.seo@keti.re.kr

Chang Mo Yang
Information Media Research Center
Korea Electronics Technology Institute
Seongnam, Korea
cmyang@keti.re.kr

***Abstract*—This paper presents a deep learning-based integrated 2D–3D video analysis system for hazardous behavior recognition in surveillance applications. A multimodal RGB–Depth dataset was constructed across urban buildings, industrial sites, and construction sites, covering 15 representative hazardous behaviors frequently observed in real-world surveillance scenarios. An R(2+1)D convolutional neural network was trained and optimized using this dataset to effectively learn spatio-temporal patterns from synchronized RGB and depth inputs. Experimental results show that the proposed system can accurately detect various hazardous behaviors in real time by jointly analyzing 2D and 3D visual information, demonstrating its potential for intelligent surveillance and rapid risk response applications.**

Keywords—hazardous behaviors recognition, integrated 2D–3D video analysis, multimodal dataset, intelligent surveillance

I. Introduction

With the rapid expansion of urban infrastructure and industrial facilities, real-time hazard recognition systems have become critical for ensuring safety and security. Environments such as urban buildings, construction sites, and industrial sites are vulnerable to various dangerous behaviors including intrusion, theft, abandonment, falldown, falling, loitering, fire, weapon detection, and safety violations. Conventional CCTV surveillance systems rely primarily on 2D RGB video. However, they often perform poorly in environments with low illumination, occlusions, or complex spatial layouts. To address these limitations, recent research has actively explored multimodal RGB–Depth video analysis to improve the robustness and accuracy of action recognition [1–3]. For example, Rahmaniboldaji et al. proposed a depth-enhanced action recognition framework to improve recognition accuracy under challenging visual conditions [1]. Kini et al. investigated egocentric RGB–Depth action recognition in industry-like environments, showing that multimodal approaches can improve action recognition performance for workplace monitoring and safety [2]. Furthermore, Zhang and Wang [3] presented a comprehensive survey on RGB–D action recognition methods, highlighting the importance of temporal modeling and multimodal feature fusion for reliable surveillance systems. Based on these advancements, this paper proposes a deep learning-based integrated 2D–3D video analysis system that utilizes synchronized RGB and depth data to detect hazardous behaviors across multiple surveillance domains. A multimodal dataset was constructed containing 15 hazard categories, and an R(2+1)D CNN was optimized to learn spatio-temporal features for accurate, real-time hazard recognition [4,5,6].

The remainder of this paper is organized as follows. Section 2 describes the multimodal dataset construction and the proposed R(2+1)D-based system architecture. Section 3 presents the training strategy and experimental evaluation. Section 4 concludes the paper and discusses future research directions.

II. Integrated 2D-3D Video Analysis System

A. Multimodal Dataset Construction

To build a robust hazard recognition system, a synchronized RGB–Depth dataset was collected in three representative physical security environments: urban buildings, industrial sites, and construction sites. These domains were selected to reflect real-world surveillance scenarios involving diverse spatial structures, activities, and lighting conditions. Fig. 1 shows the overall data collection configuration. RGB and depth sensors were installed at multiple viewpoints to simultaneously capture complementary 2D and 3D information. Each recording session included various scenarios such as day and night lighting, background complexity, and the presence of multiple actors to enhance the diversity of the collected data. Both RGB and depth streams were temporally synchronized to enable effective multimodal learning.

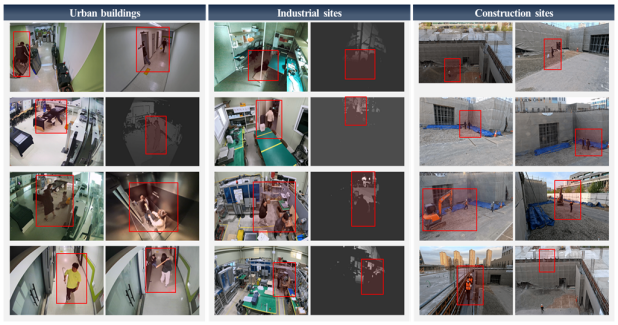


Fig. 1. Example of data acquisition setup and collected RGB–Depth samples in different domains

The dataset defines 15 hazardous behavior classes across the three domains, reflecting security-relevant activities frequently encountered in real surveillance environments. These classes include behaviors such as intrusion, theft, demage, loitering, abandonment and falldown, among others. Table 1 lists the defined hazardous behaviors categorized by domain. In total, the dataset consists of approximately 850 RGB–Depth video clips. For each hazardous behavior class,

multiple clips were recorded to ensure variation in camera viewpoints, lighting, and human actions, which enhances the model's generalization capability.

All sequences were manually annotated with temporal action boundaries and class labels to enable supervised learning. The dataset was divided into training, validation, and testing sets using a 7:2:1 ratio, where 70% of the clips were used for training, 20% for validation, and 10% for testing. This split ensures sufficient data for model training while preserving separate validation and testing subsets for unbiased evaluation. The dataset provides a balanced distribution across domains and behavior classes, enabling the model to learn both domain-invariant features and class-specific patterns. The synchronized RGB–Depth data were stored as paired, frame-indexed sequences to facilitate efficient multimodal loading and fusion during training.

TABLE I. DEFINITION OF 15 HAZARDOUS BEHAVIOR CLASSES CATEGORIZED BY DOMAIN

Number	Hazardous behavior classes	Domain
1	Intrusion	Urban buildings
2	Theft	
3	Occupant Verification Check	
4	Weapon Detection	
5	Violence	
6	Damage	Construction sites
7	Loitering	
8	Danger Exposure	
9	No Helmet	
10	Falling	
11	Abandonment	Industrial sites
12	Falldown	
13	Fire Detection	
14	Toxic Exposure Detection	
15	Compression Injury	

B. System Architecture

The core of the proposed hazardous behavior recognition system is a Residual (2+1)D Convolutional Neural Network (R(2+1)D CNN) [6] as shown in Fig. 2, which has demonstrated state-of-the-art performance in spatio-temporal video understanding tasks. Unlike conventional 3D CNNs that apply a single 3D convolution to learn spatial and temporal patterns simultaneously, the R(2+1)D architecture factorizes this operation into separate spatial and temporal convolutions.

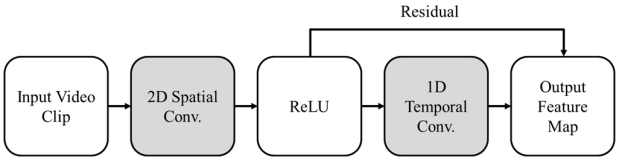


Fig. 2. Structure of the R(2+1)D convolution block. A 3D convolution is factorized into a 2D spatial convolution, a ReLU activation, and a 1D temporal convolution, followed by a residual connection

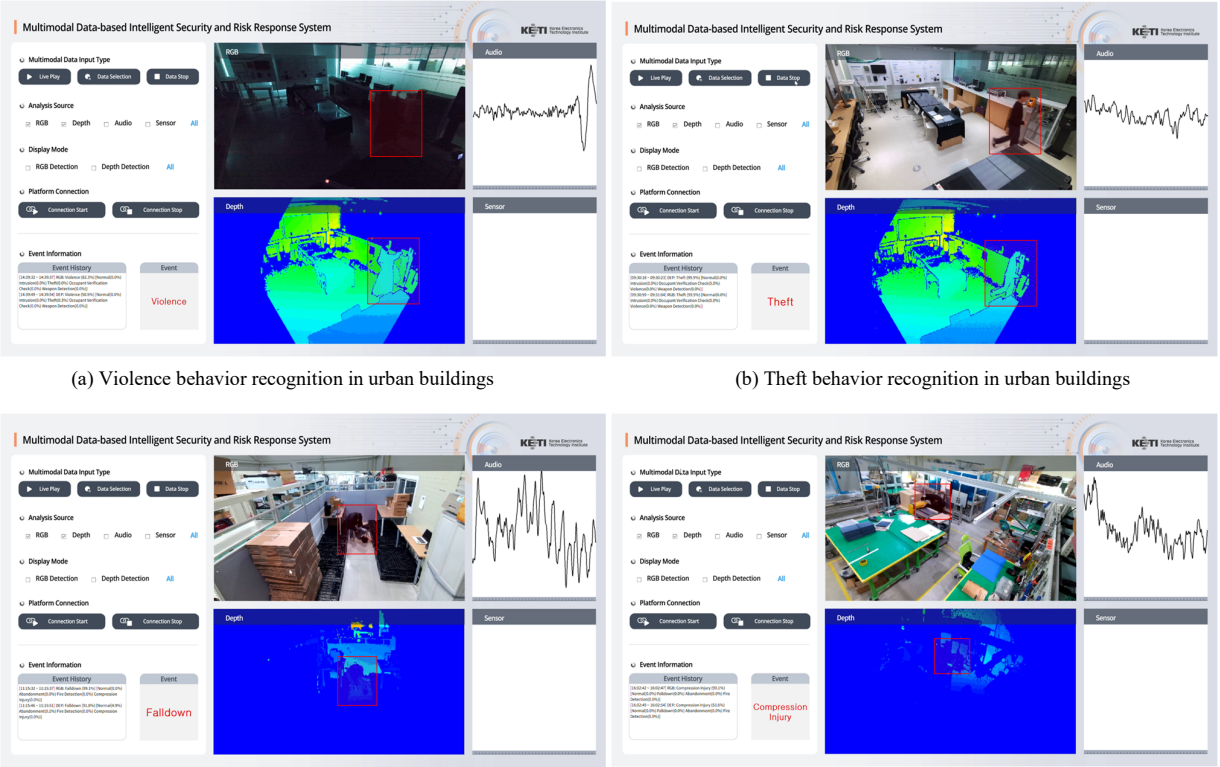
A 2D spatial convolution first extracts appearance features from individual frames, and a 1D temporal convolution then models motion across consecutive frames. The structure of an R(2+1)D convolution block, in which a standard 3D convolution is decomposed into spatial and temporal components with an intermediate non-linearity. This factorization improves training stability, reduces parameters, and enhances spatio-temporal feature representation compared to conventional 3D convolutions [6].

This factorization provides several advantages. It reduces the number of parameters compared to standard 3D convolutions, making the network more computationally efficient. It also introduces an additional non-linearity between the spatial and temporal operations, which improves optimization stability and enables the network to learn richer spatio-temporal representations. Furthermore, separating spatial and temporal processing simplifies training, allowing the model to converge faster while maintaining or improving recognition accuracy. The network adopts a ResNet-style residual structure, where factorized spatio-temporal convolutions are embedded within skip connections. These residual connections facilitate stable gradient flow and enable the training of deeper architectures without degradation. The overall structure consists of multiple stacked spatio-temporal blocks followed by temporal pooling and fully connected layers for classification. For this study, the R(2+1)D backbone was initialized with weights pre-trained on the Kinetics-400 dataset, following the approach in [6]. During fine-tuning, key hyperparameters such as learning rate, batch size, temporal window length, and network depth were optimized to maximize performance. This architecture is particularly suitable for hazardous behavior recognition because it efficiently captures both detailed spatial information (e.g., human posture, objects) and temporal dynamics (e.g., falling, intrusion), while maintaining high computational efficiency and strong generalization in complex surveillance environments.

III. TRAINING AND EVALUATION

To train the proposed system, the constructed RGB–Depth dataset was divided into training, validation, and testing sets as described in Section II. The R(2+1)D model was fine-tuned on the training set using synchronized RGB and depth streams as inputs and hazardous behavior labels as outputs. Supervised training was performed with optimized hyperparameters to ensure stable convergence and reliable recognition performance. For evaluation, the trained system was tested on unseen video sequences to verify its effectiveness in realistic surveillance scenarios.

Fig. 3 shows the experimental results of the proposed system on four hazardous behavior categories: violence, theft, falldown, and compression injury. In the violence case (Fig. 3(a)), the system demonstrates stable recognition performance even under low-illumination conditions, as depth-based analysis compensates for the lack of visual clarity in RGB frames. This enables more accurate detection of aggressive motions compared to using RGB data alone. In the theft scenario (Fig. 3(b)), the model effectively identifies suspicious human behaviors through multimodal fusion, capturing subtle motion cues that distinguish theft from normal activities. For the falldown and compression injury cases (Fig. 3(c) and (d)), the system successfully detects sudden postural changes and human–object interactions in industrial environments, showing strong temporal motion understanding and spatial reasoning capability. Overall, these experimental results confirm that the integrated 2D–3D video analysis approach robustly recognizes various hazardous behaviors across complex environments, achieving improved accuracy and reliability compared to single-modality systems.



(a) Violence behavior recognition in urban buildings

(b) Theft behavior recognition in urban buildings

(c) Falldown behavior recognition in industrial sites

(d) Compression injury behavior recognition in industrial sites

Fig. 3. Experimental results of hazardous behavior recognition for four representative cases: (a) Violence in urban buildings, (b) Theft in urban buildings, (c) Falldown in industrial sites, and (d) Compression injury in industrial sites.

IV. Conclusion

This study presented a deep learning-based hazard recognition system that integrates 2D RGB and 3D depth video analysis for surveillance applications. A multimodal dataset containing approximately 850 clips was constructed across three representative physical security domains—urban buildings, industrial sites, and construction sites—covering 15 hazardous behavior classes. To effectively learn spatio-temporal patterns from these multimodal inputs, an R(2+1)D convolutional neural network was employed, leveraging factorized 3D convolutions to improve computational efficiency and representation learning. The proposed system demonstrated robust performance in recognizing a wide range of hazardous behaviors, including intrusion, falling, and fire, under various environmental conditions. By combining RGB and depth information, the model achieved improved detection accuracy and reliability compared to RGB-only baselines, particularly in challenging scenes with occlusions or low illumination. The main contributions of this work are threefold: (1) the construction of a domain-specific RGB–Depth dataset for physical security applications, (2) the application and adaptation of an R(2+1)D architecture for multimodal hazard recognition, and (3) the demonstration of real-time surveillance event detection using RGB–Depth video streams.

Future research will focus on extending the dataset to cover more diverse environments and behaviors, optimizing the model for deployment on edge devices to enable real-time field applications, and integrating additional modalities such as audio and environmental sensors for enhanced multimodal analysis.

Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00438027, Development of an intelligent security and risk response system based on multi-modal data integration analysis).

References

[1] S. Rahmaniboldaji, F. Rybansky, Q. Vuong, F. Guerin, and A. Gilbert, “DEAR: Depth-Enhanced Action Recognition,” *arXiv preprint* arXiv:2408.15679, 2024.

[2] J. Kini, S. Fleischer, I. Dave, and M. Shah, “Egocentric RGB+Depth Action Recognition in Industry-Like Settings,” *arXiv preprint* arXiv:2309.13962, 2023.

[3] Y. Zhang and Y. Wang, “A Comprehensive Survey on RGB–D-Based Human Action Recognition: Algorithms, Datasets, and Popular Applications,” *EURASIP J. Image Video Process.*, Art. 15, 2025.

[4] D. Kim, J. Park, and J. Ryu, “Intelligent Surveillance and Anomaly Behavior Detection Based on Multimodal Data,” in *Proc. IEMEK Symp. Embedded Technol. (ISET)*, pp. 93–95, 2025.

[5] D. Kim, K. Seo, and C. M. Yang, “An Action Recognition Method for Risk Situation Recognition in Various Physical Security Applications,” in *Proc. Joint Conf. Commun. Inf. (JCCI)*, pp. 398–399, 2025.

[6] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6450–6459, 2018.