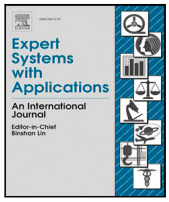
Expert Systems With Applications 254 (2024) 124391

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Review

PASS-CCTV: Proactive Anomaly surveillance system for CCTV footage analysis in adverse environmental conditions

Hobeom Jeon [a], Hyungmin Kim [a], Dohyung Kim [a,b,*], Jeahong Kim [b]

[a] University of Science and Technology, Deajeon, Republic of Korea
[b] Electronics and Telecommunications Research Institute, Deajeon, Republic of Korea

ARTICLE INFO

Keywords:
Video surveillance system
Video anomaly detection
Human tracking
Vision–language model

ABSTRACT

In recent decades, the growing deployment of Closed-Circuit Television (CCTV) systems for crime prevention and facility security has accelerated the importance of intelligent surveillance technologies. One of the primary challenges in this field includes varying viewpoints and adverse weather conditions that significantly compromise the accuracy of human tracking and anomaly detection. Moreover, conventional surveillance systems often focus only on specific events within limited scenarios, which restricts their applicability. Existing deep learning approaches also face limitations in adaptability to environmental variations, mainly due to the high maintenance costs involved in data collection. To address these challenges, we present a comprehensive surveillance system that utilizes deep learning to enhance human tracking and anomaly detection across diverse environments. Our approach includes the implementation of novel object filtering algorithms that decrease false positive rates and improve tracking precision. Additionally, our system is capable of monitoring multiple types of abnormal events, such as intrusion, loitering, abandonment, and arson. We further introduce a prompt-based recognition mechanism that enables active user participation in identifying abnormal scenes. Extensive evaluations using the Korea Internet & Security Agency CCTV datasets have demonstrated significant performance enhancements by our system, particularly under challenging weather conditions. Moreover, our system achieved competitive accuracy on the ABODA and FireNet datasets, even without additional training. This research establishes a new baseline for practical surveillance solutions that focus on comprehensive monitoring across various abnormal scenarios.

1. Introduction

Closed-circuit television (CCTV) systems have become increasingly popular for public safety and crime mitigation. These systems are valuable tools for monitoring and recording activities in public spaces, transportation hubs, and various other locations. However, the growth in the number of CCTV cameras raises concern about operator potential vigilance decrements (Donald, Donald, & Thatcher, 2015). Moreover, traditional CCTV systems often rely on human operators to identify abnormal or suspicious activities (Donald, 2019). This manual monitoring process can be labor-intensive and difficult to scale. As a result, there is a growing need for proactive surveillance systems that can automate the analysis of video data, detect abnormal situations, and alert security personnel in real-time.

Recent advancements in artificial intelligence (AI) and deep learning (DL) offer a promising solution, paving the way for the development of intelligent surveillance systems (Gan, Fernando, & Molina-Solana, 2021; Lee, Leong, Lai, Leow, & Yap, 2018; Neupane et al., 2023). By integrating computer vision algorithms, these systems are equipped to analyze video streams and detect unusual activities or patterns automatically. This capability of AI-based CCTV surveillance systems enhances their ability to identify and retrieve information about abnormal situations (Mabrouk & Zagrouba, 2018).

Despite these advancements, DL-based visual surveillance systems encounter three primary challenges. First, inaccuracies in pedestrian localization by the human tracking module significantly impair the system's ability to detect abnormal events. Second, although there is extensive DL-based surveillance research focused on single-event detection methods, the simultaneous detection of multiple events is still largely underexplored. Third, most DL-based models necessitate fine-tuning in response to environmental changes to sustain system performance post-deployment.

First, human tracking is crucial for detecting individual abnormalities, with various actor-based approaches developed (Shah, Karlsen, Solberg, & Hameed, 2023). However, CCTV environments face challenges from adverse weather conditions, such as heavy haze and snowstorms, and low illumination at dusk and night, which reduce the

* Corresponding author at: Electronics and Telecommunications Research Institute, Deajeon, Republic of Korea.
E-mail addresses: tiger@etri.re.kr (H. Jeon), khm159@etri.re.kr (H. Kim), dhkim008@etri.re.kr (D. Kim), jhkim504@etri.re.kr (J. Kim).

https://doi.org/10.1016/j.eswa.2024.124391
Received 15 March 2024; Received in revised form 16 May 2024; Accepted 1 June 2024
Available online 4 June 2024

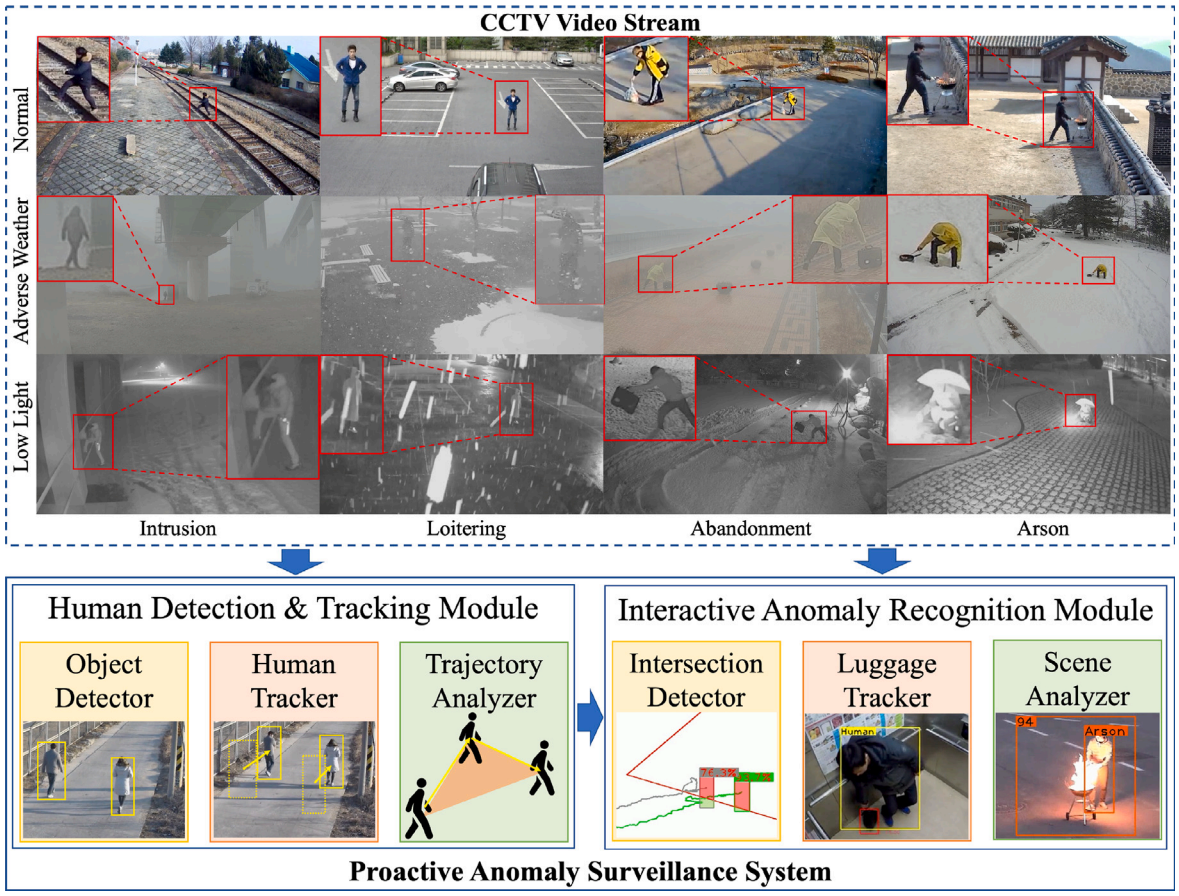


Fig. 1. Overview of the proactive anomaly surveillance system. The proposed system detects four critical scenarios — intrusion, loitering, abandonment, and arson, under challenging conditions such as adverse weather and low-light environments. The system integrates two main modules: (1) the human detection and tracking module utilizing feature coupling and object filtering for accurate anomaly identification, and (2) the interactive anomaly recognition module, which encompasses an intersection detector, a personal luggage tracker, and a hazardous scene analyzer.

reliability of pedestrian tracking (Li, Yang, & Qu, 2019). Errors in human trackers, such as treating non-human objects as humans or misclassifying humans as other objects, severely degrade the event detection performance of the system. Therefore, stable human tracking in various CCTV environments is essential to maintaining the performance of the intelligent surveillance system.

Second, the development of a unified framework capable of simultaneously detecting multiple events is crucial for proactive surveillance systems (Caruccio, Polese, Tortora, & Iannone, 2019; Lim, Tang, & Chan, 2014). Recent DL-based anomaly detection research has covered various scenarios such as fire, violence, and fallen person (Hashemzadeh & Zademehdi, 2019; Lee, Lee, Seong, Hyun, & Kim, 2023; Mahmoodi & Salajeghe, 2019). However, these studies predominantly focus on single-event detection without consideration of the interplay between different methods within the unified system. While detection methods tailored to specific scenarios can be effective, they often lack the flexibility and generality required for broader applications, resulting in underutilization across diverse environments (Lim et al., 2014). Therefore, advancing practical surveillance research demands a comprehensive framework that seamlessly integrates multiple methods for detecting diverse scenarios concurrently.

Lastly, most anomaly detection studies using DL models are limited to training and performance evaluation on pre-collected datasets. However, methods that utilize pre-trained DL models exhibit limited adaptability, particularly in previously unseen or untrained environments. To ensure consistent performance across the target domain, these anomaly detection models require continuous fine-tuning with additional data. Consequently, these training-based approaches lead to higher maintenance costs for collecting data and greater reliance on expert knowledge for re-training. Recently, Doshi and Yilmaz (2022) have adapted the continuous learning approach for anomaly detection in CCTV footage to mitigate maintenance costs. While this method effectively reduces false alarms at a lower cost, it still requires further data collection and training processes in the target domain for the analysis of complex object interactions. Therefore, developing sustainable and robust strategies remains a critical need in visual surveillance research.

To address these challenges, we propose a unified surveillance system designed to detect and alert various abnormal situations through a DL approach in CCTV environments. As illustrated in Fig. 1, the system processes video feeds across various adverse environmental conditions and generates alerts upon detecting abnormalities. It specifically targets abnormal situations caused by intentional human behaviors, such as intrusion, loitering, abandonment, and arson. Our system employs a two-stage pipeline for the effective monitoring of human-related events: the Human Detection and Tracking Module (Section 3.1), which identifies and tracks individuals within the video feed, and the Interactive Anomaly Recognition Module (Section 3.2), which analyzes behaviors for anomaly detection.

In our Human Detection and Tracking Module, we have implemented a feature coupling method along with object filtering algorithms to enhance system performance. The feature coupling approach integrates human appearance features with object-discriminate features, ensuring reliable tracking across diverse conditions. Simultaneously, our object filtering algorithm analyzes human trajectory movement at the initial entrance of humans in the field of CCTV. This

strategy minimizes false positives in human detection and improves the accuracy of detecting abnormalities.

The Interactive Anomaly Recognition Module successfully detects multiple hazards through an integrated framework, consisting of three specialized components: the Intersection Detector, Luggage Tracker, and Scene Analyzer. Each component is tailored to address a broad spectrum of visual surveillance needs, from the identification of unauthorized human presence to the tracking of personal items and the analysis of behavioral patterns and their effects on the surroundings. This study provides guidelines for the development of proactive video monitoring systems designed for general-purpose operation, enhancing the ability to recognize multiple scenarios simultaneously.

For sustainable surveillance solutions, we introduce a novel prompting interface that enables modification of text description through user intervention. This advancement allows security personnel to customize the system according to specific operational needs. Moreover, we have adapted patch processing for more accurate scene and behavior analysis, thereby improving the system's understanding of human intentions and their surroundings. Our evaluation demonstrates the effectiveness of this prompt-based approach in semantically analyzing frames with predefined abnormal text representations. This flexible approach addresses various scenarios not only enables rapid response to new events but also eliminates the need for model retraining.

The performance of our system was evaluated on the Korea Internet & Security Agency (KISA) CCTV dataset (KISA, 2017), which is purposefully designed for certification for commercial surveillance systems. The KISA CCTV dataset includes highly realistic environments for comprehensive system evaluation and has recently been enhanced to involve additional hazy weather and low-light conditions. Our experimental results have demonstrated significant improvements in system performance compared to previous research (Kim, Kim, Mok, & Paik, 2021b), especially under more challenging conditions. For comparative analysis, experiments were also conducted on the ABODA (Lin, Chen, Chen, Lin, & Hung, 2015) and the FireNet dataset (Jadon, Omama, Varshney, Ansari, & Sharma, 2019). Our system exhibits superior performance in detecting abandonment and fire scenarios, notably without requiring tailored training for each scenario.

The main contributions of our work are as follows:

- We propose the human detection and tracking module for robust performance across a wide range of environmental conditions, including diverse viewpoints, adverse weather, and low-light scenarios. Notably, it significantly reduces false positives in human tracking, thereby dramatically enhancing the accuracy of the proactive surveillance system.
- Our interactive anomaly recognition module can concurrently detect multiple abnormal situations including intrusion, loitering, abandonment, and arson. The proposed system recognizes activities ranging from human status to environmental interactions and operates effectively in real-world CCTV settings.
- Our prompt-based approach greatly enhances the flexibility and effectiveness of the surveillance system. By eliminating the need for additional training, this method reduces maintenance costs and improves the adaptability to target environments.

2. Related work

2.1. Human detection and tracking

In numerous surveillance systems, e.g. for intrusion detection and loitering prediction in the outdoors, it is essential to track each identity. Moreover, the availability of spatio-temporal information about human locations is crucial for recognizing abnormal situations caused by humans. While the integration of deep learning into human tracking has led to significant advancements, certain challenges remain when applying these technologies to surveillance systems. This paper revisits human tracking from the perspective of the surveillance solution.

2.1.1. Multiple Object Tracking

Multi-Object Tracking (MOT) research aims to extend object detection (Jocher et al., 2022; Redmon, Divvala, Girshick, & Farhadi, 2016; Ren, He, Girshick, & Sun, 2015) from images to videos by connecting detected objects in each frame to accurately track target objects. This field has evolved through competitions and challenges, including human tracking, vehicle tracking, and multi-object tracking tasks (Voigtlaender et al., 2019). Research in MOT has been actively conducted in human domains, such as analyzing pedestrian movement in crowded CCTV (Chen et al., 2019), human tracking for autonomous driving in vehicle (Galvão & Huda, 2023).

The Tracking-by-Detection paradigm (Bochinski, Eiselein, & Sikora, 2017), which allows object tracking using only location information based on high-speed object detection, has constituted the most efficient method in multi-object tracking. Among the widely used approaches, the SORT(Simple Online and Realtime Tracking) algorithm (Bewley, Ge, Ott, Ramos, & Upcroft, 2016), uses the Intersection over Union (IoU) method to calculate bounding box similarity and utilizes the Kalman filter and Hungarian algorithm for object tracking. Similarly, motion-based tracking approaches focus on efficiently handling detected position information to maintain accurate tracking even in situations with object overlap or occlusion (Cao, Pang, Weng, Khirodkar, & Kitani, 2023). However, since these methods solely rely on positional information and do not utilize the visual appearance of objects, Deep SORT (Wojke, Bewley, & Paulus, 2017) introduces deep-learning that uses a separate CNN to extract object appearance features and perform matching based on similarity instead of IoU. Subsequent research has proposed methods like FairMOT (Zhang, Wang, Wang, Zeng, & Liu, 2021), which utilizes low-level features from object detection networks, and Bottom-Up based GSDT (Wang, Kitani, & Weng, 2021) and RelationTrack (Yu, Li, Han, & Wang, 2022) based on decoupling field, which perform object detection and tracking simultaneously.

Human tracking in surveillance systems still faces fundamental challenges in real-world environments. Object detectors often prioritize positional accuracy over reducing false positives due to mean Average Precision (mAP) measurements (Bolya, Foley, Hays, & Hoffman, 2020). Consequently, these detectors sometimes fail to accurately classify objects, leading to misidentifying non-human objects as humans. While erroneously tracking non-humans as real humans, the surveillance system analyzes the non-humans to detect abnormal situations. This fundamental issue of incorrect object detection primarily contributes to unnecessary false alarms and increases in computational costs. To tackle this issue, we introduce an object filtering method designed to exclude these false positive human results. By providing highly reliable human object identification, our approach significantly improves the efficiency and performance of anomaly detection systems.

2.1.2. Person re-identification

Person re-identification (re-ID) is a pivotal area of research in intelligent video surveillance, especially significant in multi-camera environments with non-overlapping surveillance zones. The objective of re-ID technology is to precisely match individuals captured on different cameras by comparing them against a pre-constructed gallery of previously recorded individuals. The re-ID process involves tracking all persons visible in CCTV footage to build the gallery set and matching by similarity between query and gallery (Ye et al., 2021). The accuracy of re-ID is evaluated using the Cumulative Matching Characteristics metric, which measures the precision of correctly ranking the same individual in search results.

Addressing the challenges of accurately identifying individuals in multi-camera surveillance systems is complex due to variations in capturing angles and other environmental factors. Several specific issues contribute to these difficulties: low resolution (Wang et al., 2018), lighting changes (Karanam, Li, & Radke, 2015), occlusions and truncations (Wang et al., 2018), cluttered backgrounds (Song, Huang,

Ouyang, & Wang, 2018), and bounding box inaccuracies (Martinel, Das, Micheloni, & Roy-Chowdhury, 2016).

Most re-identification (re-ID) evaluation metrics adhere to a closed-world assumption, where all pedestrians are expected to be present within the gallery set. In contrast, real-world applications require the simultaneous human tracking during the integration of individuals into the gallery set. This paper utilizes the re-ID model (Zhou, Yang, Cavallaro, & Xiang, 2019, 2021) to assess the similarity between individuals captured in previous frames and persons detected in the current frame. To achieve precise tracking within the same video, we introduce the feature coupling method, which integrates information extracted from detectors. Our proposed method enables the robust extraction of identity vectors for individuals under diverse environmental conditions.

2.2. Abnormal event detection

Our objective is the concurrent recognition of multiple events by amalgamating diverse detection methods. This section provides the overview of current research trends in detecting specific events, including intrusion, loitering, abandonment, and arson.

2.2.1. Intrusion and loitering detection

Advances in object detection and tracking technologies have significantly enhanced the capability to analyze human trajectories in CCTV environments. For instance, the tripwire-based intrusion detection method (Chen, Chen, Yuan, & Kuo, 2016), which enables users to define a virtual "wire" in video footage, facilitates intuitive alarm generation. This approach simplifies the intrusion detection process using simultaneous quadratic equations, making it suitable for integration within camera-embedded software. The method is widely supported in numerous commercial surveillance systems due to its simplicity and efficiency. Moreover, Kim et al. (2021b) further extends the diversity of surveillance methods by leveraging polygonal boundary regions for intrusion detection.

Effective identification of loitering through manual observation requires surveillance operators to recall individuals' appearances and monitor video footage over periods. Therefore, surveillance systems require the capability to assist in the identification of human loitering by tracking the same individuals over extended durations and detecting suspicious individuals who continue to remain in a particular area. Several methods have been proposed for modeling the loitering attributes of suspected individuals, such as trajectory probability distribution-based approaches (Kang & Kwak, 2014), user analysis of activity areas (Höferlin, Höferlin, Weiskopf, & Heidemann, 2011), and classification of trajectory activity areas (Huang et al., 2019). Although intrusion detection in surveillance systems is typically straightforward, the analysis of loitering behavior presents a unique challenge due to the diverse interpretations of 'loitering.' Various definitions of loitering behavior lead to inconsistent standards for performance evaluation, complicating the comparison of methods and impeding further advancements.

This study adopted KISA CCTV criteria to analyze intrusion and loitering detection performance improvement. We introduce an intersection ratio function to enhance the accurate recognition intruders in the alert area. By integrating advanced human tracking and the intersection ratio function, we have achieved significant improvements in the precision of intrusion and loitering detection. Our system allows operators to designate specific areas for intrusion and loitering surveillance, providing high-performance proactive surveillance.

2.2.2. Personal luggage abandonment detection

Detecting abandoned objects is critical for preventing unauthorized littering and identifying potential security threats, such as explosives. Traditional approaches have utilized background subtraction techniques to identify stationary objects that newly appear in the video (Zivkovic, 2004). These methods involve comparing current and previous frames to distinguish the foreground from the background. Advancing this approach, dual background subtraction techniques have been developed, combining Short-term and Long-term Masks to more accurately identify changes in the foreground (Lin et al., 2015; Wahyono, Filonenko, & Jo, 2016). These background subtraction techniques effectively designate certain foreground pixels as static objects that have newly appeared.

The integration of object detection models with foreground information, as proposed by Park, Park, and Joo (2019) and Shyam, Kot, and Athalye (2018), focuses on reducing the impact of lighting changes on detection accuracy. Further, HLDNet (Kim, Kim, Mok, & Paik, 2021a) presents an advanced method, utilizing human pose estimation to link objects to their probable owners, specifically near the wrist area. Despite its advancements, this method faces computational challenges due to the complexity of pose estimation. Moreover, while Liao, Yang, Ying Yang, and Rosenhahn (2017) proposed a theft detection method utilizing ownership information, it was validated only in three video sequences.

Our research presents a top-down approach to tracking luggage and identifying the moment of abandonment. Our method detects abandoned objects without foreground information, thereby enhancing its robustness against lighting variations and its applicability in outdoor settings. Notably, our top-down approach enables tracking even small baggage and facilitates faster detection of abandonment by eliminating the latency associated with generating foreground information.

2.2.3. Arson detection

Our objective is to identify arson activities for effective early fire detection. Arson detection requires not only detecting the early signs of a fire but also recognizing behaviors that indicate intentional ignition. However, to the best of our knowledge, no research focuses on identifying arsonists and recognizing ignition in fire prevention and surveillance.

Predominant fire detection studies have concentrated on the classification and recognition of fire instances, primarily within wildfire contexts. Initial investigations in this field applied image classification to detect fire presence (Vipin, 2012), evolving to more sophisticated approaches that extract fire texture features from images for a refined identification of fire regions (Chino, Avalhais, Rodrigues, & Traina, 2015; Yuan, Liu, & Zhang, 2015). The integration of DL paradigms has marked a significant leap forward, offering refined algorithms capable of discerning fire regions with high precision in both still images and video streams (Cao, Yang, Tang, & Lu, 2019; Jadon et al., 2019; Park & Ko, 2020; Wu, Xue, & Li, 2022). These advancements have been instrumental in wildfire management, aiding in the precise localization of extensive fire outbreaks for targeted firefighting interventions.

Nevertheless, the aforementioned methods are confined to detecting well-established fires, with limited utility in the early stages of fire development or the identification of arson perpetrators. Inception of fires, particularly captured in outdoor surveillance, flames tend to be small and elude early detection. This delineates a critical research gap between detecting fires and identifying arsonists. To bridge this gap, our approach adopts a holistic scene analysis strategy, with a specialized focus on behaviors indicative of arson. By examining individual actions within the scene, our framework not only aims to identify nascent fires but also to pinpoint the arsonists responsible for igniting fires. This approach enables the proactive response to fire incidents by detecting arson behaviors at flame onset and identifying the culprits, thereby significantly enhancing the efficacy of surveillance systems.

3. Proactive anomaly surveillance system

3.1. Human tracking and trajectory analysis

In the Human Detection and Tracking Module of our proactive surveillance system, we integrate object detection, human tracking, and

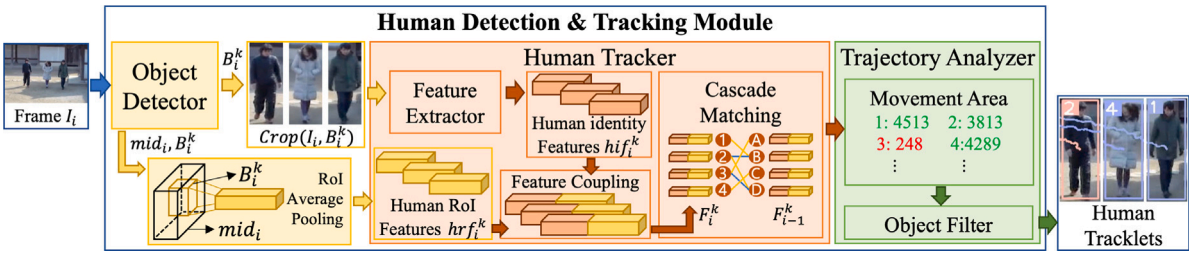


Fig. 2. Operational diagram of the human detection and tracking module. This module processes incoming frames and undertakes three primary tasks: object detection, human tracking, and trajectory analysis. The object detector localizes human positions in 2D coordinates. The human tracker, utilizing our feature coupling method, combines identity and RoI features to manage each human's tracklet. The trajectory analyzer filters out false positive humans by analyzing their movement patterns.

person re-identification to ensure precise individual monitoring. As depicted in Fig. 2, the operational diagram of the module comprises three core components: an object detector, a human tracker, and a trajectory analyzer. We introduce a feature coupling method that combines identity vectors from both a re-ID model and the object detector for robust human tracking. Moreover, our trajectory analyzer leverages polygon area calculation to remove false positives and improve abnormal event detection efficiency.

3.1.1. Augmenting identity feature by Object Detector

For the initial step, the *Object Detector*, positioned at the leftmost part of the operational diagram in Fig. 2, retrieves objects classified as humans within the video frames and relays their location information to the Human Tracker. The operation of the *Object Detector* is formalized as follows:

$$B_i^k = [(x_{tl}, y_{tl}), (x_{br}, y_{br})] \in ([0, W], [0, H]) \tag{1}$$

The variable B_i^k denotes the location of the kth pedestrian within the ith frame, represented as a bounding box (Bbox). This bounding box is defined within the constraints of the frame's maximum dimensions, $W \times H$, and includes two pairs of coordinates: the top-left (tl) and bottom-right (br) corners. To achieve consistent identification across frames, the *Human Tracker*, positioned centrally in Fig. 2 and receiving input directly from the *Object Detector*, utilizes the deep association metric (Wojke et al., 2017). This method leverages the *Feature Extractor*, located within the *Tracker* as a model to extract deep features, to produce the human identification vector, hif, from these B_i^k regions:

$$hif_i^k = Extractor(Crop(I_i, B_i^k)) \in \mathbb{R}^{(1\times dim)} \tag{2}$$

We implement the re-identification model OSNet (Zhou et al., 2021) to serve as the *Feature Extractor* within our *Human Tracker*. However, performance degradation is observed in top-view camera configurations, attributed to the scarcity of overhead perspectives in the training datasets. Additionally, the identity representation space of the *Feature Extractor* is restricted during inclement weather or low-light scenes due to the uniform visual appearances. These adverse conditions compromise the model's effectiveness, often leading to the misidentification of distinct individuals and the degradation of overall system performance.

Our system introduces the *Feature Coupling* method for reliable tracking in adverse environmental conditions. This method, functioning as a key component of the *Tracker* and depicted in Fig. 2, combines hif_i^k from the *Feature Extractor* with visual patterns derived from the *Object Detector*. Drawing inspiration from the mid-level deep pattern mining (Li, Liu, Shen, & Hengel, 2017), the *Tracker* generates discriminative features from the mid-level representations provided by the *Object Detector*. Extensive training on datasets, enriched by a diversity of backgrounds and object shapes, empowers object detection models to identify and recognize a wide variety of objects. Therefore, the *Detector* can analyze spatial and textural data to produce discriminative features essential for precise object localization and identification. Our tracking module extracts intermediate-level representation from the *Object Detector* to augment feature vectors for human tracking. In this extraction process, we apply Region of Interest (RoI) average pooling (Girshick, 2015) to the mid-level results for extracting individual features. This repurposing approach enables the extraction of *Human RoI Features*, labeled as hrf, from each identified human location B_i^k in the mid-level representation mid_i:

$$hrf_i^k = RoIPooling(mid_i, B_i^k) \in \mathbb{R}^{(1\times dim)} \tag{3}$$

For the detail, the hrf_i^k is extracted into a 256-dimension vector through pooling from the 20×16 resolution space of mid-level representation from the *Object Detector*. Subsequently, the Tracker concatenate hif_i^k and hrf_i^k to create the comprehensive pedestrian tracking vector F_i^k:

$$F_i^k = Concat(hif_i^k, hrf_i^k) \in \mathbb{R}^{(1\times dim)} \tag{4}$$

Our *Feature Coupling* method constructs the tracking vector F_i^k with 512-dimension by concatenating the identity feature vector hif_i^k and the human region feature vector hrf_i^k. Finally, the tracking vectors F_i^k connect with the preceding vectors F_{i-1}^k by the *Cascade Matching* algorithm (Wojke et al., 2017) to assign identities. The *Object Detector*, trained on larger datasets than the re-ID model, can provide another perspective on visual appearance. This advantage allows our *Feature Coupling* method to enhance stability in human tracking under various viewpoints, lighting, and weather conditions. Furthermore, this method minimizes computational demand by reusing results previously generated by the *Object Detector*. This strategy enhances the human tracking efficiency of the system to manage real-time human tracking with minimal computational overhead.

3.1.2. False positive filter using trajectory movement

Human False Positive (HFP) occurs when detection and tracking methods misidentify non-human objects as humans. In surveillance systems, the accuracy and computational efficiency of identifying abnormal situations are significantly compromised by HFPs. Therefore, we introduce a reliable *Trajectory Analyzer* to eliminate HFPs, as shown in the rightmost module of Fig. 2.

In this paper, we first assume that actual humans must present movement over an initial period to enter the field of surveillance view. Accordingly, we assess the amount of movement over some duration after the onset of tracking to identify HFPs. However, the following formulas commonly used for measuring distance cannot be applied to the estimated trajectories.

$$\mathcal{T}_k^{tl} = B_0^k, B_1^k, \ldots, B_n^k \in (x_{tl}, y_{tl}) \tag{5}$$

$$Length(\mathcal{T}_k^{tl}) = \sum_{i=1}^{n} \| \overrightarrow{B_{i-1}^k, B_i^k} \| \tag{6}$$

$$Displacement(\mathcal{T}_k^{tl}, t) = \| \overrightarrow{B_{n-t}^k, B_n^k} \| \tag{7}$$

The $\mathcal{T}_k^{tl}$ in Eq. (5) represents trajectory information that extracts the top-left coordinates of tracked humans. The Length function in Eq. (6) calculates the total movement distance of $\mathcal{T}_k^{tl}$ using Euclidean distances between positions in successive frames. Eq. (7) computes the displacement as the magnitude of the movement vector over time t.

Fig. 3. Visualization of Human False Positive results. The bounding boxes and trajectories illustrate the observations from the human tracker, including instances where objects are misidentified as humans. These Human False Positives frequently occur in adverse environments such as low lighting as shown on the left and under severe snowstorm conditions on the right. The trajectories of incorrect instances demonstrate rapid and repetitive movements across long distances in random directions. Best viewed in color (Human: green bbox, False positive: red bbox, Trajectories: yellow lines).

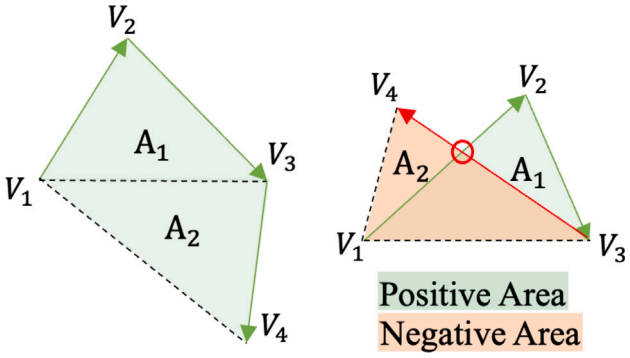


Fig. 4. Calculating areas in trajectory polygons. On the left, vertices V_1 to V_4 define a simple polygon (i.e. non-overlapping polygon), with the total area as the sum of A_1 and A_2. On the right, the polygon demonstrates self-overlapping, with $V_1 \rightarrow V_2$ crossing $V_3 \rightarrow V_4$. This intersection reduces the total area by subtracting the 'Negative Area' from the 'Positive Area' segments.

The overall tracked objects demonstrate continuous noise movement influenced by the environmental factors of outdoor CCTV. These conditions contribute to the challenges in accurately calculating the movement of objects or individuals. We observed that HFPs are primarily characterized by two distinctive movement patterns: continuous oscillation in place, and repetitive crossing movements caused by ID switches. The characteristic movements of High False Positives (HFP), as exemplified in Fig. 3, demonstrate that length and displacement metrics frequently assess substantial movement to objects despite no actual locomotion. This discrepancy is largely due to the inability of these common distance functions for quantifying motion to differentiate between genuine spatial transitions and stationary oscillations or ID-switching, leading to an overestimation of movement metrics. To exclude noise movements in measurement, our human trajectory analyzer utilizes polygonal area calculations as follows:

$$Area(Polygon) = \frac{1}{2}\sum_{i=1}^{n} x_i(y_{i+1} - y_{i-1}) \tag{8}$$

For the precise movement evaluation, we utilize Gauss's area calculation formula in Eq. (8). This formula is valid for simple polygons in constructive solid geometry. Therefore, the vertices of the polygon must be arranged clockwise or counterclockwise for the computation to be accurate. When processing self-overlapping polygon through Eq. (8), intersections within the trajectories introduce negative area. In Fig. 4, we visually contrast simple and self-overlapping polygons. Building on the area function, we conceptualize the object tracklet as a polygon to calculate the movement. By considering trajectories as polygons, we can subtract overlapping areas caused by intersecting vectors, and this metric is formulated in Eq. (9).

$$ATM(\mathcal{T}_k^{tl}) = \frac{1}{2} \cdot \left| \sum_{i=1}^{n} \mathcal{T}_k^{tl} \times \sigma(\mathcal{T}_k^{tl}) \right| \tag{9}$$

The Area of Trajectory Movement (ATM) function calculates movement by applying the cross product between trajectory $\mathcal{T}_k^{tl}$ and its circular shifted version $\sigma(\mathcal{T}_k^{tl})$. This process reduce movement areas by deducting overlaps in self-intersecting trajectories. Consequently, trajectories with significant self-overlap are calculated as small areas in ATM calculations despite extensive travel distances. The *Trajectory Analyzer* distinguishes between actual human movements and HFPs by using ATM. This trajectory area approach allows for effective HFP filtration without relying on length and displacement metrics. The *Object Filter* method within the *Trajectory Analyzer* classifies reliable human objects by comparing movement areas against a minimum threshold during the initial period. Our filtering method significantly enhances the performance of abnormal recognition by eliminating incorrectly identified human objects.

3.2. Interactive Anomaly Recognition

The Interactive Anomaly Recognition Module comprises three integral components: the Intersection Detector, Luggage Tracker, and Scene Analyzer to enhance the detection of multiple abnormal events. The Intersection Detector observes individuals implicated in intrusion or loitering within designated alert zones. The Luggage Tracker identifies abandoned luggage by tracking smaller items and associating them with their owners. Leveraging vision–language techniques, the Scene Analyzer detects arson incidents by interpreting textual information about fire. All three components work together to identify abnormal situations such as intrusion, loitering, abandonment, and arson.

Each component of the Anomaly Recognition Module is designed to interact with security personnel through the user-control panel to enhance detection capabilities. In the Intersection Detector, operators can define and modify the shapes of polygonal alert zones tailored to the geography of the monitored area, which allows for precise calibration of intrusion or loitering detection parameters. For the Luggage Tracker, operators can specify categories of interest objects, enabling the system to concentrate on relevant items based on the location for enhanced surveillance of potential abandonment scenarios. Lastly, the Scene Analyzer utilizes customizable text prompts that personnel can adapt to target specific anomalies, such as identifying potential arson by textual cues related to fire and smoke within the surveillance footage. These interactive features empower the operators and ensure that the system's response is finely tuned to the immediate security requirements.

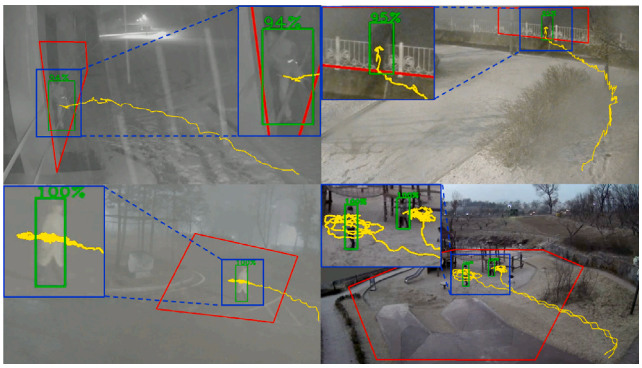


Fig. 5. Intersection detection in various outdoor settings: Alert zones are outlined in red with individual trajectories in yellow. The examples demonstrate the system's capability of intrusion and loitering detection in fog, rain, and low-light conditions.

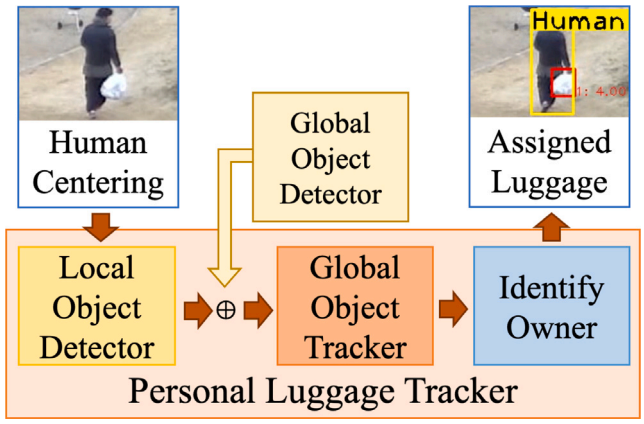


Fig. 6. Operational diagram of the personal luggage tracker. The top-down tracking approach dynamically detects personal luggage by combining global detection with localized individual detectors and identifies the owner through bipartite matching.

3.2.1. Alert zone intersection detection

The Intersection Detector evaluates positions of tracked human to identify instances of intrusion and loitering within designated polygonal alert zones. The detector assesses the intersection rate to quantify the degree of overlap between the human bbox and the alert area. For segmenting the overlapping polygon between the human bbox and the alert zone, we utilize the polygon clipping algorithm (Sutherland & Hodgman, 1974). The intersection ratio is defined by Eq. (10):

$$Intersection(B_i^k, Z) = \frac{Area(B_i^k \cap Z)}{Area(B_i^k)} \tag{10}$$

This formulation presents the intersection ratio as the quotient of the overlap area ($Area(B_i^k \cap Z)$) relative to the total area of the human bbox ($Area(B_i^k)$). This quantification enables the Intersection Detector to accurately determine the levels of intrusion.

$$Intersection(B_i^k, Z_{intrusion}) > \theta_{intrusion} \tag{11}$$

$$\begin{aligned} &Intersection(B_{i-t}^k, Z_{loitering}) > \theta_{loitering} \\ \&\quad &Intersection(B_i^k, Z_{loitering}) > \theta_{loitering} \end{aligned} \tag{12}$$

The intersection ratio function calculates the ratio of overlap with the intrusion zone ($Z_{intrusion}$) for each bbox (B_i^k). An intrusion event is detected when any individual ratio surpasses the threshold value ($\theta_{intrusion}$), as delineated in Eq. (11). Extending this principle to loitering detection, the function evaluates the duration an individual remains within the loitering zone ($Z_{loitering}$) against a predefined duration threshold (10 s in the KISA dataset), as formulated in Eq. (12). Unlike the previous approach in Kim et al. (2021b) that relied on object coordinate detection within the intrusion zone, our approach leverages the advanced intersection ratio method. This method enhances the precision in determining intrusion times by quantitatively assessing the extent of each entry into designated alert areas.

Fig. 5 graphically illustrates the process of identifying intrusion and loitering events in the application of intersection detection. Red alert zones, designated by control personnel, are prominently displayed within the visual frame, facilitating targeted monitoring of specific regions of interest. This strategic approach significantly enhances the system's responsiveness to potential intrusions and loitering incidents, thereby augmenting the overall efficacy of the proactive surveillance system.

3.2.2. Personal luggage abandonment detection

Detecting abandoned luggage in CCTV environments requires tracking human trajectories and identifying their luggage. The detection of small objects, such as handbags and plastic bags, poses challenges at distances due to the minimum anchor size restrictions of object detectors. To enhance the detection for small and distant objects, each camera location demands customized datasets and models. However, the optimization for each object size is resource-intensive and degrades performance for other sizes.

To overcome these challenges, we propose a top-down luggage tracking method that dynamically identifies abandoned luggage from any distance. This top-down method adapts to the scale and distance of humans and eliminates the need for extensive datasets tailored to luggage size. In the Human-Centering step, the re-acquire individual-centered images from the current frame. Illustrated in Fig. 6, the Luggage Tracker combines the Global Object Detector with localized detection to merge duplicate instances by proximate individuals. Objects detected by top-down are consolidated into globally detected items by the inverse matrix and eliminating duplicates using non-maximum suppression (NMS). Our top-down approach allows the detection of small luggage items that remain unnoticed by the global detector, ensuring thorough surveillance regardless of object size or human distance.

For owner identification, our system utilizes bipartite matching and the Intersection over Union (IoU) metric by measuring the spatial relationship between luggage and potential owners in a video frame. This approach, contrasting with background subtraction methods (Krusch, Bochinski, Eiselein, & Sikora, 2017), detects abandonment by tracking the presence or absence of belongings. The capability to identify belongings offers a foundational approach for detailed surveillance that analyzes various attributes from individuals. This feature enables applications such as determining the theft of possessions or the carriage of prohibited items and enhancing security measures by monitoring the movement and ownership of items.

The Luggage Tracker utilizes a dynamic threshold approach to determine luggage abandonment, calculating the real-time spatial distance between the object and its owner. The dynamic threshold adjusts based on the distance of individuals from the camera, using twice the width of the human's Bbox. This adaptability is crucial for responding to the diverse range of distances captured by CCTV cameras. In cases where an individual is near the edge of the camera's view, potentially hiding some body parts, the system applies the lastly calculated threshold to preserve accuracy in its decisions. This approach ensures consistent detection of luggage abandonment throughout the entire camera coverage area. By integrating top-down object tracking with dynamic threshold, our Luggage Tracker addresses small and distant objects in real-world abandonment surveillance.

3.2.3. Text prompt based human arson detection

The proposed Text Prompt Scene Analyzer aims to comprehensively analyze fire scenes and arson activities within CCTV environments.

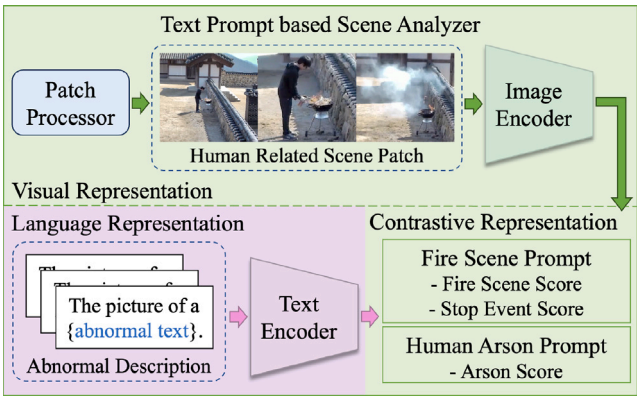


Fig. 7. Operation diagram of arson detection with the vision–language model. The text encoder converts user-provided text into the contrastive space, while the image encoder converts patches extracted from the CCTV frame into the same space. The contrastive representations are compared by cosine similarity, the analyzer can identify the scene of fire and arson.

By incorporating the Contrastive Language–Image Pre-training (CLIP) model from Radford et al. (2021), we have developed an analyzer capable of effectively recognizing scenes and actions. Our fire and arson detection method encompasses four key operations: patch collection, image and text encoding, image–text contrastive inference, and fire alarm generation. The overall schematic diagram of the proposed fire and arson detection method by text prompt is illustrated in Fig. 7.

Our approach to arson detection in CCTV environments leverages the CLIP model to overcome dataset scarcity, enabling the system to recognize fire and arson incidents in adverse environmental conditions. By integrating image and text encoders as the foundational backbone, the CLIP model achieves an alignment of feature vectors within a contrastive representation space. This alignment allows for zero-shot scene identification via image–text similarity, significantly broadening the detection capabilities. Our novel strategy enhances detection capabilities under diverse conditions, such as nighttime or adverse weather, by utilizing text descriptions of fire scenarios alongside normal conditions, thereby addressing the challenges of dynamic fire sizes and diverse background settings.

Fig. 8 presents the operation diagram of our patch processor, which is instrumental in collecting image patches for detailed scene analysis. Initially, the processor resizes the entire video frame to the input size of the image encoder. Subsequently, the processor uses the human tracklet to extract trajectory patches from the current and past positions. These trajectory patches are adjusted to input size by maintaining the aspect ratio. Additionally, the processor determines each individual's stationary status to isolate patches from areas where movement has ceased. The detection of stationary individuals involves calculating the IoU between their consecutive positions. Integration with the ATM, as detailed in Eq. (9), enables precise assessment of the human stationary state. To augment arson detection capabilities, the processor enlarges the region of the individual stop positions to 1.5 times the height of each person. The scene patches collected from areas of human presence are converted by the Image Encoder. This collective approach for multi-region reasoning permits detailed observation of both entire frames and specific areas where potential threats. Consequently, the analysis of patches from trajectory and stop region improves the early detection of fires and facilitates arson activity identification.

For arson detection, we introduce two types of prompts, as detailed in Table 1. These prompts include textual descriptions of both fire scenarios and human arson behaviors to enable comprehensive analysis across scene and action domains. Within the prompts, each description consists of text along with a corresponding risk score. In the arson detection process, the Analyzer calculates the average of the risk scores associated with the top three similarity scores for each patch. This

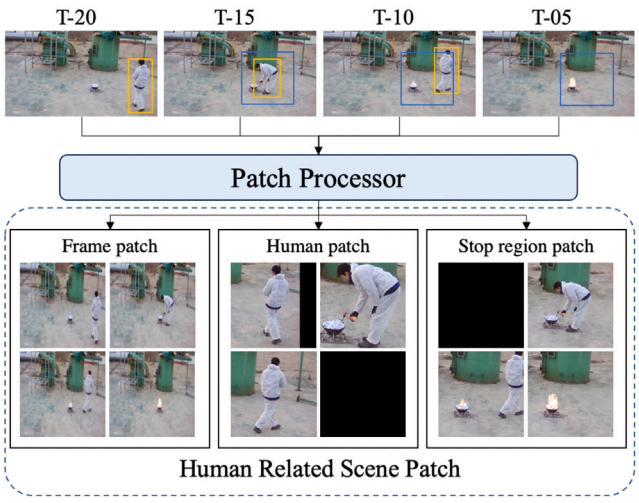


Fig. 8. Operation diagram of patch processor: Given an original video frame, the processor collects human detection result and analyze each person's stationary state. The example frames show the human bounding box and human stop region. The patch processor cropped the patches from each frame according to the information collected. we show the configured human-related scene patches at the bottom. Even if human patch cannot be cropped in the square, the processor maintains the aspect ratio by zero padding.

Table 1
Prompt about fire scene and human arson.

Fire scene prompt: There is {label}	
Non-fire descriptions	Fire descriptions
No person A moving person A person doesn't move	Smoke rising An arsonist Something that shines brightly A campfire Flame and fire soaring
Human arson prompt: A photo of {label}	
Non-fire descriptions	Fire descriptions
A standing person A walking person A bending person A squats person	An arsonist A man fighting a fire Flames and fires soaring A campfire A burning cooking pan A man doing a barbecue

approach leverages both the descriptive and quantitative aspects of the prompts to accurately assess and identify potential arson activities.

The Fire Scene Prompt and Arson Scene Prompt provide comprehensive perceptions for analyzing fire scenes and arson-related activities captured in CCTV footage. The Fire Scene Prompt utilizes both non-fire and fire descriptions to compare entire frames and specific stop-event patches, detecting normal and fire scenes. Meanwhile, the Arson Scene Prompt focuses on human behavior, utilizing arson-related textual descriptions and trajectory patches to infer specific actions, including those related to arson. Therefore, this dual perspective allows for the effective identification of potential fire incidents and arsonists through simultaneous analysis of scene context and individual actions. Moreover, our detection method offers significant scalability by requiring minimal computational resources to integrate additional text prompts for calculating cosine similarity. This approach of using multiple prompts enables the easy incorporation of new crime scene prompts according to user preferences, facilitating the customization of alarms to meet specific detection needs. This system design allows users to directly input their detection requirements into the system without the need for specialized expertise.

We propose prompt-based recognition as a foundational element for a user-friendly and explainable surveillance system. By leveraging textual descriptions for scene analysis, our approach not only enhances the system's adaptability to varied user requirements but also contributes to improving its accessibility and interpretability.

4. Experiments

This section presents a detailed evaluation of the proficiency of our system in real-world scenarios. Utilizing Real Time Streaming Protocol (RTSP) from a test streaming server, we aim to assess the performance in abnormal situation recognition and proactive event reporting.

4.1. Evaluation datasets

The Korea Internet & Security Agency (KISA) provides the CCTV dataset for the assessment of commercial software. Our study leverages the KISA CCTV dataset to verify system performance and conduct ablation studies on proposed methods. This dataset encompasses several scenarios, including Intrusion, Loitering, Abandonment, and Arson, segmented into four distinct subsets: Distribution, Validation, Certification, and Abroad. The Distribution subset is in an inclusive relationship of the Verification set and discloses for development. The Validation and Certification subsets, exclusively filmed within Korea, are confidentially maintained for certification examinations, whereas the Abroad subset, comprising videos recorded outside Korea, is publicly accessible. A comprehensive analysis and evaluation of each scenario within the KISA CCTV dataset utilized a total of 1765 videos. The detailed allocation of videos for each specific scenario is summarized in Table 2. All videos within the KISA dataset are captured at a resolution of 1280 × 720, ranging in duration from a minimum of 3 min to a maximum of 15 min. The recording perspectives were secured from a fixed camera at varying distances to actors: short (10~15 m), medium (15~20 m), and long (20~30 m), across six distinct time zones (sunrise, 9:00, noon, 15:00, sunset, and night), with half of the video captured under adverse weather conditions such as snow, rain, and fog. This comprehensive evaluation not only demonstrates the robustness and effectiveness of our system across varied environments but also validates the feasibility of our approach for integration into actual commercial systems.

The ABODA dataset (Lin et al., 2015) serves as a benchmark for evaluating our abandonment detection methodology against existing techniques. Illustrated in Fig. 9, the ABODA dataset encompasses 11 videos captured under various conditions, including indoor and outdoor environments, nighttime settings, and scenarios with lighting changes. This dataset enables a detailed evaluation of the proposed luggage tracking method, focusing on its capacity to detect abandoned objects. Through comparison with traditional background subtraction techniques, our method demonstrates superior early detection capabilities.

Similarly, the FireNet dataset (Jadon et al., 2019), depicted in Fig. 10, provides a basis for comparison with existing fire detection methods. This dataset integrates data from previous studies (Foggia, Saggese, & Vento, 2015; Sharma, Granmo, Goodwin, & Fidje, 2017) with images sourced from the Internet to ensure various backgrounds. The training set was constructed by random sampling of 1124 fire images and 1301 non-fire images from the previous datasets. The images in the test dataset consist of 593 fire images and 278 non-fire images extracted from 46 fire videos (19,094 frames) and 16 non-fire videos (6747 frames). Unlike traditional approaches, our arson detection method demonstrates the achievement of high performance without reliance on training data.

Table 2
Number of videos per subset on the KISA CCTV dataset.

Scenario	Distribution	Validation	Certification	Abroad
Intrusion	30	150	150	170
Loitering	30	150	150	325
Abandonment	10	50	50	400
Arson	10	50	50	70

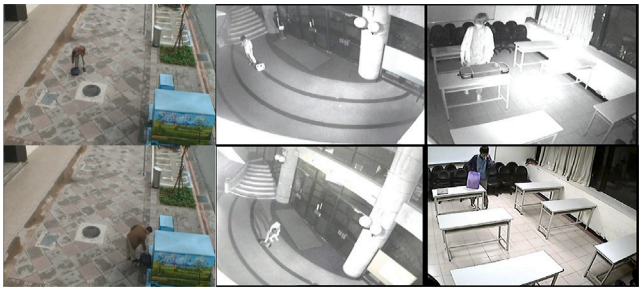


Fig. 9. Sample frames from ABODA dataset.

Fig. 10. Sample frames from FireNet dataset.

4.2. Evaluation metrics

The detection criteria for the KISA dataset are detailed in Table 3. These criteria establish a precise standard for identifying and alerting to anomalies By specifically defining the critical timing for accurate detection. An abnormal detection is registered as a true positive if it alarms within a 12-second window, starting 2 s before the event, as represented by $GT - 2 < Pred < GT + 10$. The evaluation method employs the F1 score that combines Precision ($\frac{TP}{TP+FP}$) and Recall ($\frac{TP}{TP+FN}$), to thoroughly assess both the correctness of abnormal events and identification timing.

$$\text{F1 score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{13}$$

The previous studies on the ABODA dataset assess performance by identifying the positions of abandoned objects. In contrast to the KISA dataset, the ABODA metric does not focus on the temporal accuracy of detecting occurrence instances. Consequently, studies relying on long-term background modeling have been unable to promptly detect abandoned objects. In our experimental approach, we adopt a stricter criterion, similar to the strategy from Park et al. (2019), which includes setting a detection time limit for abandoned objects to enable early identification. This evaluation strategy aligns with the abandonment detection criteria described for the KISA dataset in Table 3. By performance evaluation that more rigorous assessment than prior research, we demonstrate comparative results of our system.

The FireNet dataset is evaluated through fire classification in individual extracted frames, using the F1 score as the evaluation metric. Despite our system's focus on processing video data, we restrict our method to only use the Scene Analyzer for comparison with conventional fire detection methods. We tailor the fire scene prompt to the dataset configuration of fire and non-fire images, disregarding factors associated with human presence or location. Unlike conventional

Table 3
Criteria for evaluating performance in the KISA dataset.

Scenario	Definition Ground truth event time
Intrusion	Occurs when one or more individuals invade a specific area Refers to the time when the entire human body enters the area of intrusion
Loitering	One or more people roam in a specific area for more than 10 s Refers to the time 10 s after the entire human body enters the loitering area
Abandon	A situation in which a person disappears after dumping garbage or bags Refers to the time 10 s after a human puts trash or bags on the ground
Arson	A situation in which smoke or flames are produced by arson Refers to the time 10 s after a human tries to generate smoke or flames

Table 4
Performance result of our system on KISA CCTV dataset: all scores are F1 score (%).

Scenario	Validation	Certification	Abroad
Intrusion	96.7%	96.7%	89.4%
Loitering	99.3%	100%	92.9%
Abandon	96.0%	94.9%	96.3%
Arson	98.0%	90.0%	92.9%

Table 5
Ablation results of our human tracking module on the distribution subset market scenario.

Algorithm	TP↑	FP↓	FN↓	F1↑
Baseline	468	82	57	87.07%
+ Object feature	512	71	13	92.41%
+ Object filter	511	18	14	96.96%

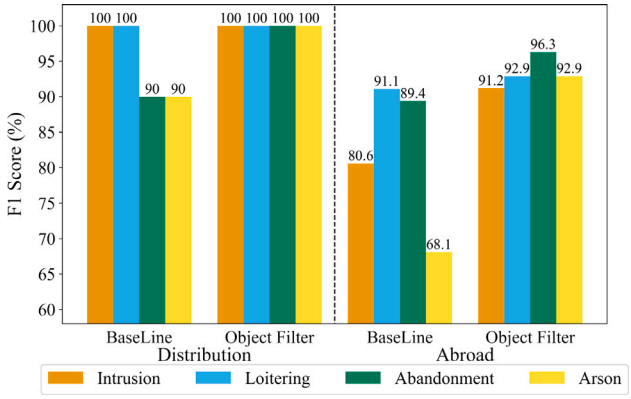


Fig. 11. Comparison of F1 scores (%) between Baseline and Object Filter configurations for the object filtering algorithm on the KISA CCTV dataset.

methods that hinge on training datasets (Jadon et al., 2019; Kumar & Sankarasubramanian, 2021; Xu, Guo, & Saleh, 2020), our approach takes the zero-shot approach. We modify the text within the scene prompt and assess performance without additional training efforts. For this experiment, the prompt configuration encompasses keywords such as "sky","sea", "forest", and "mountain", designated as non-fire labels, while "fire" and "flames" represent the fire label. Furthermore, unlike the typical approach in CLIP prompt engineering, which employs 80 different templates to enhance performance (Radford et al., 2021), we use the single template "The picture of label" within our prompt engineering. Importantly, our scene analyzer classifies an image as indicating a fire when the two highest similarities correspond to the fire label.

4.3. Experimental results

Performance on KISA CCTV Dataset is described in Table 4, shows F1 scores for scenarios including Intrusion, Loitering, Abandonment, and Arson. Our system consistently demonstrated high performance across most scenarios, with F1 scores exceeding 90%. By robust human tracking on low-light conditions, our system outperforms the previous result (Kim et al., 2021b) and demonstrates enhanced monitoring in Intrusion and loitering identification. Challenges in the Abandonment scenarios stemmed mainly from the misclassification of certain object categories, such as pink bags and military duffel bags, and from object occlusion by the owner. In the Arson scenario, false positives were primarily triggered by misidentifications involving an abandoned red fuel tank or the intermittent illumination of street lights under foggy conditions, illustrating the system's sensitivity to specific environmental stimuli. This comprehensive analysis demonstrates the system's robust capability in accurately identifying security incidents within a diverse range of environmental settings.

Ablation Study on Human Tracking Module assesses the tracking robustness within market scenarios with high pedestrian traffic in the KISA CCTV dataset. This scenario includes 25 people-counting videos and 50 people-queuing videos for each Validation and Certification subset, applying a 4-second time window ($GT - 2 < Pred < GT + 2$) to evaluate the precision of entry and exit reporting. Notably, our system achieved 97.7% F1 score in verification and 98.4% in certification, demonstrating reliable human tracking capabilities across various settings. Detailed results from the ablation study, shown in Table 5, focus on the Distribution subset to examine the impact of algorithmic enhancements on pedestrian counting accuracy. Starting with a baseline F1 score of 87.07% using deep sort tracking, the addition of object features significantly improved system performance to 92.41%. This enhancement markedly reduced false negatives from 57 to 13 due to tracking failures, although 71 false positives emphasize the need for further refinements to reduce incorrect identifications. Subsequent refinement with the Object Filter effectively reduced false positives, elevating the F1 score to 96.96% by filtering out tracked HFPs, and demonstrated significant improvement in human tracking within densely populated market environments.

Impact of the Object Filtering is shown in Fig. 11, which details the ablation results across scenarios on the KISA CCTV dataset. The scenarios include Intrusion, Loitering, Abandon, and Arson, evaluated in both Distribution and Abroad subsets. Introducing the Object Filter (OF) improved the F1 scores across all scenarios and subsets. Without object filter in the Distribution subset, the Abandon and Arson scenarios each fail to detect the correct event in one sample. In the Abroad subset, the Object Filter demonstrated its effectiveness, particularly in the Intrusion and Arson scenarios, where it significantly outperformed the Baseline (BL). In the case of Intrusion, if the HFP enters the alert zone even briefly, the alarm can be triggered and fatal. In the case of Arson, the patch processor may generate and analyze background scenes unrelated to the fire caused by the HFP, resulting in false alarms.

Ablation Study on Hand Luggage Tracker within the KISA abandonment scenario, as shown in Fig. 12, provides a comprehensive evaluation of our algorithmic interventions: Baseline, Object Filter, Top-Down method, and combination of both. Notably, the integration of the Object Filter and Top-Down approach achieved superior F1

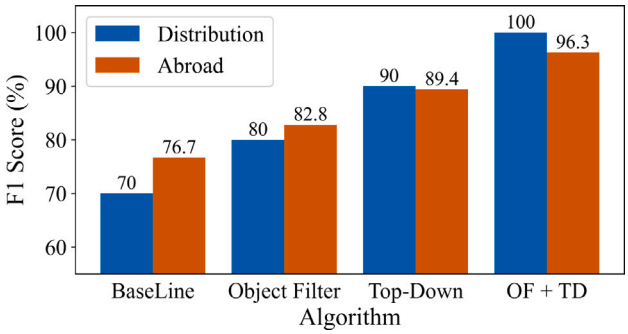


Fig. 12. Ablation results for our Top-Down (TD) approach and Object Filter (OF) in the Abandonment scenario on the KISA CCTV dataset, using the Distribution and Abroad subsets. The Distribution subset comprises 10 videos, and the Abroad subset contains 400 videos.

Table 6
Delay analysis: Time from event occurrence to alert generation by scenario, with average delays and standard deviations (s).

Intrusion	Loitering	Abandon	Arson
1.598 ± 1.766	0.808 ± 0.824	6.843 ± 2.052	2.394 ± 2.095

scores in both the Distribution and Abroad subsets, reaching 100% and 96.3%, respectively. In particular, the Top Down method enabled stable tracking of small luggage items in outdoor settings, thereby ensuring precise abandonment detection. Moreover, the fusion of the Top-Down method with object filters not only escalated detection accuracy but also minimized computational demands by effectively filtering out false positives. This synergy demonstrates a resource-efficient enhancement, significantly benefiting the overall system performance by balancing accuracy with computational efficiency.

Alert Generation Delay Times are presented in Table 6 for each scenario: Intrusion, Loitering, Abandonment, and Arson. This assessment utilized the Abroad subset for measuring the time elapsed from the ground truth (GT) event to the generation of an alert. It was confirmed that the average time for the Intrusion scenario is 1.6 s. This result reflects the Object Filter's latency in validating individuals who rapidly enter the intrusion area immediately upon appearing on video. In the Loitering scenario, where individuals remain stationary for extended periods, alerts were almost instantaneous. For Abandonment, our system could generate alerts more rapidly than traditional foreground methods. The 6.8-second delay includes the time it takes for the person to leave the object, demonstrating our system's efficiency in quickly confirming abandonment. In the case of Arson, the system effectively detected arson-related activities, not just the presence of fire, generating alerts within a short time of 2.4 s. Our system's ability to promptly generate alerts enables monitoring centers to expect rapid responses.

Optimizing Computational Speed through variable batch sizes is analyzed using three models: the object detector, human tracker, and image encoder, as shown in Fig. 13, focusing on both batch processing time and frame delay time. Our system was tested on an Intel Core i7-10700K CPU and an NVIDIA RTX 2080 Ti GPU. The graph indicates that while batch processing times generally lengthen as the batch size increases, this trend does not uniformly lead to improved delay times per frame. Specifically, the batch processing time for the Object Detector more than doubles, increasing from 112.2 ms to 237 ms as the batch size expands from 16 to 32. Interestingly, all models demonstrate optimal frame delay times at a batch size of 16, indicating a peak in computational efficiency. Therefore, we implemented variable batch processing with a maximum batch size of 16, our system accumulates frames that are received from the camera via RTSP until the subsequential process. This approach ensures effective real-time frame handling and enhances overall system responsiveness and reliability.

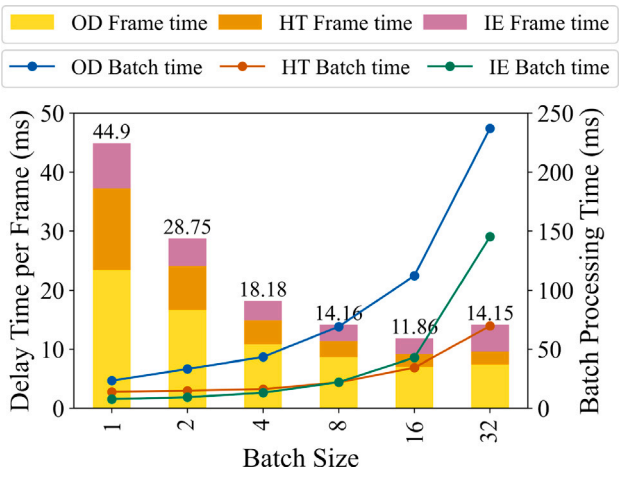


Fig. 13. Computation speed of three core models: the Object Detector (OD), Human Tracker (HT), and Image Encoder (IE) across various batch sizes. The optimal frame delay times are achieved at a batch size of 16.

Table 7
Comparison with other abandonment detection methods on the ABODA.

Method	TP↑	FP↓	F1↑
Wahyono et al. (2016)	7	0	73.7%
Lin et al. (2015)	12	6	80.0%
Liao et al. (2017)	12	4	85.7%
Krusch et al. (2017)	11	2	88.0%
Park et al. (2019)	11	0	95.7%
Shyam et al. (2018)	11	0	95.7%
Kim et al. (2021a)	11	0	95.7%
Ours	11	0	95.7%

4.4. Comparison result

4.4.1. Comparison on the ABODA dataset

The ABODA dataset's comparison evaluation for abandonment detection methods, as shown in Table 7, reveals a spectrum of performances. Lin et al. (2015) set a baseline with an F1 score of 80.0%. Wahyono's approach (Wahyono et al., 2016) scored 73.7%, limited by its background referencing under abrupt lighting changes. Liao et al. (2017) achieved an 85.7% F1 score and demonstrated capability in theft detection, while Krusch et al. (2017) enhanced detection precision to 88.0% with Person and Size Filtering methods. Notably, DL-based detection model approaches by Kim et al. (2021a), Park et al. (2019), and Shyam et al. (2018) each achieved an impressive 95.7% F1 score with zero false positives. Our system only employs the COCO dataset and without relying on foreground information, matches this high performance and demonstrates effective detection of abandonment scenarios by the Top-Down approach.

4.4.2. Comparison on the FireNet dataset

The evaluation of fire detection methods on the FireNet dataset presented diverse approaches and results described in Table 8. The foundational research in computer vision for fire image classification, conducted by Filonenko, Hern'andez, and Jo (2017) established early benchmarks with F1 scores of 90%, as reported by Saponara, Elhanashi, and Gagliardi (2020). FireNet (Jadon et al., 2019) exhibits a promising approach for fire region detection. Moreover, FireNet-v2 (Shees, Ansari, Varshney, Asghar, & Kanwal, 2023), significantly reduces false positives and achieves a precision rate of 99.28% even without training on the FireNet dataset. Light-ResNet (Ayala, Lima, Fernandes, Bezerra, & Cruz, 2019) presents comparative performance analysis across diverse fire datasets and exhibits an accuracy of 96.33 when train with FireNet. Despite the lightweight model, FireNet-Micro (Marakkaparambil, Rameshkumar, Dinesh, Aslam, & Ansari, 2023) demonstrated a high

Table 8
Comparison with other fire detection methods on the FireNet dataset.

Method	Accuracy↑	Specificity↑	Recall↑	Precision↑	F1↑
Filonenko et al. (2017)	85	[b]63	96	85	90
R-CNN (Saponara et al., 2020)	93.6	[b]80	**100**	92.4	96.04
FireNet (Jadon et al., 2019)	93.91	93.88	93.93	97.04	95.46
FireNet-v2[a] (Shees et al., 2023)	94.95	**98.56**	93.24	**99.28**	96.17
Light-ResNet (Ayala et al., 2019)	96.33	–	–	–	–
FireNet-Micro (Marakkaparambil et al., 2023)	96.78	95.32	97.47	97.8	97.64
UFS-Net[a] (Hosseini et al., 2022)	97.13	[b]96.40	97.47	98.29	97.83
Ours[a]	**99.08**	97.12	**100**	98.67	**99.33**

[a] Indicated without training on FireNet dataset.
[b] Denote recalculated based on the number of test samples.

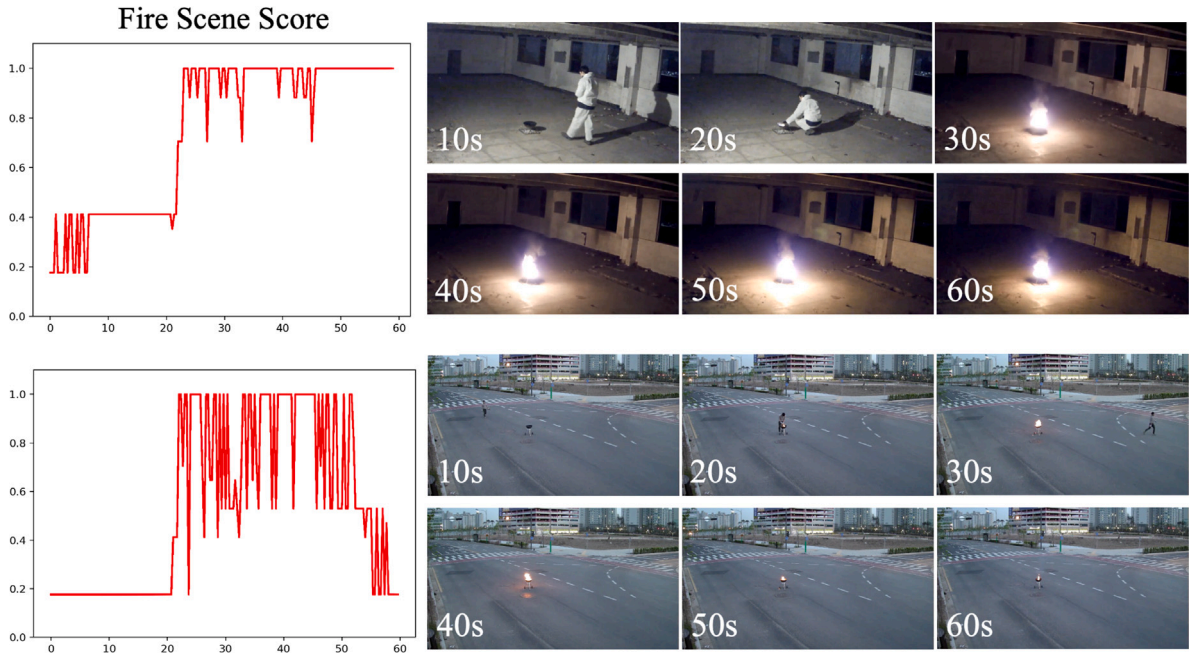


Fig. 14. Qualitative Results: Processed Fire Scene Scores from the KISA CCTV Distribution subset. The arson event begins at 20 s, and the score graph demonstrates fire detection capabilities, even when the fire starts small.

accuracy of 96.78. UFS-Net (Hosseini, Hashemzadeh, & Farajzadeh, 2022), which also performs smoke detection, proposed a state-of-the-art result with an F1 score of 97.83%. Our approach achieves state-of-the-art performance with the accuracy of 99.08% and F1 score of 99.33, even without requiring any training on fire dataset. This approach presents the efficiency and adaptability of our system in recognizing and responding to specific scenarios based on text prompts. Consequently, text-based surveillance demonstrates the potential to attain high accuracy in detection tasks, circumventing the traditional dependence on extensive training datasets.

4.5. Qualitative result

Qualitative Results on Arson Detection, shown in Fig. 14, present video samples from the KISA distribution subset alongside Fire Scene Score across a timeline. These videos were specifically modified to include an arson incident starting at 20 s and a GT event marked at 30 s. The graph reveals that the score fluctuated in the range of 0.2 to 0.4 before the arson incident. However, the score sharply rises upon the ignition of the fire and maintains a high level until the GT event. Our prompt-based methodology excels in generating explainable scores derived from textual descriptions, serving as an intuitive feature for users. This strategy not only bolsters the reliability of arson detection but also furnishes users with a transparent and understandable explanation for alarm triggers. Such clarity in the detection rationale significantly augments the system's overall efficacy and fosters user trust. By integrating textual analysis, our approach ensures that users are not just alerted to potential dangers but are also provided with contextually rich, comprehensible insights, thereby enhancing both the responsiveness and the interpretability of the system.

Visualization of Attention Maps using the relevancy-based method proposed by Chefer, Gur, and Wolf (2021), as illustrated in Fig. 15, demonstrates the extent of activation in calculating similarities between text and image within the CCTV domain. The figure consists of two parts: (A) depicts a person igniting a substantial fire next to a traditional house, while (B) captures a person committing arson on a road and then escaping the scene. In (A), the relevancy maps for the text "fire" and "man doing a barbecue" focus on the fire's location, resulting in high similarity scores of 23.28 and 26.98, respectively. The heatmap for "walking person" directs attention to the movement of the legs, demonstrating that the model can focus on the person to calculate similarity based on specific actions. In (B), as the arsonist flees, the similarity score for "walking person" is higher at 26.12 compared to "fire" at 22.94, reflecting the reduced visual prominence of the fire. The relevancy maps show distinct attention to both the fire and person areas but prioritize similarity related to human action over the scene, suggesting the need for our multi-prompt strategy. Additionally, the heatmap for "Man doing a barbecue" in part (B) demonstrates the model's capability to simultaneously observe both the person and the fire area, capturing the dynamics of human–object interaction. This illustration exemplifies how the image encoder differentiates between

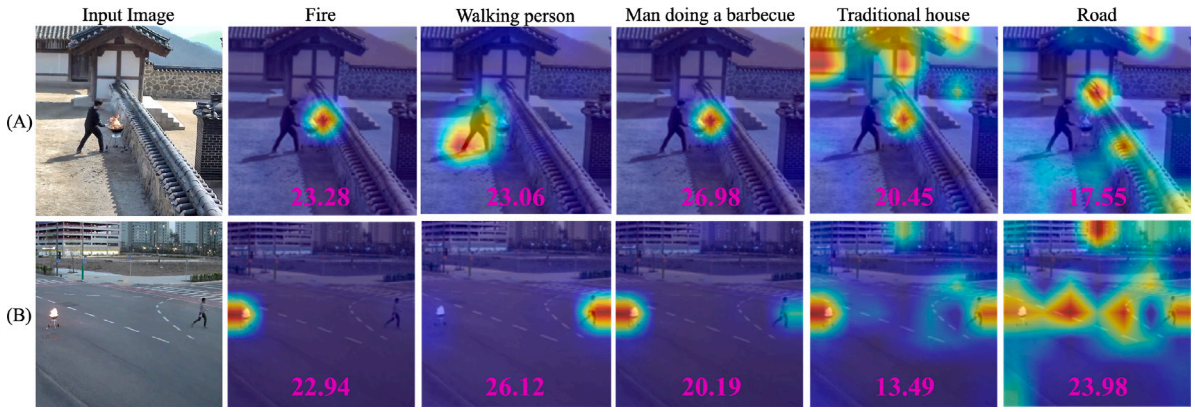


Fig. 15. Visualization of Text–Image Attention Maps in CCTV Footage. (A) shows an individual igniting a fire next to the fence of a cultural heritage site. (B) depicts an individual committing arson on a road and then fleeing the scene. The heatmap uses a color gradient from blue to red to visualize the intensity of attention, with pink numbers indicating measured similarity scores. Each heatmap appropriately focuses on the fire area, the individual actions, and the surrounding area to calculate the similarity with the given text.

elements like human actions and environmental features to assess textual similarity, effectively classifying dynamic situations and decoding complex interactions in the surveillance footage.

4.6. Discussion

In sparsely populated areas, where potential hazards may often remain unnoticed, our surveillance system strategically focuses on detecting unusual activities of individuals rather than relying on direct reports from crowds. However, this strategic focus reveals limitations in environments with high population density, with reduced speed efficiency that computational delays become noticeable. For instance, our tests indicated a delay of approximately 10 s when the system was tasked with continuously tracking 10 individuals over a period exceeding two minutes. Such delays stem primarily from the increased computational cost, which scales with the number of people the system attempts to monitor. To address these challenges in crowded environments, implementing appropriate frame skipping could be necessary to balance the trade-off between performance and speed effectively.

Currently, our surveillance system operates without the capability to autonomously learn from past detections, relying solely on manual updates and configurations to maintain its effectiveness. Recognizing the limitations this imposes, we are planning significant enhancements for the system's development. Our primary goal is to integrate advanced learning algorithms that will enable the system to independently adapt and learn from new anomalies and environmental changes. This pivotal development is expected to lead to the creation of autonomous detection agents. These agents will be designed to operate based on predefined user goals and will be equipped to autonomously adapt to new and varying operational environments. The realization of these advancements will reduce the need for frequent manual interventions, thus streamlining operations, and will also extend the system's versatility and applicability, enabling it to function effectively across a broader range of scenarios and settings.

5. Conclusion

This paper introduces the proactive anomaly surveillance system for multiple abnormal recognition in challenging outdoor CCTV environments. Our human tracking approach involves robust human feature extraction and advanced object filtering algorithms. Additionally, the framework efficiently detects human-related abnormal situations such as intrusion, loitering, abandonment, and arson with high performance. Our system offers the flexibility to define alert areas for intrusion and loitering and precisely calculates intrusion degree through the intersection function. Furthermore, our top-down approach ensures the accurate monitoring of pedestrian luggage, even over long distances. Leveraging the visual-language model in our scene analyzer, we analyze fire scenes and arson behavior for early arson detection. Our surveillance system and comprehensive evaluation demonstrate strong potential for practical applications in real-world environments, providing a holistic solution for diverse requirements.

CRediT authorship contribution statement

Hobeom Jeon: Methodology/Study design, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Hyungmin Kim:** Methodology/Study design, Methodology/Study design, Software, Validation, Resources, Writing – original draft. **Dohyung Kim:** Conceptualization, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jeahong Kim:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government foundation (24ZB1200, Human-centered autonomous intelligence system source technology research, 50%) and the Ministry of Trade, Industry and Energy (MOTIE) in the year 2024 Robot Industrial Technology Project "Development of companion robot technologies capable of emotional connection based on Human–Robot physical and cognitive interaction."(No. 20018513, 50%)

References

Ayala, A., Lima, E., Fernandes, B., Bezerra, B. L., & Cruz, F. (2019). Lightweight and efficient octave convolutional neural network for fire recognition. In *2019 IEEE latin American conference on computational intelligence* LA-CCI, (pp. 1–6). IEEE.

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

Bochinski, E., Eiselein, V., & Sikora, T. (2017). High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance* AVSS, (pp. 1–6). IEEE.

Bolya, D., Foley, S., Hays, J., & Hoffman, J. (2020). Tide: A general toolbox for identifying object detection errors. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part III 16* (pp. 558–573). Springer.

Cao, J., Pang, J., Weng, X., Khirodkar, R., & Kitani, K. (2023). Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9686–9696).

Cao, Y., Yang, F., Tang, Q., & Lu, X. (2019). An attention enhanced bidirectional LSTM for early forest fire smoke recognition. *IEEE Access*, *7*, 154732–154742.

Caruccio, L., Polese, G., Tortora, G., & Iannone, D. (2019). EDCAR: A knowledge representation framework to enhance automatic video surveillance. *Expert Systems with Applications*, *131*, 190–207.

Chefer, H., Gur, S., & Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 397–406).

Chen, Z., Cai, H., Zhang, Y., Wu, C., Mu, M., Li, Z., et al. (2019). A novel sparse representation model for pedestrian abnormal trajectory understanding. *Expert Systems with Applications*, *138*, Article 112753.

Chen, Y.-W., Chen, K., Yuan, S.-Y., & Kuo, S.-Y. (2016). Moving object counting using a tripwire in H. 265/HEVC bitstreams for video surveillance. *Ieee Access*, *4*, 2529–2541.

Chino, D. Y., Avalhais, L. P., Rodrigues, J. F., & Traina, A. J. (2015). Bowfire: detection of fire in still images by integrating pixel color and texture analysis. In *2015 28th SIBGRApI conference on graphics, patterns and images* (pp. 95–102). IEEE.

Donald, F. M. (2019). Information processing challenges and research directions in CCTV surveillance. *Cognition, Technology & Work*, *21*(3), 487–496.

Donald, F., Donald, C., & Thatcher, A. (2015). Work exposure and vigilance decrements in closed circuit television surveillance. *Applied Ergonomics*, *47*, 220–228.

Doshi, K., & Yilmaz, Y. (2022). Rethinking video anomaly detection-a continual learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3961–3970).

Filonenko, A., Hern'andez, D. C., & Jo, K.-H. (2017). Fast smoke detection for video surveillance using CUDA. *IEEE Transactions on Industrial Informatics*, *14*(2), 725–733.

Foggia, P., Saggese, A., & Vento, M. (2015). Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(9), 1545–1556.

Galvão, L. G., & Huda, M. N. (2023). Pedestrian and vehicle behaviour prediction in autonomous vehicle system—A review. *Expert Systems with Applications*, Article 121983.

Gan, H. M., Fernando, S., & Molina-Solana, M. (2021). Scalable object detection pipeline for traffic cameras: Application to Tfl JamCams. *Expert Systems with Applications*, *182*, Article 115154.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

Hashemzadeh, M., & Zademehdi, A. (2019). Fire detection for video surveillance applications using ICA K-medoids-based color model and efficient spatio-temporal visual features. *Expert Systems with Applications*, *130*, 60–78.

Höferlin, M., Höferlin, B., Weiskopf, D., & Heidemann, G. (2011). Uncertainty-aware video visual analytics of tracked moving objects. *Journal of Spatial Information Science*, *2011*(2), 87–117.

Hosseini, A., Hashemzadeh, M., & Farajzadeh, N. (2022). UFS-Net: A unified flame and smoke detection method for early detection of fire in video surveillance applications using CNNs. *Journal of Computer Science*, *61*, Article 101638.

Huang, T., Han, Q., Min, W., Li, X., Yu, Y., & Zhang, Y. (2019). Loitering detection based on pedestrian activity area classification. *Applied Sciences*, *9*(9), 1866.

Jadon, A., Omama, M., Varshney, A., Ansari, M. S., & Sharma, R. (2019). FireNet: a specialized lightweight fire & smoke detection model for real-time IoT applications. arXiv preprint arXiv:1905.11922.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Fang, J., et al. (2022). *ultralytics/yolov5: v6. 1-TensorRT, TensorFlow dge TPU and OpenVINO export and inference*. Zenodo.

Kang, J., & Kwak, S. (2014). Loitering detection solution for CCTV security system. *Journal of Korea Multimedia Society*, *17*(1), 15–25.

Karanam, S., Li, Y., & Radke, R. J. (2015). Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE international conference on computer vision* (pp. 4516–4524).

Kim, D., Kim, H., Mok, Y., & Paik, J. (2021a). HLDNet: Abandoned object detection using hand luggage detection network. *IEEE Consumer Electronics Magazine*, *11*(4), 45–56.

Kim, D., Kim, H., Mok, Y., & Paik, J. (2021b). Real-time surveillance system for analyzing abnormal behavior of pedestrians. *Applied Sciences*, *11*(13), 6153.

KISA (2017). Korea internet and security agency's Korea intelligent camera testing center. https://www.kisa.or.kr/1050607. (Accessed 21 August 2023).

Krusch, P., Bochinski, E., Eiselein, V., & Sikora, T. (2017). A consistent two-level metric for evaluation of automated abandoned object detection methods. In *2017 IEEE international conference on image processing* ICIP, (pp. 4352–4356). IEEE.

Kumar, V. K. S., & Sankarasubramanian, P. (2021). Realtime pipeline fire & smoke detection using a lightweight CNN model. In *2021 IEEE international conference on machine learning and applied network technologies* ICMLANT, (pp. 1–4). IEEE.

Lee, S., Lee, S., Seong, H., Hyun, J., & Kim, E. (2023). Fallen person detection for autonomous driving. *Expert Systems with Applications*, *213*, Article 119242.

Lee, W.-K., Leong, C.-F., Lai, W.-K., Leow, L.-K., & Yap, T.-H. (2018). ArchCam: Real time expert system for suspicious behaviour detection in ATM site. *Expert Systems with Applications*, *109*, 12–24.

Li, Y., Liu, L., Shen, C., & Hengel, A. v. d. (2017). Mining mid-level visual patterns with deep CNN activations. *International Journal of Computer Vision*, *121*, 344–364.

Li, G., Yang, Y., & Qu, X. (2019). Deep learning approaches on pedestrian detection in hazy weather. *IEEE Transactions on Industrial Electronics*, *67*(10), 8889–8899.

Liao, W., Yang, C., Ying Yang, M., & Rosenhahn, B. (2017). Security event recognition for visual surveillance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, 19–26.

Lim, M. K., Tang, S., & Chan, C. S. (2014). Isurveillance: Intelligent framework for multiple events detection in surveillance videos. *Expert Systems with Applications*, *41*(10), 4704–4715.

Lin, K., Chen, S.-C., Chen, C.-S., Lin, D.-T., & Hung, Y.-P. (2015). Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance. *IEEE Transactions on Information Forensics and Security*, *10*(7), 1359–1370.

Mabrouk, A. B., & Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, *91*, 480–491.

Mahmoodi, J., & Salajeghe, A. (2019). A classification method based on optical flow for violence detection. *Expert Systems with Applications*, *127*, 121–127.

Marakkaparambil, S. I., Rameshkumar, R., Dinesh, M. P., Aslam, A., & Ansari, M. S. (2023). FireNet-micro: Compact fire detection model with high recall. In *UK workshop on computational intelligence* (pp. 65–78). Springer.

Martinel, N., Das, A., Micheloni, C., & Roy-Chowdhury, A. K. (2016). Temporal model adaptation for person re-identification. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part IV 14* (pp. 858–877). Springer.

Neupane, D., Bhattarai, A., Aryal, S., Bouadjenek, M. R., Seok, U.-M., & Seok, J. (2023). SHINE: Deep learning-based accessible parking management system. arXiv preprint arXiv:2302.00837.

Park, M., & Ko, B. C. (2020). Two-step real-time night-time fire detection in an urban environment using Static ELASTIC-YOLOv3 and Temporal Fire-Tube. *Sensors*, *20*(8), 2202.

Park, H., Park, S., & Joo, Y. (2019). Robust detection of abandoned object for smart video surveillance in illumination changes. *Sensors*, *19*(23), 5114.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.

Saponara, S., Elhanashi, A., & Gagliardi, A. (2020). Exploiting R-CNN for video smoke/fire sensing in antifire surveillance indoor and outdoor systems for smart cities. In *2020 IEEE international conference on smart computing* SMARTCOMP, (pp. 392–397). IEEE.

Shah, S. H. H., Karlsen, A. S. T., Solberg, M., & Hameed, I. A. (2023). An efficient and lightweight multiperson activity recognition framework for robot-assisted healthcare applications. *Expert Systems with Applications*, Article 122482.

Sharma, J., Granmo, O.-C., Goodwin, M., & Fidje, J. T. (2017). Deep convolutional neural networks for fire detection in images. In *Engineering applications of neural networks: 18th international conference, EANN 2017, athens, Greece, August 25–27, 2017, proceedings* (pp. 183–193). Springer.

Shees, A., Ansari, M. S., Varshney, A., Asghar, M. N., & Kanwal, N. (2023). FireNet-v2: Improved lightweight fire detection model for real-time IoT applications. *Procedia Computer Science*, *218*, 2233–2242.

Shyam, D., Kot, A., & Athalye, C. (2018). Abandoned object detection using pixel-based finite state machine and single shot multibox detector. In *2018 IEEE international conference on multimedia and expo* ICME, (pp. 1–6). IEEE.

Song, C., Huang, Y., Ouyang, W., & Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1179–1188).

Sutherland, I. E., & Hodgman, G. W. (1974). Reentrant polygon clipping. *Communications of the ACM*, *17*(1), 32–42.

Vipin, V. (2012). Image processing based forest fire detection. *International Journal of Emerging Technology and Advanced Engineering*, *2*(2), 87–95.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., et al. (2019). Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7942–7951).

Wahyono, Filonenko, A., & Jo, K.-H. (2016). Unattended object identification for intelligent surveillance systems using sequence of dual background difference. *IEEE Transactions on Industrial Informatics*, *12*(6), 2247–2255.

Wang, Y., Kitani, K., & Weng, X. (2021). Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE international conference on robotics and automation* ICRA, (pp. 13708–13715). IEEE.

Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., et al. (2018). Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8042–8051).

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

Wu, Z., Xue, R., & Li, H. (2022). Real-time video fire detection via modified YOLOv5 network model. *Fire Technology*, *58*(4), 2377–2403.

Xu, Z., Guo, Y., & Saleh, J. H. (2020). Tackling small data challenges in visual fire detection: A deep convolutional generative adversarial network approach. *IEEE Access*, *9*, 3936–3946.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), 2872–2893.

Yu, E., Li, Z., Han, S., & Wang, H. (2022). Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*.

Yuan, C., Liu, Z., & Zhang, Y. (2015). UAV-based forest fire detection and tracking using image processing techniques. In *2015 international conference on unmanned aircraft systems* ICUAS, (pp. 639–643). IEEE.

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, *129*, 3069–3087.

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3702–3712).

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2021). Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5056–5069.

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. *vol. 2*, In *Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004* (pp. 28–31). IEEE.