2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

DEVELOPING A MULTI-CAMERA ANALYSIS METHODOLOGY FOR PEOPLE COUNTING IN BUILDINGS

Yelin Ryu[1], Insoo Jeong[1], Seungmo Lim[1], Junghoon Kim[1], and Seokho Chi[1,2]
[1]Seoul National University, Seoul, Korea, Republic of (South Korea)
[2]Institute of Construction and Environmental Engineering, Seoul, Korea, Republic of (South Korea)

Abstract

Determining the number and positions of individuals inside a building is critical for effective data management and rescue operations during building disasters. Although recent vision-based methods have been developed for people counting and tracking, they often struggle to process video data from non-overlapping zones. To overcome these challenges, this study proposes a novel multi-camera analysis methodology that integrates a camera network with Re-Identification (Re-ID) models to manage personnel information. Experimental results demonstrate that the system provides precise zone-specific counts and locations of individuals. This approach can significantly improve emergency response by enabling strategic resource allocation and prioritized access.

Introduction

Accurately verifying the number and locations of occupants in a building before a disaster occurs is crucial for effective search and rescue operations while preventing further casualties in its aftermath. Building disasters represent a severe social crisis that causes significant casualties and property damage. For example, building fires account for the largest proportion (65%) of all fire incidents (National Fire Agency, 2023). Notably, 95 percent of the property damage caused by all social disasters is attributed to building disasters (Ministry of the Interior and Safety, 2023). Specifically, the intricate internal structures and high occupant density of the buildings are the primary reasons for the substantial damages during disasters.

Generally, to initiate search and rescue operations immediately at a disaster site, rescue teams gather information on the number and precise locations of occupants inside the building before arrival by referencing statements from building managers or records of previous rescue activities (Ko et al., 2021). However, this approach is prone to frequent errors because it heavily relies on memory and experience. These errors can significantly impact the golden time for life rescue operations. Therefore, an automated system for accurately counting people by floor and zone which can be seamlessly integrated into efficient rescue operations is essential.

Traditional studies on people counting employed various sensors, such as Radio Frequency Identification (RFID) tags, Wi-Fi, and environmental sensors for counting people (Ding et al., 2019; Tang et al., 2020; Choi et al., 2021; Segura and Albert, 2023; Otto et al., 2024). However, these sensors were inconvenient for users and incurred additional costs for installation and management. Furthermore, the accuracy of sensor-based people counting systems is highly sensitive to external factors, such as lighting, temperature, and network connectivity, which can significantly degrade performance under adverse conditions. Recently, with advances in deep learning technologies, several researchers have focused on computer vision-based people counting systems using Closed-Circuit Television (CCTV) video to efficiently acquire and process data. However, most of these systems still suffer from counting errors due to insufficient consideration of blind spots and limited use of correction mechanisms for counting errors.

Improving people counting requires additional assistance; for this purpose, a people matching algorithm is employed. Through this algorithm, the same person appearing in various locations at different times can be identified with a single ID. However, most research efforts using people matching algorithms utilized cameras with overlapping Fields of View (FOV), which limits their ability to analyze videos capturing entirely separate zones. Furthermore, insufficient consideration of the camera network has resulted in low accuracy for people matching.

To overcome the issue, this paper proposes a multi-camera analysis methodology for real-time people information management in buildings by combining a people matching algorithm with people counting. This methodology integrates a camera network with non-overlapping views into people matching and counting processes to improve the accuracy of people information. By implementing the proposed approach in actual buildings and integrating it with data management platforms, rescue teams can enhance situational awareness and prioritize which zones to enter based on the

real-time occupancy data, thereby enabling more efficient search and rescue operations. Furthermore, the system may also support routine facility operations, including safety monitoring and maintenance.

Literature Review

Previous Studies on People Counting System

People counting in spaces has been recognized as a key technology in various fields. In particular, the results of people counting in buildings are utilized for diverse purposes, such as building disaster response, energy management, behavior analysis, and interior design (Chun et al., 2023).

In the early stages of research, most people counting systems employed sensors. Some studies used wearable sensors like RFID tags and Bluetooth to count individuals in specific areas and analyze their behaviors. Wearable sensors monitor people by utilizing signals from devices dedicated to users or specific locations. Ding et al. (2019) proposed and installed the R# system using an RFID reader and 20 passive RFID tags in a supermarket aisle to estimate the number of individuals in the area. Segura and Albert (2023) applied Bluetooth devices to detect and count people, monitoring their concentration in particular zones. These sensors collect and analyze detailed information such as locations and movement paths. However, their use increases inconvenience for users by requiring them to wear sensors, making it challenging to ensure consistent usage in large-scale areas and identify random anonym people without sensors. Additionally, collecting personal data, including physical, locational, and biometric information, raises privacy concerns.

Non-wearable sensors such as Wi-Fi and Radar have also been used for counting people. Tang et al. (2020) proposed passive Wi-Fi sensors installed in rooms to identify individuals and estimate crowd size in controlled experimental environments. Choi et al. (2021) developed an IR-UWB Radar-based people counting system that integrates various domain features. Empirical results demonstrated that these systems can reliably determine the number of people in a specific Region of Interests (ROIs). Even so, if some zones are not connected to the same network or have weak signals, the sensors fail to detect them, which reduces accuracy. Some researchers have explored environment-based sensors. Otto et al. (2024) employed CO_2 levels data from sensors in various rooms to estimate presence and count the number of people. Nevertheless, environment-based sensors are highly sensitive to external factors such as lighting, temperature, and humidity. Since building disaster scenarios differ significantly from normal conditions and involve many environmental variables, it is hard for these sensors to achieve high accuracy.

Recently, rapid advancements in video analysis algorithms based on computer vision technologies have been achieved. The most convenient way to capture video data for a building is by utilizing the CCTV cameras installed on the premises. Consequently, research on vision-based people counting systems has shown more accurate and efficient results using CCTV footage. Chae et al. (2020) proposed an algorithm for counting the number of individuals entering a building in real time using video footage captured at building entrances with a pyramid-type hierarchical area structure. Hong and Mazlan (2023) developed an automated people counting methodology that incorporates object detection and tracking from video inputs. The system includes notification features to send updates about the number of occupants in a target area. However, these studies are limited by their use of a single camera. Since single cameras have a restricted FOVs, these methods cannot detect or count individuals in areas with blind spots.

To address this issue, some approaches have introduced people counting systems in multi-camera environments. Zhang et al. (2021) proposed a Cross-View Cross-Scene (CVCS) system that selects and fuses images from different perspectives, showing improved accuracy in crowd counting. Hu et al. (2022) combined a feature extraction network with a three-stage detection module to estimate building occupancy. Park et al. (2024) introduced a queue-buffer algorithm to correct multi-camera-based counting errors. These studies commonly state the importance of analyzing data from multiple cameras, including those covering non-overlapping areas for reliable people counting. Even so, while many methods perform people counting at the building or floor level, few have addressed zone-level counting. Moreover, most do not provide detailed information such as individual identification or localization.

Multi-camera People Re-Identification

People matching algorithm generally employs Re-Identification (Re-ID) techniques, the process of recognizing and associating the same object across multiple views. Initial research efforts on Re-ID mainly focused on hand-crafted features from body structures (Liao et al., 2015), but it is difficult to handle occluded parts or individuals with various poses.

In recent years, studies on Re-ID using deep learning for detailed feature extraction and precise similarity comparison have been actively conducted. Zhang et al. (2021) proposed a Deep Re-Identification Occlusion Processing (DROP) framework to address occlusion issues in multi-camera videos by re-identifying and monitoring objects after full occlusion. Jang et al. (2022) utilized ground projection to extract movement information and Dynamic Time Warping (DTW) for multi-object matching and tracking method under indoor multi-camera settings. However, existing studies primarily focus on multiple videos with overlapping regions and insufficiently account for non-overlapping views.

Some methods for multi-camera people Re-ID in non-overlapping areas have been proposed, using tracklets and correlation clustering to enhance the accuracy of feature extraction (Riachy et al., 2019; Wu et al., 2020). These studies not only detect people but also infer trajectories using appearance features. Liu et al. (2021) enhanced matching performance by leveraging multi-scale graphs

to capture spatial-temporal dependencies across non-overlapping views. Peng et al. (2024) proposed continuous Re-ID to ensure consistent retrieval of the target person across all camera views, considering camera location and temporal information. However, existing methods still face several limitations. Most of them suffer from low accuracy, as they rely heavily on visual features while paying little attention to external factors such as temporal and the camera network information. Moreover, many are not optimized for real-time processing, hindering their application in building occupancy management. To overcome the limitations of previous studies, this research proposes an innovative multi-camera analysis methodology that combines people counting and people matching based on the camera network matrix for the real-time collection and management of occupancy information in buildings.

Methodology

Figure 1 presents the overall pipeline of the proposed methodology, which consists of three modules: (1) people detection and tracking, (2) camera network matrix mapping, (3) people matching and counting. The outputs are the real-time number of people in each floor and zone of the building and the current location of individuals (ID).

People Detection and Tracking

People detection and tracking module detects and tracks people who are entering and moving in the building. At this stage, the module utilized a tracking-by-detection approach, where a tracking model assigns unique IDs to individuals based on the data provided by the detection model, enabling efficient tracking across consecutive frames. For accurate and real-time processing, the You Only Look Once (YOLOv8)-ByteTrack algorithm was implemented.

YOLO is a single-stage object detection model known for its fast detection speeds compared to other models like Regions with Convolutional Neural Networks (R-CNN) or Single Shot Detector (SSD). YOLOv8 incorporates architectural improvements that increase reliability, providing higher accuracy and reduced latency in object detection and segmentation (Shyaa and Hashim, 2024). This study deployed a pre-trained YOLOv8 model on the Common Objects in Context (COCO) dataset.

ByteTrack is a Multi-Object Tracking (MOT) method recognized for its high tracking accuracy. It integrates seamlessly with YOLO-based object detection models, enabling rapid simultaneous people detection and tracking. ByteTrack assigns unique IDs to individuals detected, enabling efficient tracking across frames, which makes it ideal for real-time applications.

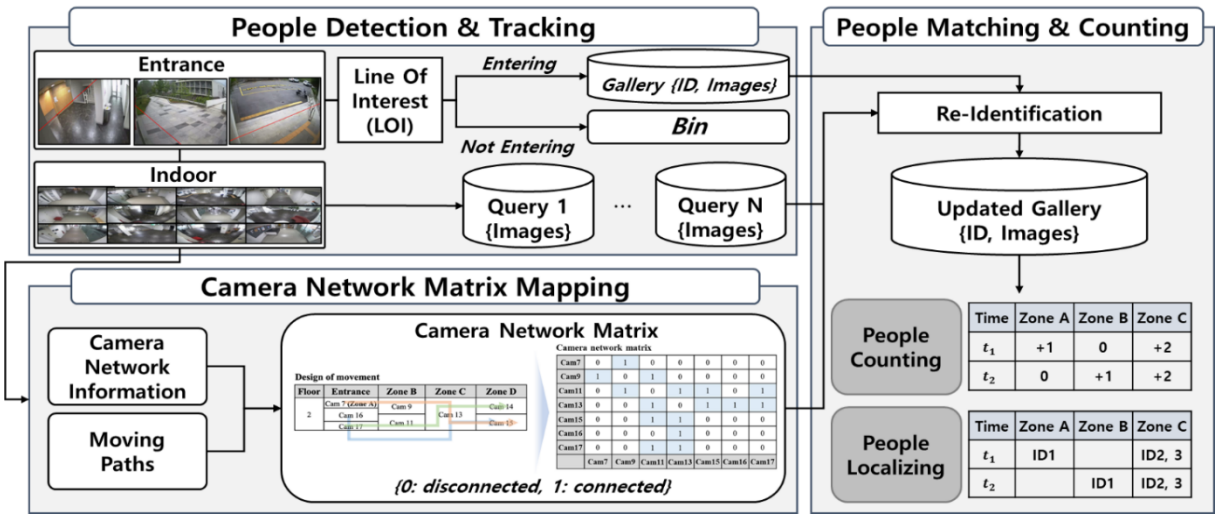


Figure 1: Overall Pipeline of the Proposed Methodology

Each bounding box from the YOLOv8-ByteTrack algorithm is associated with three attributes: timestamp, camera index, and unique person ID. These outputs are then divided into query and gallery datasets for people matching module. The query dataset consists of consecutive images of tracked individuals per frame to be matched, while the gallery dataset is a continuously updated database for matching. In this approach, individuals tracked in CCTV footage of building entrances are designated as the initial gallery dataset, and those tracked indoors are classified as the query dataset.

In constructing the initial gallery dataset, a Line of Interest (LOI) is defined at the building entrance video to accurately detect people entering from outside. The LOI system is widely used in people counting studies to calculate the total number of occupants in restricted access areas such as stores, metros, and malls (Zheng et al., 2019; Tien et al., 2021; Chun et al., 2023). Considering the entrance locations and movement directions, LOIs are drawn as straight lines defined by the x and y coordinates of their start and end points. Unique IDs are assigned to individuals crossing a virtual line (LOI), and these IDs form the initial gallery dataset for the matching module to prevent matching errors caused by duplicated IDs. For example, if a person is detected but does not cross the LOI, they are considered as an outsider and are not assigned an ID. Figure 2 illustrates the process of detecting and assigning IDs to individuals using the LOI.

Figure 2: Example of Constructing Initial Gallery using LOI

Camera Network Matrix Mapping

Camera network matrix mapping module is designed to establish sequential relationships between multiple cameras in non-overlapping areas, where blind spots make people matching challenging. In such large-scale environments, inferring camera network topology is crucial for improving tracking accuracy. Traditional appearance-based methods overlook the structure of the camera network, resulting in reduced matching performance. While some studies focus on walking speeds to predict transitions (Cho et al., 2017) is difficult when people move at significantly different speeds. Others used camera calibration (Cho & Yoon, 2019), but internal camera parameters are often inaccessible in building CCTV systems due to security constraints.

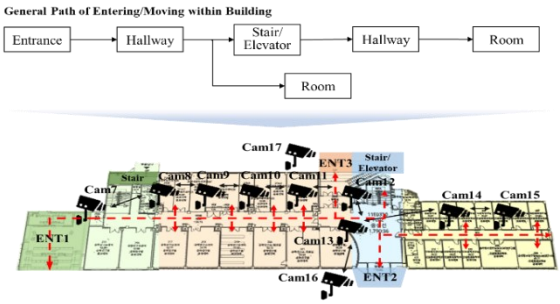


Figure 3: Manual Design of Moving Patterns

A key characteristic of the building is that it is a confined space, limiting the movement paths of people. Therefore, this study infers the camera network based on observed movement patterns, modeling it as a matrix. Four common facility types in buildings were defined: entrances, stair/elevators, hallways, and rooms to model typical movement flows within buildings. Figure 3 illustrates the representative movement paths constructed using this classification. By identifying entrances and connecting hallways, plausible routes among cameras can be simulated. The resulting camera network structure reflects ordered relationships between zones and corresponding cameras. Cameras in stair or elevator zones create vertical transitions, while those in hallways and rooms exhibit horizontal connectivity. This modeling of spatial transitions enables a more realistic and adaptable camera network representation for indoor spaces.

In the next step, the movement patterns can be converted into a matrix. Figure 4 shows an example of mapping the camera network matrix on a single floor based on the designed movements. The camera network matrix is calculated using binary values (0 and 1). A value of 0 indicates that no continuous relationship exists between two cameras, while a value of 1 indicates the presence of a continuous relationship. For example, in Figure 4, people in the building can appear consecutively on cameras 7 and 9, so they are considered sequentially connected (assigned a value of 1). In contrast, camera 13 and camera 9 are assigned 0, as people cannot move continuously between these two locations. This matrix helps predict the next camera where an object may appear after disappearing from the view of a previous camera, thereby improving matching and counting accuracy.

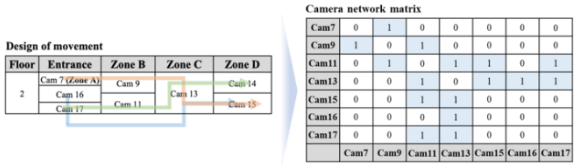


Figure 4: Example of Camera Network Matrix Mapping

People Matching and Counting

Figure 5 depicts the conceptual design of the people matching module. The module identifies the same individual who appears across various zones with non-overlapping views and assigns a unified ID. In the people matching module, a Re-ID technique is mainly utilized. Typically, constructing a person Re-ID system for a specific scenario involves a query-gallery search method. Specifically, given a person of interest (query) in matching and a pool of candidates (gallery), the query is matched to the gallery entry with the highest feature similarity. This process consists of two main steps: (1) feature extraction and (2) feature similarity calculation. In this study, the query dataset is generated from the people detection and tracking results in indoor videos. Each query is assigned with a random along with corresponding frame images before matching. Queries are input into the matching module and compared with the gallery dataset. The module then replaces the random ID with the gallery ID when it identifies the same person. As a result, the image pools of each ID in the gallery dataset are continuously updated over time.

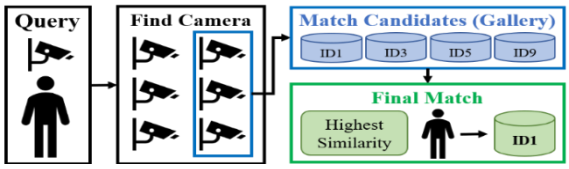


Figure 5: People Matching Module with Query and Gallery

The first step of the people matching module is extracting features from the dataset. Feature extraction involves identifying and representing both global and local features of a person using information within bounding boxes. Robust and reliable feature extraction not only enables accurate classification of different individuals but also mitigates the impact of complex environments. Traditional manual feature extraction methods (e.g. color, shape) are sensitive to changes in brightness and camera angles. Therefore, it is challenging to apply them across diverse scenarios. Consequently, deep learning-based feature extraction methods have become more prevalent. Deep learning models are less affected by environmental variations, allowing them to extract more hierarchical features (Yang et al., 2022).

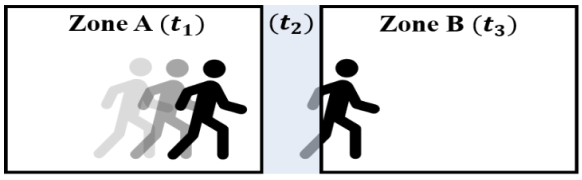
Three representative feature extraction models: ResNet50, Part-based Convolutional Baseline (PCB), and Swin Transformer V2 (Swinv2) were selected for comparison to identify the best model for extracting individual features within buildings. ResNet50 uses global features through a Convolutional Neural Network (CNN), PCB enhances local discriminability with part-based representations, and Swinv2 captures both local and global context. All models were trained and evaluated on the Market-1501 dataset (Zheng et al., 2015), a large-scale public benchmark for person Re-ID tasks. It contains 32,668 images of 1,501 individuals from six cameras. Of these, 12,936 images were used for training, and 3,368 query images were used to find matches across 19,732 gallery images. Based on the training results, Swinv2, with the highest mean Average Precision (mAP) of 0.781, was selected as the feature extractor for this study. Swinv2 is a transformer-based model that leverages a shifted window-based attention mechanism. Adaptive Split Self-Attention (ASA) module in the model enhances global feature modeling by autonomously focusing on important regions in an image. This allows Swinv2 to effectively address common Re-ID challenges, such as occlusion, pose variations, and cross-view appearance differences.

The next step in the people matching process is similarity calculation. After extracting features from the query and gallery datasets, cosine similarity is calculated by comparing query feature vector with those in the gallery to find the best match. Cosine similarity measures the similarity between vectors by calculating the cosine of the angle between them with the range from [-1, 1]. A higher cosine similarity value indicates a greater similarity. To reduce errors in people matching, a cosine similarity threshold is applied. If the similarity score between the query and a gallery exceeds the threshold, the gallery ID is classified as a potential matching candidate; otherwise, it is excluded. The query is ultimately matched to the gallery ID with the highest similarity score among the potential candidates. The formula for calculating the cosine similarity is:

$$sim(X, Y) = cos\ (\theta) = \frac{X \cdot Y}{\|X\|\|Y\|} \quad (1)$$

X and Y are the vectorized features of the query and gallery, respectively; θ denotes the angle between them in the vector space. In summary, the people matching system selects candidates from the gallery based on the camera network. Only the IDs detected in cameras located along the movement paths can be included as candidates. Then, it assigns the final ID which has the highest feature similarity. Once the matching of query is completed, the query is incorporated into the real-time updated gallery dataset.

Then, the number of people is counted by each floor and zone based on the results of people matching. The rules of the people counting system are depicted in Figure 6. When an individual previously located in Zone A (t_1) moves through a blind spot and cannot be tracked during t_2, their count remains in Zone A. However, once the person is detected in a Zone B (t_3), one count is removed from Zone A and added to Zone B.



Time	Zone A	Zone B
t_1	+1	0
t_2	+1	0
t_3	0	+1

Figure 6: People Counting Module

Results and Discussion

Experimental Setup

This study applied the proposed multi-camera analysis methodology for people counting in a three-story building at Seoul National University. The layout and camera locations are shown in Figure 7. Cameras that were closed during the experiment, such as the example in Figure 8, were excluded from the analysis. The experiment dataset includes 14 CCTV videos: three from the building entrances and 11 from indoor areas. All cameras are positioned in non-overlapping areas, resulting blind spots between views since no two video feeds share a common field of view. The dataset provides an appropriate testing environment for evaluating the proposed method in scenarios where direct visual continuity is not guaranteed. Each video is 10 minutes long and was recorded during the morning rush hour (8:45:00~8:55:00 a.m.). During the experiment, a total of 12 individuals (12 unique IDs) entered the building, with no one exiting. The video resolution is 1280 × 960, with a frame rate of 30 FPS.

To generate zone-based counting results, the study divided the building into four sections (A-D) per floor, resulting in 12 zones. This zoning was implemented based on the previously defined process of entering and moving within a building. The division specifically considered areas that support vertical movement (stairs, elevators) and those that allow horizontal movement (hallways, rooms). Accordingly, Zone A was designated as a stair area, Zone B as a hallway and room area, Zone C as a stair/elevator area, and Zone D as another hallway and room area. Since this study involves analyzing human images from CCTV footage, it received approval from the Institutional Review Board (IRB) (IRB No. 2408/004-018).

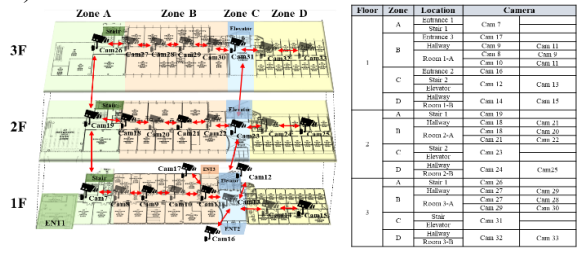


Figure 7: Layout of Target Building and Camera Location

Figure 8: Example of Camera Views

Evaluation Metrics

This study evaluated results of people matching and counting obtained using the proposed multi-camera analysis methodology. Precision, recall, and F1-score were utilized as evaluation metrics for both tasks. Specifically, the results for people matching employed ID Precision (IDP), ID Recall (IDR), and ID F1-score (IDF1) by incorporating the concept of ID. This extension is motivated by the inherent connection between matching and tracking, where issues like ID switches and object misidentification can significantly affect performance. While standard precision and recall focus on single-frame evaluations, IDP and IDR account for temporal consistency across frames. The equations for IDP, IDR, and IDF1 are as follows:

$$ID\ Precision\ (IDP) = \frac{IDTP}{IDTP+FP} \quad (2)$$

$$ID\ Recall\ (IDR) = \frac{IDTP}{IDTP+FN} \quad (3)$$

$$ID\ F1-score\ (IDF1) = 2 \cdot \frac{IDP \cdot IDR}{IDP+IDR} \quad (4)$$

IDTP presents the IDs correctly matched by the system to their corresponding ground truth IDs. FP refers to IDs that are incorrectly assigned, and FN occurs when a person is present and has an ID, but the system fails to give any ID or completely misses the person.

Additionally, Mean Absolute Error (MAE) is used as an additional evaluation metric for people counting. MAE, defined as the average of the absolute differences between predicted and actual values, provides an intuitive measure of error in the counting system for both overestimation and underestimation of the number of people. The formula for MAE is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \quad (5)$$

N represents the total number of zones in the building. $\hat{y}_i$ refers to the predicted number of people in the i-th zone, and y_i is the ground truth number of people in that zone.

Experimental Results and Discussions

In the people detection and tracking phase, several experiments were conducted with different configurations of the YOLOv8-ByteTrack algorithm, and the values that produced the best results are shown in Table 1. Since the initial detection system generated frequent errors due to distortion from fisheye cameras, three such views were excluded. As a result, the average precision, recall, and F1-score for people detection were 94.24%. 75.71%. and 82.93%, respectively. The Multiple Object Tracking Accuracy (MOTA) was then 84.13%.

Table 1: YOLOv8-ByteTrack algorithm Parameter Settings

Configuration	Parameter Value
Track buffer	30
Track thresholds	0.25 (high), 0.1 (low)
New track threshold	0.25
Match threshold	0.8
Min box area	10

Overall, the proposed people matching system successfully matched 19 out of 22 Re-ID cases (86.36%). To localize individuals, the people matching results from each camera were aggregated to generate ID lists by predefined zones over time. Table 2 presents the results of IDP, IDR, and IDF1 evaluations for people matching across building zones. Zones 3B and 3D were excluded as no one appeared, resulting in no tracking outcomes.

The average IDP was 96.46%, the IDR was 87.20%, and the IDF1 was 90.81%, demonstrating accurate real-time ID localization through the camera network and Re-ID. The camera network matrix significantly improved the matching performance by effectively establishing relationships between cameras, demonstrating its effectiveness in the challenging context of non-overlapping areas. These results enable rescue teams to identify individuals' locations during building disasters and prioritize vulnerable groups. However, recall in some zones was lower than other metrics due to individuals appearing smaller in the frame when farther from the camera, making detection and tracking more difficult.

Table 2: Results of People Matching by Zone

Zone	IDP	IDR	IDF1
1A	0.9804	0.9577	0.9688
1B	0.9943	0.8396	0.9027
1C	0.9582	0.8238	0.8804
1D	0.9897	0.9478	0.9678
2A	1.0000	0.7416	0.8259
2B	0.9773	0.8933	0.9258
2C	0.9372	0.9954	0.9645
2D	0.9657	0.8370	0.8910
3A	0.9449	0.8329	0.8809
3C	0.8857	0.8292	0.8556

Figure 9 illustrates an example of a mismatched result in cam 15. In the people matching system, both the camera network and feature similarities are considered. In cases like this, where a person lacks clearly distinguishable features (e.g., wearing dark-colored clothing), the similarity scores among available candidates are often very close, resulting in incorrect matches.

Figure 9: Example of Mismatched ID (Camera 15)

Table 3 summarizes the people counting results by zone. The average precision. recall, and F1-score were 92.54%, 97.51%, and 92.87%, respectively, and the average MAE was 0.0940. Compared to previous multi-camera-based people counting studies (Chun et al., 2023; Park et al., 2024), the accuracy improved by approximately 10%, while the MAE decreased by around 0.1.

Table 3: Results of People Counting by Zone

Zone	Precision	Recall	F1-score	MAE
1A	0.7436	1.0000	0.8529	0.0293
1B	0.9981	1.0000	0.9991	0.3596
1C	0.9980	1.0000	0.9990	0.5908
1D	1.0000	1.0000	1.0000	0.0000
2A	0.875	1.0000	0.9333	0.0073
2B	1.0000	1.0000	1.0000	0.0000
2C	0.1920	1.0000	0.3220	0.1467
2D	1.0000	0.6083	0.7564	0.3651
3A	1.0000	0.9939	0.9969	0.0055
3C	1.0000	1.0000	1.0000	0.0000

Figure 10 shows the people counting results on the 1st floor of the building and its visualization on a 2D model in a specific time. The image coordinates of the floor plan were used to map the results to each zone. Differences between the dashed and solid lines in the graph indicate points where counting mistakes occurred. The trends of results showed various patterns across zones. In some cases, low matching accuracy did not affect the count, as alternative IDs were correctly matched (e.g., Zone 1B and 2A). Conversely, in other instances, where matching was highly precise, the counting accuracy appeared relatively low (e.g., Zone 1A and 2C). This issue arose from counting errors caused by incorrectly matched IDs that should belong to other zones, eventually affecting subsequent counting processes. Nevertheless, aside from these special scenarios, it was observed that accurate matching results in each zone significantly contributed to achieving more precise people counting outcomes under realistic non-overlapping areas. These results underscore the importance of people matching in ensuring reliable people counting by mitigating errors and enhancing system performance. These findings highlight the proposed system's potential as a valuable tool for rescue teams, aiding in the strategic deployment into zones with the highest density during building disasters.

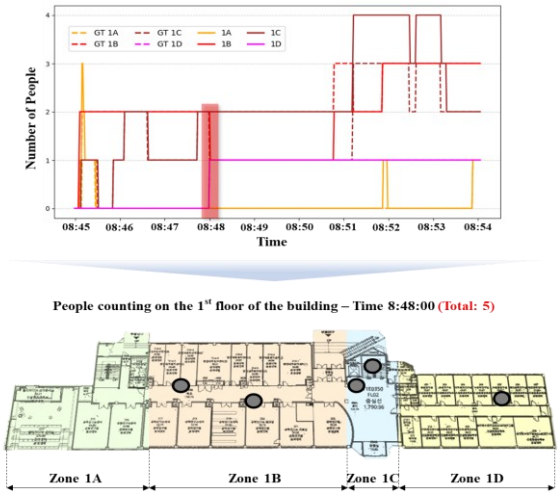


Figure 10: People Counting Results on the 1st Floor

Conclusions

This study proposed a multi-camera analysis methodology for real-time people counting in buildings, integrating tracking and matching techniques. Through an experiment conducted in a three-story building equipped with 14 CCTV cameras, the proposed system achieved counting precision, recall, and F1-scores of 92.54%, 97.51%, and 92.87%, respectively, with an average MAE of 0.0940. For the matching performance, the system demonstrated average IDP of 96.46%, IDR of 87.20%, and IDF1 of 90.81%. These results indicated that the proposed system enables highly accurate people counting even in non-overlapping areas by applying the matching algorithm.

The developed system provides real-time people localization and counting results at the zone level, visualized on a 2D floor map with high accuracy. By leveraging a camera network, the system mathematically analyzes movement trajectories and camera correlations to incorporate external factors into the counting process. This approach offers more granular counting results and demonstrates superior performance compared to other studies. Furthermore, the system can serve as a valuable tool for efficient decision-making during disaster response by providing rescue teams with real-time data of people in a building. Such information enables swift resource allocation and prioritization of rescue operations, ultimately saving lives and minimizing damage.

Further studies could address the limitations of the current system and improve scalability. The current approach treats individuals moving between zones as remaining in their previous zones, reducing the granularity and reliability of the counting results. Incorporating movement vector analysis could enable localization into categories such as "in transit" or "in a room". Additionally, integrating the proposed method with Building Information Modeling (BIM) could be pursued. Linking tracking data with spatial and structural information would support building management tasks like predictive building maintenance and energy optimization based on real-time occupancy patterns.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2003696, No. RS-2023-00241758).

References

Chae, S., Kwon, H., Park, S., Cho, W., Kwon, O., & Lee, J. (2020). CCTV high-speed analysis algorithm for real-time monitoring of building access. Journal of the Korean Society of Hazard Mitigation, 20(2), p.pp.113-118.

Cho, Y., & Yoon, K. (2019). Distance-based camera network topology inference for person re-identification. Pattern Recognition Letters, 125, p.pp.220-227.

Cho, Y., Park, J., Kim, S., Lee, K., & Yoon, K. (2017). Unified framework for automated person re-identification and camera network topology inference in camera networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, p.pp. 2601-2607.

Chun, H., Park, C., Chi, S., Roh, M., & Connie, S. (2023). Developing an Occupants Count Methodology in Buildings Using Virtual Lines of Interest in a Multi-Camera Network. KSCE Journal of Civil and Environmental Engineering Research, 43(5), p.pp. 667-674.

Hong, C., & Mazlan, M. (2023). Development of Automated People Counting System using Object Detection and Tracking. International Journal of Online & Biomedical Engineering, 19(6). p.pp.18-30.

Hu, S., Wang, P., Hoare, C., & O'Donnell, J. (2022). Building occupancy detection and localization using CCTV camera and deep learning. IEEE Internet of Things Journal, 10(1), p.pp.597-608.

Jang, J., Seon, M., & Choi, J. (2022). Lightweight indoor multi-object tracking in overlapping FOV multi-camera environments. Sensors, 22(14), 5267, p.pp.1-18.

Ko, Y., Shin, Y., Yoo, S., & Shin, D. (2021). Real-time location identification of indoor rescuees at accident sites and location-based rescue response. Journal of the Korean Institute of Gas, 25(3), p.pp.46-52.

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, p.pp.2197-2206.

Liu, J., Zha, Z. J., Wu, W., Zheng, K., & Sun, Q. (2021). Spatial-temporal correlation and topology learning for person re-identification in videos. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p.pp.4370-4379.

Ministry of the Interior and Safety. (2023). 2022 Disaster Annual Report.

National Fire Agency. (2023). 2023 National fire agency statistical yearbook.

Otto, R., Guedey, M., Pohler, B., & Uckelmann, D. (2024). Evaluating Room Occupancy with CO2 Monitoring in Schools: A Student-Participative Approach for Presence-Based Heating Control. In International Conference on Science and Technology Education, p.pp.23-31.

Park, C., Chun, H., & Chi, S. (2024). Multi-Camera People Counting Using a Queue-Buffer Algorithm for Effective Search and Rescue in Building Disasters. KSCE Journal of Civil Engineering, p.pp.1-15.

Peng, Y., Li, Y., & Zheng, W. (2023). Revisiting Person Re-Identification by Camera Selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5), p.pp.2692-2708.

Riachy, C., Khelifi, F., & Bouridane, A. (2019). Video-based person re-identification using unsupervised tracklet matching. IEEE Access, 7, p.pp.20596-20606.

Sánchez, S. (2023). Bluetooth technology for people counting and research and rescue missions. Bachelor's thesis, Universitat Politècnica de Catalunya.

Shyaa, T., & Hashim, A. (2024). Enhancing real human detection and people counting using YOLOv8. In BIO Web of Conferences, 97, p.pp.61.

Tien, P., Wei, S., Calautit, J., Darkwa, J., & Wood, C. (2021). Vision-based human activity recognition for reducing building energy demand. Building Services Engineering Research and Technology, 42(6), p.pp.691-713.

Wu, G., Zhu, X. and Gong, S. (2020). Tracklet Self-Supervised Learning for Unsupervised Person Re-Identification, Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), p.pp.12362-12369.

Yang, B., Shan, Y., Peng, R., Li, J., Chen, S., & Li, L. (2022). A feature extraction method for person re-identification based on a two-branch CNN. Multimedia Tools and Applications, 81(27), p.pp.39169-39184.

Zhang, X., Wang, X., & Gu, C. (2021). Online multi-object tracking with pedestrian re-identification and occlusion processing. The Visual Computer, 37(5), p.pp.1089-1099.

Zheng, H., Lin, Z., Cen, J., Wu, Z., & Zhao, Y. (2018). Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation. IEEE transactions on circuits and systems for video technology, 29(3), p.pp.787-799.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision p.pp. 1116-1124.