

# Exploring visual techniques for indoor intrusion detection using detectron2 and Faster RCNN

**Changhao Dong**

International School of Information Science & Engineering, Dalian University of Technology, Dalian, 10141, China

1483108987@mail.dlut.edu.cn

**Abstract.** As societal evolution marches forward, there's an escalating emphasis on indoor security concerns. Within the safety realm, indoor intrusion detection emerges as a pivotal challenge, given its direct implications on safeguarding human lives and assets. Leveraging visual methods for indoor intrusion detection holds promising potential, particularly due to its straightforward deployment advantages. This study zeroes in on surveillance video footage, a prevalent medium in the security domain, as its experimental muse. Post an initial preprocessing phase utilizing a video difference map, the research introduces a pedestrian detection algorithm hinged on the synergy of detectron2 and Faster R-CNN. Insights gleaned reveal that this combined algorithm, when augmented with the video difference map preprocessing, exhibits commendable accuracy, robustness, and real-time efficacy on surveillance footage, especially in scenarios bereft of significant target occlusion. Moreover, this algorithm showcases an adeptness at discerning diminutively sized targets, demonstrating resilience against varying light magnitudes and maintaining impeccable accuracy amidst intricate lighting conditions. By harnessing this methodology, the enhancement of indoor environment safety monitoring becomes feasible, thereby bolstering the provision of dependable protection for individuals.

**keywords:** indoor intrusion detection, vision, pedestrian detection, deep learning, surveillance video.

## 1. Introduction

With societal evolution and progress, the focus on indoor safety has magnified considerably. Notably, in the realm of security, addressing indoor intrusions stands out as a critical challenge, given its direct implications for the safety of individuals and their assets. Over recent decades, the surge in urbanization coupled with population growth has witnessed an ascending trajectory in the frequency and intensity of such intrusion incidents.

Among the available indoor intrusion detection technologies, options span from infrared and GPS to vision-based systems and sensor networks. Detection mechanisms relying on infrared or GPS typically necessitate specific markers or equipment either on the target or within the system, posing material constraints that hinder widespread adoption. Deploying sensor network-based systems often demands extensive node device installations, translating to elevated costs and maintenance challenges, making sustainability and widespread promotion problematic. In stark contrast, vision-based indoor intrusion detection techniques stand out, characterized by minimal hardware prerequisites, ease of maintenance,

and straightforward deployment, making them apt for broader implementation. Additionally, given the escalating quality of surveillance footage and the strides in deep learning since its inception by Hinton et al. in 2006, vision-based techniques for surveillance video intrusion detection have attained commendable maturity.

This study harnesses surveillance video as its experimental medium. The Faster RCNN network, an amalgamation of RPN and Fast RCNN, processes the differential between consecutive video frames to accentuate areas of change and diminish the prominence of static zones. Subsequently, this modified image undergoes processing via the pretrained model of Detectron2, Facebook's open-source computer vision platform. The output is an image adorned with bounding boxes, highlighting potential areas of human presence, denoted with their respective probability percentages. This innovative approach paves the way for enhanced indoor intrusion detection, rendering it a viable contender for security applications..

## 2. Methods

### 2.1. *Detectron2 pre-training model*

In the intricate landscape of indoor intrusion detection, the choice of a model significantly influences overall system efficacy. Detectron2 pretrained models have been judiciously employed as the foundation for this intrusion detection exploration [1]. Such a decision emerges from comprehensive evaluations and considerations, aiming to drive exceptional results in this domain [2].

*2.1.1. Rationale for model selection.* From the diverse palette of object detection models on the market, Detectron2 stands out, primarily due to its laudable performance on the COCO dataset. This model's fusion of unparalleled accuracy and efficiency earmarks it as the ideal candidate for indoor intrusion detection tasks. Furthermore, Detectron2 embodies the cutting-edge in deep learning and computer vision, equipping it with the capability to accurately identify objects within intricate settings, a requirement that aligns perfectly with the unique challenges posed by intrusion detection.

*2.1.2. Advantages of the detectron2 platform.* Detectron2, an open-source computer vision library by Facebook AI Research, boasts several commendable advantages that equip it with exceptional adaptability in intrusion detection. Its openness and flexibility facilitate researchers in extending and enhancing it to suit diverse application scenarios. Detectron2 supports various model architectures, offering a wider scope for innovation in our research. Furthermore, the platform provides a rich set of preprocessing and post-processing tools that aid in handling input data and output results, thereby enhancing the efficiency and reliability of our research.

*2.1.3. Model application in intrusion detection.* Pretrained models from Detectron2 demonstrate significant promise in intrusion detection applications. When fine-tuned on a specific dataset, these models acquire the proficiency to discern intrusive behaviors. They autonomously extract pivotal information from surveillance footage, accurately identifying a range of targets and ensuring precise classification. Such enhancements elevate the efficiency and intelligence of intrusion detection systems. Irrespective of the indoor setting, this model astutely demarcates potential intrusion zones, equipping security teams with timely and trustworthy indicators essential for informed security actions.

### 2.2. *Introduction to Faster RCNN*

Faster R-CNN, an outstanding object detection network, demonstrates formidable potential in indoor intrusion detection [3]. Its efficient architecture and precise detection capabilities render it an essential component of our research.

*2.2.1. Working principle of Faster R-CNN.* Faster R-CNN operates as a deep learning-based object detection framework, succinctly characterized as "Region Proposal Network (RPN) + Fast R-CNN". In Faster R-CNN, the RPN plays a pivotal role, generating a sequence of candidate regions deemed likely to contain objects. Subsequently, Fast R-CNN utilizes these candidate regions for object classification and precise localization [4]. This two-stage design empowers Faster R-CNN to achieve efficient object detection without sacrificing accuracy.

Specifically, RPN employs a sliding window approach to generate anchor boxes on the image, representing potential object locations. Subsequently, RPN employs convolutional operations to determine whether each anchor box contains an object. This process employs anchor boxes of varying scales and aspect ratios to effectively capture objects of different sizes. During this phase, RPN learns the position offsets of anchor boxes and the scores for object classification. Consequently, based on these classification scores and positional information, RPN filters out high-quality candidate regions, which are likely to contain actual objects.

The Fast R-CNN stage utilizes the candidate regions generated by RPN and performs feature extraction using RoI pooling layers. RoI pooling maps candidate regions of different sizes to fixed-size feature maps, preserving spatial information. These features subsequently undergo classification and regression via fully connected layers, achieving object localization and classification. Fast R-CNN leverages RoI pooling and shared convolutional features among the RoIs, significantly enhancing computational efficiency.

*2.2.2. Synergy between RPN and Fast R-CNN.* The core strength of Faster R-CNN lies in the synergy between RPN and Fast R-CNN. RPN is responsible for generating candidate regions using anchor boxes to represent potential object locations, followed by classification and positional offset prediction through convolutional operations. This step provides crucial information for subsequent object detection, effectively reducing the number of candidate regions and improving computational efficiency.

Building upon the candidate regions generated by RPN, the Fast R-CNN stage extracts features using RoI pooling layers, followed by object localization and classification through classification and regression branches. The organic fusion of RPN and Fast R-CNN equips Faster R-CNN with efficient candidate region generation and accurate object detection capabilities [5]. This dual structure allows Faster R-CNN to achieve rapid object detection while maintaining high precision.

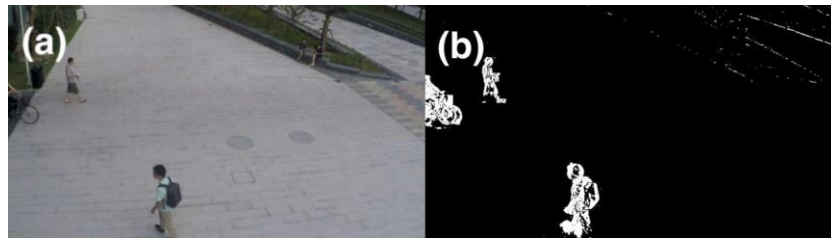
*2.2.3. Applicability of Faster R-CNN in intrusion detection.* The applicability of Faster R-CNN in intrusion detection is evident on multiple fronts. Firstly, it accurately delineates potential intrusion areas, providing crucial reference information to security personnel and enhancing response speed. Secondly, Faster R-CNN exhibits outstanding multi-object detection capabilities, simultaneously handling detection of multiple intruders, thereby enhancing the comprehensiveness of the monitoring system.

Moreover, the efficiency of Faster R-CNN enables real-time detection, offering robust support for preempting intrusive activities. In various indoor environments, Faster R-CNN consistently identifies potential intrusion areas, bolstering technical support for security management. Through its widespread application in intrusion detection, Faster R-CNN demonstrates significant potential in the field of indoor security, paving the way for more intelligent and secure indoor monitoring systems.

### *2.3. Video preprocessing*

Within the realm of indoor intrusion detection, the preprocessing of surveillance videos emerges as a crucial phase, instrumental in bolstering algorithmic precision and performance [6]. Through meticulously crafted preprocessing sequences, potential intrusion zones are accentuated, background distractions are minimized, and enriched data is channeled into subsequent object detection stages.

**2.3.1. Generation of video difference maps.** Generating video difference maps serves as a crucial preprocessing phase, aiming to discern variations between sequential frames within surveillance videos. This technique is rooted in a straightforward concept: by juxtaposing pixel values of successive frames, changes between these frames become evident. For every pixel, the difference between its values across two frames is computed, producing a pixel value that encapsulates the observed change. Multiple algorithms can facilitate this process. For instance, the absolute pixel value difference method provides one approach, while more intricate techniques rooted in optical flow track pixel modifications over durations, bolstering sensitivity to dynamic shifts. Irrespective of the chosen methodology, the outcome is typically a grayscale image wherein the extent of pixel value change signifies the prominence of distinct regions, as illustrated in Figure 1.



**Figure 1.** Comparison of video images before and after differencing (Photo/Picture credit: Original).

**2.3.2. Enhancing change regions with video difference maps.** Generating video difference maps effectively emphasizes regions experiencing alterations in surveillance footage, indicating potential areas of intrusion. Such alterations could represent moving individuals, shifting objects, or other unusual activities. Nonetheless, merely illuminating areas with changes doesn't sufficiently support precise object detection. Thus, the difference maps are integrated with the original video frames to offer a comprehensive view. This combination is realized through pixel-level overlay, visually amplifying the prominence of altered regions and offering enhanced clarity for the ensuing object detection phase [7].

**2.3.3. Increasing sensitivity and accuracy by attenuating invariant areas.** While video difference maps stand out in emphasizing areas of change, they can sometimes display variations in invariant zones due to external influences like shifting lighting conditions or static backgrounds. Such occurrences can trigger false alarms, undermining detection precision. To navigate this hurdle, an innovative approach is employed: diminishing the prominence of invariant regions to amplify the system's sensitivity and precision. This approach harnesses various image enhancement methods, from fine-tuning pixel intensities to leveraging advanced filters. By applying these techniques to the invariant zones, distractions stemming from these areas are significantly reduced, thereby honing the detection's focus on truly dynamic regions [8]. The generation and use of video difference maps accentuate evolving areas, yielding crucial intrusion data. Concurrently, by minimizing invariant regions, there's a marked improvement in detection accuracy, delivering superior inputs for ensuing object detection algorithms. These carefully orchestrated preprocessing stages collaboratively forge a formidable foundation ensuring the robustness and efficiency of indoor intrusion detection systems [9]. With such optimized preprocessing workflows, it becomes feasible to adeptly manage a wide array of intricate situations, ultimately elevating the efficacy of indoor security surveillance systems.

### **3. Result analysis**

The data for the experiment in this paper is the surveillance video of a factory. The factory surveillance video provides a variety of complex backgrounds and various target types [10], so it is possible to evaluate the ability of the algorithm to deal with complex environments, including knowing whether it can effectively distinguish the target from the background, and in this case accuracy. In order to reflect complex situations in the real world such as changing light conditions, target occlusion, dense targets, and small target scales, the performance of the test algorithm in these situations helps to identify the

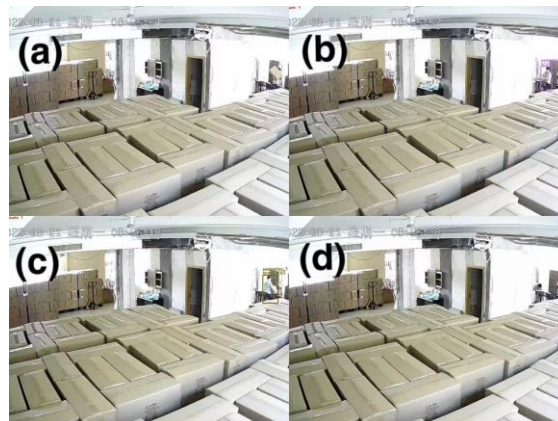
potential limitations of the algorithm in specific application scenarios. The author displays and analyzes the target detection results by testing the surveillance videos of Detectron2+FasterRCNN in different scenarios in the factory building.

The experimental results are shown in Figures 2 to 4, where the bounding boxes mark the position of the human body, and these bounding boxes are distinguished by different colors and line styles. "pedestrian" on the bounding box is the category label for the currently detected human body, and the value after "score" is the confidence that the detected object is the "pedestrian" category, that is, the probability that the object belongs to the "pedestrian" category. The upper left corner of the image is the relevant information for processing the image. "frame" indicates which frame of the video stream the image comes from, "fps" indicates the number of frames per second, indicating the frame rate for processing the frame, and "num" indicates the human body successfully detected by the detectron2+Faster R-CNN model in the current image frame, and "Total count" represents the total number of human bodies detected during the entire video stream processing process.



**Figure 2.** Results of stair landing object detection (Photo/Picture credit: Original).

Figure 2 shows the target detection results of Detectron2+FasterRCNN on the stairway scene. It can be seen that for stairwells where the target moves frequently and the target is low-density (such as Figure 2(a) and Figure 2(c)), detectron2+FasterRCNN can effectively locate and detect pedestrians, with an accuracy rate of nearly 100%. However, when the target is dense and severely occluded (such as Figure 2(b) and Figure 2(d)), the feature extraction ability of Detectron2+FasterRCNN is limited, and some pedestrians are missed.



**Figure 3.** Results of warehouse object detection (Photo/Picture credit: Original).

Figure 3 shows the target detection results of Detectron2+FasterRCNN on the warehouse scene. It can be seen that the algorithm performs relatively well for warehouses where the target size is small (as shown in Figure 3(b) and Figure 3(c)). However, when the target is severely occluded (as shown in Figure 3(a) and Figure 3(d)), the phenomenon of missed detection.



**Figure 4.** Results of square object detection (Photo/Picture credit: Original).

Figure 4 shows the target detection results of the algorithm on the square. It can be seen that when the illumination changes significantly, the accuracy of the algorithm does not change, and it has good robustness.

After the preprocessing of the video difference map, the Detectron2+FasterRCNN algorithm in this paper can accurately distinguish the target human body from other objects, and there is no false detection. However, due to the insufficient ability of the algorithm to extract the features of the target with a large degree of occlusion, the phenomenon of missed detection occurs.

## 4. Discussion

### 4.1. Feasibility analysis of indoor intrusion detection

Indoor intrusion detection requires algorithms to be accurate, robust and real-time, to avoid false detections and missed detections under complex conditions such as light changes, and to process images and videos in a short period of time. So as to achieve the purpose of timely detection of intrusion behavior.

When invading indoors, the intruding object usually moves frequently, and there are no large obstacles such as trees and buildings indoors, so the degree of human body occlusion is not large. Since the indoor camera is relatively far away from the human body, the target scale is relatively small. However, the human detection algorithm in this paper performs better in the case of "small target size", "frequent target movement" and "small target occlusion", so it is feasible to perform indoor intrusion detection in an indoor environment with few occluders.

### 4.2. Suggestions

But although occlusion is usually small, other items such as furniture may still cause some occlusion in indoor environments. When there is a large occlusion, the algorithm cannot take into account the requirements of "accuracy" and "real-time". It shows that in the case of dense pedestrians or severe occlusion, the algorithm feature extraction is limited, and there are cases where some targets are missed. Therefore, multiple feature extraction and feature fusion technology can be used for target occlusion, so as to better capture the target context information, and help to accurately detect the target when the target is occluded to improve the accuracy of the algorithm.

Not only that, in order to improve the real-time performance of the algorithm, it can be combined with YOLOv3 for human detection, first use YOLOv3 to perform preliminary detection of the target, perform target screening, and then extract the area where the target is located as the input of Faster R-CNN for refinement detection, and finally result fusion. In this way, "YOLOv3" with better real-time performance and "Faster R-CNN", which is more powerful in fine positioning, can be combined to ensure accuracy and real-time performance at the same time.



#### 4.3. Limitation analysis

It should be noted that since the data used in this study comes from the same factory, although the target is small, the flow of people is large, and the occlusion is serious, the data sample is still limited and there are limitations. First, factory environments are different from other indoor environments, and performance in this particular environment may not be directly applicable to other environments. Secondly, subtle differences in different indoor environments may affect the performance of the algorithm, so further verification of the robustness of the algorithm is needed. Furthermore, although the experiment involves severe occlusion, indoor intrusion detection may face more complex occlusion in reality.

#### 5. Conclusion

This paper proposes an indoor surveillance video intrusion detection algorithm based on the detectron2+Faster R-CNN algorithm after video difference map preprocessing. Through research, this paper finds that the detectron2+Faster R-CNN human detection algorithm after differential image processing on the video can show good accuracy, robustness and real-time performance in surveillance videos where human body occlusion is not serious. Due to the processing of the video differential image, the method can accurately distinguish between dynamic objects and static objects, and further can distinguish between human bodies and objects such as doors and windows, thereby realizing accurate human body detection. It shows that this method has a better processing for the human body detection of video with fixed shooting equipment such as surveillance video, and meets the needs of indoor real-time detection. This method can be used for indoor personnel intrusion detection, and can realize on-site real-time detection. Due to the small burden of hardware foundation, convenient maintenance and simple deployment, this method has a good application prospect. However, the human detection ability of this method needs to be improved when people are densely populated or under severe occlusion. Future research can improve the accuracy of detection by further refinement of human feature extraction and feature fusion when human body occlusion is serious or pedestrians are dense.

#### References

- [1] Farheen C, Harald K. Detectron2 for Lesion Detection in Diabetic Retinopathy[J]. Algorithms, 2023,16(3).
- [2] Bobomirzaevich A A, Siful M B I,Rashid N, et al. An Improved Forest Fire Detection Method Based on the Detectron2 Model and a Deep Learning Approach[J]. Sensors, 2023, 23(3).
- [3] He Y,Xiaotang W, Yuhan L, et al. A new face detection method based on Faster RCNN [J]. Journal of Physics: Conference Series,2021,1754(1).
- [4] Hua W, Shifa J, Yang G. Improved Object Detection Algorithm Based on Faster RCNN [J]. Journal of Physics: Conference Series,2022,2395(1).
- [5] Hu L,Wei C, Yang X, et al. Special faster-RCNN for multi-objects detection[P]. International Workshop on Pattern Recognition,2018.
- [6] Leyuan Z, Bihui L,Chuanhui L. SHIP target image recognition based on FAST detector and faster-RCNN[P]. Naval Aeronautical and Astronautical Univ. (China),2021.
- [7] Zhang X,Qiu Z, Jiao L, et al. A method for centroid extraction based on Faster-RCNN [P]. International Symposium on Multispectral Image Processing and Pattern Recognition, 2020.
- [8] Lin Z, Guo Z, Yang J. Research on Texture Defect Detection Based on Faster-RCNN and Feature Fusion [P]. Machine Learning and Computing,2019.
- [9] Wei-Jie X,Shuaiqi L, Fei-Hong Y, et al. Disturbance recognition for -OTDR based on Faster-RCNN [P]. Southern Univ. of Science and Technology of China (China), 2022.
- [10] Bao J, Wei S,Lv J, et al. Optimized Faster-RCNN in Real-time Facial Expression Classification [C]//Advanced Science and Industry Research Center. Proceedings of 2019 2nd International Conference on Communication,Network and Artificial Intelligence(CNAI 2019). IOP Publishing, 2019:1012-1019.