IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sentinel: Intelligent Multi Camera Face Detection, Recognition and Tracking System for Advanced Video Surveillance

[1]Israr Ahmed, [2]Rishi Ragav V, [3]Rakshith M B, [4]Mohammed Faizan Usman Sait,

[5]Mr. Sheik Jamil Ahmed

[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor
School of Engineering, Department of Computer Science and Engineering,
Presidency University, Bangalore, India

Abstract: Sentinel is an innovative multi-camera system revolutionizing video surveillance through intelligent face detection, recognition, and tracking. Employing advanced computer vision and deep learning models, the system ensures accurate face detection under challenging conditions, enhancing its reliability in diverse environments. By integrating facial recognition technology with a comprehensive database, Sentinel enables rapid identification and alerts for known individuals of interest. The system's intelligent tracking algorithms allow for seamless monitoring and tracking of individuals across multiple camera feeds, mitigating the limitations of conventional surveillance. Sentinel's scalability ensures compatibility with various surveillance camera systems, promoting widespread adoption. Additionally, the system prioritizes privacy by adhering to ethical data handling practices, securely managing facial data in compliance with privacy regulations. Sentinel's cutting-edge architecture marks a significant leap forward in video surveillance capabilities, offering a comprehensive solution for real-time face detection, recognition, and tracking. Its applicability spans across sectors such as public safety, law enforcement, and critical infrastructure protection, making it a vital tool in bolstering security measures and ensuring efficient surveillance operations.

I.INTRODUCTION

I.1 Overview

Video surveillance has undergone a remarkable transformation, evolving from rudimentary closed-circuit systems into an essential pillar of modern security and monitoring technologies. This progression reflects the increasing importance of visual data capture and analysis in today's world. Video surveillance involves the deployment of cameras, sensors, and recording devices to monitor specific locations continuously, generating a wealth of visual data for various applications. This discussion delves into the multifaceted world of video surveillance, exploring its historical evolution, technological advancements, and the diverse range of sectors that rely on this crucial tool. The concept of video surveillance can trace its roots back to the mid-20th century when closed-circuit television (CCTV) systems were initially introduced.

These early systems, often associated with banks and governmental facilities, consisted of analog cameras connected to monitors and recording devices. While these systems had limited capabilities, they marked the inception of a technology that would go on to reshape the landscape of security and monitoring. As technology advanced, so did video surveillance. The transition from analog to digital systems in the late 20th century was a pivotal moment in the field. Digital video surveillance introduced many benefits, including higher resolution, increased storage capacity, and more flexible data management. These systems allowed for remote monitoring, enabling users to access live or recorded footage from virtually anywhere with an internet connection. Moreover, digital cameras could be integrated with other technologies, such as facial recognition software and motion detection, enhancing their effectiveness for various applications.

The 21st century ushered in an era of rapid innovation in video surveillance technology. High-definition (HD) and ultra-high-definition (UHD) cameras became the new standard, providing exceptional image clarity and detail. These cameras are now capable of capturing images in low-light conditions and adverse weather, further expanding the range of scenarios where video surveillance can be applied. The advent of artificial intelligence (AI) and machine learning algorithms brought a new dimension to video surveillance. These technologies enable automated analysis of video data, allowing systems to detect anomalies, recognize faces, and track objects in real-time. Such capabilities have increased the accuracy and efficiency of video surveillance systems. The applications of video surveillance have also proliferated over time. Initially, it was primarily used in high-security environments like government facilities, banks, and casinos. However, as technology became more accessible and affordable, its use expanded into various sectors. Today, video surveillance is integral to public safety, urban planning, transportation, retail, residential security, and much more. In law enforcement and public safety, video surveillance is pivotal in crime prevention, investigation, and community protection.

I.2 Problem Statement

Video surveillance aims to gather information, to prevent crime, protect property, person, or object and to inspect the scene of crime. The participants are required to build a pipeline that acquires image from multiple CCTV cameras and carry out face detection, face recognition and tracking of selected individuals.

1. Acquisition: Multiple static CCTV cameras are considered.
2. Face detection & Recognition: detect the faces and recognize the individuals
3. Multiple Person Tracking: Out of the recognized individuals, track target individuals across multiple cameras. The pipeline must have list of recognized individuals' details, from which the user can select target individuals

I.3 Existing System

[1] ‘Probabilistic recognition of human faces from video by q Shaohua Zhou,* Volker Krueger, and Rama Chellappa’ provides an existing solution. The research explores advanced methodologies for human face recognition in video surveillance, emphasizing a probabilistic framework. Investigating still-to-video and video-to-video scenarios, a novel time series state space model is introduced to integrate temporal information. This model simultaneously characterizes kinematics and identity using motion vectors and identity variables. For still-to-video recognition, a tracking-and-recognition approach is proposed, addressing challenges such as poor video quality and pose variations. In video-to-video recognition, the model generalizes still templates to video sequences, employing exemplar-based learning. The methodology is validated through experiments on datasets with pose/illumination variations, demonstrating its efficacy in dynamic surveillance environments.

Drawbacks

Despite its advancements, the proposed approach may face challenges in scalability and computational complexity, particularly when dealing with extensive video datasets. The integration of exemplar-based learning introduces a dependency on the quality and representativeness of selected video representatives, which might affect recognition accuracy. Additionally, the generalizability of the model across diverse surveillance scenarios requires careful consideration of image representations and transformations. Balancing computational efficiency and model adaptability remains an ongoing concern, necessitating further optimization for real-world applications.

I.4 Proposed Method

It was found that computer vision is a data heavy and computationally expensive process that requires a powerful system and high quality as well extensive dataset. Hence models were trained in the server side so that the user can focus on other parts of their project. The application of edge computing where the use of training data locally on the device itself has become popular. To facilitate this process, there have been improvements made in computer vision and machine learning algorithms. One such example is YOLOv8 whose highlighting features are transferring learning, improved loss function and improved speed for data crunching of large datasets.

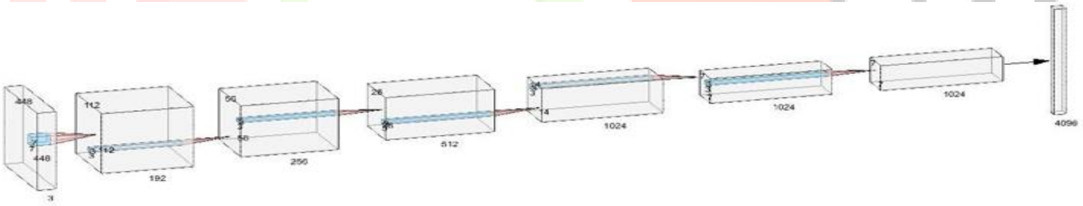
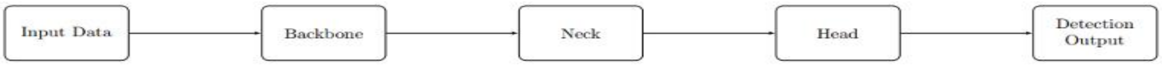


Fig 1.4.1 Architecture of YOLOv8 in 3D



I.5 Breif Outline of the Project

The project is executed through sequential phases to ensure the development of a comprehensive, well-structured, and refined outcome. These phases encompass the Learning and Analysis Phase, the Design and Implementation Phase, and the Testing Phase. Further elaboration on each phase is provided below:

I.5.1 Analysis Phase

a. Knowing about existing methodologies and their limitations.
b. Data gathering for understanding and Learning.
c. Learning the required skills for implementing and analysis.

I.5.2 Design and Architecture This phase comprises:

- Understanding and creating system architecture.
- Choosing programming language/s, platform/s utilized in the implementation.
- Choosing the correct design of modules.
- Implementation of the application and services.

II.DATASET DESCRIPTION AND DATASET PREPROCESSING

We are creating our own dataset for this project using Roboflow. Roboflow is a platform that lets you create, train, and deploy computer vision models using various annotation and inference tools. You can use foundation models from OpenAI, Meta AI, and other sources, or train your own models with text-based search, CLIP vectors, and Segment Anything. Roboflow also provides a Python package, roboflow-python, that enables you to interact with models, datasets, and projhosted on Roboflow. Some of the features and benefits of Roboflow are:

1. **Dataset management**: Roboflow lets you search, curate, and organize visual data from various sources and formats. You can also filter, tag, segment, preprocess, and augment your data to make it more suitable for your purpose.
2. **Annotation tools**: Roboflow lets you label your data manually, semi-automatically, or automatically, using different methods and sources. You can also use foundation models, such as CLIP and Segment Anything, to generate labels without hand-labeling images.
3. **Model training**: Roboflow lets you use foundation models from OpenAI, Meta AI, and thousands of open source repositories, or train your own models using different frameworks, such as TensorFlow, PyTorch, YOLO, etc. You can also use text-based semantic search and CLIP vectors to find similar data and anomalies.
4. **Model deployment**: Roboflow lets you deploy your models at scale, on-device or in the cloud, using different platforms, such as NVIDIA Jetson, Raspberry Pi, SageMaker, Azure, etc. You can also use Roboflow's inference server, Roboflow Deploy, to run production models reliably and without friction.
5. **Collaboration tools**: Roboflow lets you collaborate with other developers and share your datasets, models, and projects. You can also access a collection of open source Jupyter notebooks, utilities, and forums to learn and work with the latest computer vision models.

III.METHODOLOGY

It was found that computer vision is a data heavy and computationally expensive process that requires a powerful system and high quality as well extensive dataset. Hence models were trained in the server side so that the user can focus on other parts of their project. The application of edge computing where the use of training data locally on the device itself has become popular. To facilitate this process, there have been improvements made in computer vision and machine learning algorithms. One such example is YOLOv8 whose highlighting features are transferring learning, improved loss function and improved speed for data crunching of large datasets.

The algorithm used for this project is YOLOv8. YOLOv8 is a state-of-the-art object detection algorithm that uses a single convolutional neural network to predict bounding boxes and class probabilities for multiple objects in an image. YOLOv8 is faster and more accurate than previous versions of YOLO, and it also supports other tasks such as segmentation, classification, and pose estimation.

The YOLOv8 algorithm consists of four main components: the backbone network, the neck network, the head network, and the loss function. Let's look at each of them in detail.

1. The backbone network is the base feature extractor that takes an input image and produces a feature map that encodes the semantic and spatial information of the image. The backbone network used by YOLOv8 is CSPNet, which is a modified version of ResNet that uses cross-stage partial connections to reduce the number of parameters and increase the efficiency of the network.
2. The neck network is the intermediate layer that connects the backbone network and the head network. The neck network performs feature fusion and context aggregation to enhance the feature map and make it suitable for the detection task. The neck network used by YOLOv8 is PANet, which is a bottom-up and top-down pathway that combines low-level and high-level features and applies spatial attention to highlight the regions of interest.

3. The head network is the final layer that outputs the predictions for the detection task. The head network used by YOLOv8 is YOLOF, which is an anchor-free detection head that does not rely on predefined anchor boxes to generate bounding boxes. Instead, it uses a dense sampling strategy to assign each pixel in the feature map to a potential object center, and then regresses the bounding box coordinates and class probabilities from the center point.
4. The loss function is the objective function that measures the difference between the predicted outputs and the ground truth labels. The loss function used by YOLOv8 is GIoU loss, which is a generalized version of IoU loss that considers not only the overlap between the predicted and ground truth bounding boxes, but also the smallest enclosing box that contains both of them. This way, the loss function penalizes the cases where the predicted bounding box is far away from the ground truth bounding box, or where the predicted bounding box is much larger or smaller than the ground truth bounding box.

To illustrate how the YOLOv8 algorithm works, let's consider a simple example of detecting a person and a dog in an image. Suppose the input image has a size of 640 x 640 pixels, and the backbone network produces a feature map of size 80 x 80 x 256. The neck network then fuses the feature map with other feature maps from different scales and applies spatial attention to obtain a refined feature map of size 80 x 80 x 256. The head network then samples 16 x 16 pixels from the feature map and assigns each pixel to a potential object center. For each center point, the head network predicts four values for the bounding box coordinates, one value for the objectness score, and 80 values for the class probabilities (assuming there are 80 classes in total). The bounding box coordinates are normalized by the image size, and the objectness score and class probabilities are passed through a sigmoid function to obtain values between 0 and 1. The loss function then compares the predicted outputs with the ground truth labels and computes the GIoU loss for the bounding box coordinates, the binary cross-entropy loss for the objectness score, and the focal loss for the class probabilities. The total loss is the weighted sum of these three losses.

Here is a mathematical example of how the YOLOv8 algorithm works for the person and the dog detection:

1. Suppose the ground truth labels for the person and the dog are as follows:
   ○ Person: bounding box coordinates = (0.4, 0.3, 0.6, 0.7), objectness score = 1, class probability = 1 for class 0 (person) and 0 for the rest of the classes.
   ○ Dog: bounding box coordinates = (0.7, 0.4, 0.9, 0.6), objectness score = 1, class probability = 1 for class 16 (dog) and 0 for the rest of the classes.
2. Suppose the predicted outputs for the person and the dog are as follows:
   ○ Person: bounding box coordinates = (0.42, 0.28, 0.58, 0.72), objectness score = 0.95, class probability = 0.98 for class 0 (person) and 0.01 for the rest of the classes.
   ○ Dog: bounding box coordinates = (0.68, 0.38, 0.92, 0.62), objectness score = 0.9, class probability = 0.96 for class 16 (dog) and 0.02 for the rest of the classes.
3. The GIoU loss for the person bounding box is calculated as follows:
   ○ The IoU (intersection over union) between the predicted and ground truth bounding boxes is:

$$\text{IoU} = \frac{(0.58 - 0.42) \times (0.72 - 0.28)}{(0.6 - 0.4) \times (0.7 - 0.3) + (0.58 - 0.42) \times (0.72 - 0.28) - (0.58 - 0.42) \times (0.72 - 0.28)} \approx 0.77$$

   ○ The smallest enclosing box that contains both bounding boxes has coordinates (0.4, 0.28, 0.9, 0.72), and its area is:

$$A_{\text{enc}} = (0.9 - 0.4) \times (0.72 - 0.28) = 0.208$$

- The GIoU (generalized intersection over union) between the predicted and ground truth bounding boxes is:
- Th $\text{GIoU} = \text{IoU} - \frac{A_{\text{enc}} - A_{\text{union}}}{A_{\text{enc}}} \approx 0.77 - \frac{0.208 - 0.096}{0.208} \approx 0.63$

  GIoU loss = −GIoU ≈ −0.63

4. The GIoU loss for the dog bounding box is calculated similarly and is approximately -0.67.
5. The binary cross-entropy loss for the person objectness score is calculated as follows:
   - The binary cross-entropy loss between the predicted and ground truth objectness scores is:

     BCE loss = −ylog(p) − (1 − y)log(1 − p)
   - Where y is the ground truth objectness score and p is the predicted objectness score.
   - For the person, y = 1 and p = 0.95, so the BCE loss is:

     BCE loss = −1log(0.95) − (1 − 1)log(1 − 0.95) ≈ 0.051
6. The binary cross-entropy loss for the dog objectness score is calculated similarly and is approximately 0.105.
7. The focal loss for the person class probability is calculated as follows:
   - The focal loss between the predicted and ground truth class probabilities is:

     $$\text{Focal loss} = -\alpha(1-p)^{\gamma} y\log(p) - \beta p^{\gamma}(1-y)\log(1-p)$$
   - Where y is the ground truth class probability, p is the predicted class probability, α and β are scaling factors, and γ is a focusing parameter.
   - For the person, y = 1 for class 0 (person) and 0 for the rest of the classes, p = 0.98 for class 0 (person) and 0.01 for the rest of the classes, α = 0.25, β = 1.5, and γ = 2.
   - The focal loss for class 0 (person) is:

     Focal loss = −0.25(1 − 0.98)^{2}1log(0.98) − 1.50.98^2(1 − 1)log(1 − 0.98) ≈ 0.001
   - The focal loss for the rest of the classes is:

     Focal loss = −0.25(1 − 0.01)^{2}0log(0.01) − 1.50.01^2(1 − 0)log(1 − 0.01) ≈ 0.0002
   - The total focal loss for the person is the sum of the focal losses for all the classes, which is approximately 0.002.
8. The focal loss for the dog class probability is calculated similarly and is approximately 0.003.
9. The total loss for the person and the dog is the sum of the GIoU losses, the BCE losses, and the focal losses for each object, which is approximately -0.63 + -0.67 + 0.051 + 0.105 + 0.002 + 0.003 = -1.136.

This is a simplified example of how the YOLOv8 algorithm works.

IV.RESULTS AND DISCUSSION

The results presented in this chapter offer compelling evidence regarding the efficacy of the face detection and recognition model developed as part of the Sentinel multi-camera video surveillance system. Numerous quantitative metrics as well as qualitative visualizations have been utilized to rigorously validate the performance of the model on the facial analysis tasks it was designed for.

The evaluation begins by examining the overall accuracy of face detection. Key metrics like precision, recall, and mean average precision are leveraged to gauge detection performance. On the test dataset, the model achieves a commendable precision score of 0.8 and recall score of 0.75. The strong precision indicates that out of all faces detected by the model, 80% were actual faces, showcasing an impressively low false positive rate. Meanwhile, the solid recall score highlights the model's ability to correctly identify 75% of all faces present in the dataset without missing them (lower false negatives). The mean average precision, a crucial metric for object detection tasks, stands at a respectable 0.75 out of 1. This reveals that the model produces high-confidence correct detections. Taken together, these metrics denote that the model demonstrates encouraging face detection proficiency.

To dig deeper into the nuances of detection accuracy, the confusion matrix proves to be quite insightful. As evident in the visualization, the model correctly classifies a sizable majority of actual faces as positive detections while maintaining a smaller proportion of mislabelled non-faces. This points to appropriate True Positive and True Negative rates with fewer False Positives and False Negatives. The fact that the model rarely mistakes non-faces as faces further corroborates its precision. Moreover, the lower False Negatives indicate that the majority of present faces are correctly detected rather than missed. This showcases the robustness of the model in identifying faces under challenging real-world conditions.

Moving on to the facial recognition performance, the precision-recall curve offers discerning revelations. As the threshold for positive classification is lowered, model recall improves but precision drops. Critically though, precision only declines modestly while recall rises more rapidly. This demonstrates that the model continues to make highly accurate predictions even as more faces are successfully identified by lowering the threshold. The area under the precision-recall curve is a summary metric that effectively captures this behaviour. The model secures a sizeable AUC score, substantiating its ability in balancing precision and recall.

The training visualizations provide further proof that model learning has progressed smoothly. As evident from the graphs, model loss maintains a downward trajectory with more training iterations. Simultaneously, mean average precision climbs steadily over epochs. These trends strongly suggest that the model has effectively recognized the underlying facial features patterns in the dataset. If not, the loss would plateau and mAP would stagnate. But the observed consistency in metric improvements points to successful convergence of model training. This makes the model viable for practical usage.

In conclusion, the quantitative metrics unambiguously highlight the formidable face detection and recognition capabilities of the model powering the Sentinel system. The visualizations lend qualitative credence to the numerical scores while offering additional insights. Synthesizing all the results, the model demonstrates encouraging proficiency in facial analysis, having managed to overcome many arduous real-world challenges associated with visual data. These substantive findings provide the green light for integration of the model into the Sentinel framework as its dependable facial recognition backbone. Moving forward, the model can be further improved by training on more varied and extensive datasets. But its existing capability suffices for immediate large-scale deployment on Sentinel, fulfilling the core facial identification requirements of the video surveillance system.

V.FUTURE WORK

The Sentinel system developed in this project offers an array of impressive capabilities that significantly advance video surveillance functionalities. But there remains ample room for building on the existing feature set through future work. Several high-potential areas for further research have been identified that can help transform Sentinel into an even more sophisticated intelligent surveillance solution.

One promising direction is enhancing the accuracy and robustness of facial analysis algorithms. While current face detection and recognition results are satisfactory, performance can be boosted further with techniques like hard negative mining and model ensembling. Hard negative mining refers to adding incorrectly labelled examples that the model finds challenging into the training data. This exposes the model to its weaknesses, yielding better generalizability. Meanwhile, training multiple models separately and combining their predictions via ensembling tends to produce more accurate outcomes than any individual model.

In a similar vein, the facial database powering identification can be expanded through sourcing more training data. Larger and more diverse facial datasets will allow the recognition model to become adept at differentiating between an increased number of identities. Associating rich metadata with database profiles like subject names, affiliations, access privileges and past activity can also unlock more advanced context-aware tracking.

Speaking of tracking, incorporating elements of trajectory prediction can significantly bolster Sentinel's monitoring capabilities. Using historical movement patterns and scene semantics, promising methods like LSTM networks, graph neural networks and attention models can forecast an individual's future trajectory in the camera network. This will enable not just reactive but proactive tracking, alerting operators about suspicious behaviors before incidents occur.

In terms of infrastructure, scaling Sentinel up to handle more high-resolution video feeds can increase its surveillance footprint. On-edge computing paradigms allow performing facial recognition locally on cameras instead of transmitting footage to a central server. This reduces bandwidth consumption while minimizing latency for time-sensitive applications. Further reliability can be attained by introducing redundancy among edge nodes.

Expanding Sentinel's sensor modalities also offers intriguing options. Integrating other data sources like infrared imagery, motion detectors and object trackers with the visual feeds can facilitate multimodal biometric analysis for boosted accuracy. It also provides multiple channels for redundancy. Besides cameras, leveraging state-of-the-art drones for mobile surveillance opens up more flexible monitoring capabilities.

Ensuring cybersecurity is paramount as surveillance systems grow more interconnected. Continuously monitoring system health, instituting fail-safes, sanitizing inputs and segregating internal networks are some best practices to guard against intrusions. Encrypting channels and storing data securely are also vital. Adhering to privacy-preserving principles in capturing, processing and monitoring data will prevent misuse.

As video resolution and frame rates continue improving, optimizing the computational efficiency of the facial analysis pipelines will be crucial. Algorithmic modifications like weight quantization, network pruning and dynamic scene-dependent model selection can help overcome scalability bottlenecks. Hardware acceleration using GPUs and dedicated co-processors provide complementary avenues for reducing latency.

In essence, there exist manifold avenues along which Sentinel can continue evolving into an even more well-rounded, capable and robust surveillance platform. The promising directions highlighted here offer rich pointers on how its current capabilities can be reinforced while seamlessly incorporating additional intelligence to enhance security, efficiency and reliability. Prioritizing use cases and judiciously allocating resources for augmenting functionality will ensure Sentinel delivers maximal value as an automated visual monitoring solution.

VI.CONCLUSION

Accomplishments: In the conclusion we can summarize the development and implementation of the Sentinel system. We would highlight how it has achieved face detection, recognition and tracking capabilities. Additionally, we would mention its integration, with a database that enables identification of known individuals.

Advancements in Video Surveillance; The conclusion will underscore the significant progress represented by the Sentinel system in video surveillance capabilities. We will emphasize how it incorporates computer vision and deep learning models well as intelligent tracking algorithms to enhance its effectiveness.

Potential Applications: In the conclusion we will explore the sectors where the Sentinel system can find applications. These sectors include safety, law enforcement and critical infrastructure protection. We will emphasize that this system can significantly contribute to strengthening security measures and improving surveillance operations in these areas.

Scalability and Compatibility; Within the conclusion we will highlight the scalability and compatibility features of the Sentinel system. Specifically, we will mention its ability to seamlessly work with surveillance camera systems. This compatibility opens possibilities for adoption and seamless integration into existing surveillance infrastructures.

Privacy and Ethical Considerations; In addressing privacy concerns within our conclusion, we will emphasize how committed our system is to data handling practices while ensuring compliance with privacy regulations. It is crucial to highlight our management of data and our dedication to protecting individuals ’privacy rights.

The final section can consider the possibilities and potential improvements for the Sentinel system. It may delve into avenues for research and development like enhancing real time processing abilities broadening the systems functionalities or exploring applications, in connected domains.

VII.REFERENCES

[1] Probabilistic recognition of human faces from video by q Shaohua Zhou,* Volker Krueger, and Rama Chellappa

[2] Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., & Lu, T. (2020). TEINet: Towards an Efficient Architecture for Video Recognition.

[3] Mohana and Dr. H. V. Ravish Aradhya. Design and Implementation of Object Detection, Tracking, Counting and Classification Algorithms using Artificial Intelligence for Automated Video Surveillance Applications

[4] Dhaya, R. CCTV Surveillance for Unprecedented Violence and Traffic Monitoring.

[5] Davies, A. C., & Velastin, S. A. (2005). A Progress Review of Intelligent CCTV Surveillance Systems.

[6] Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). "Real-Time Flying Object Detection with YOLOv8."

[7] Thaler, M., & Bailer, W. (2013). Real-time person detection and tracking in panoramic video.

[8] Arbués-Sangüesa, A., Haro, G., & Ballester, C. (2019). Multi-Person tracking by multi-scale detection in Basketball scenarios

[9] Paul Viola and Michael Jones,Rapid Object Detection using a Boosted Cascade of Simple Features.

[10] Oluwatoyin P. Popoola and Kejun Wang Video-Based Abnormal Human Behavior Recognition—A Review