Suspicious People Detection and Tracking in Thermal Video Using Machine Learning: A survey

Volume :6 year: 2024

مجلة العلوم والتطـبيقات الهـندسية

Journal of Science and Engineering Application

ISSN 2521-3911

Suspicious People Detection and Tracking in Thermal Video Using Machine Learning: A survey

by

Hassanein yarob albakaa

Hassanein.albakaa@gmail.com

Ali Abdulkarem Habib Alrammahi

alia.alramahi@uokufa.edu.iq

Abstract

In recent years, the global rise in crimes involving weapons has become more pronounced, particularly in areas with weak law enforcement or where firearm possession is legal. To address this issue, early identification of suspicious behavior related to weapon possession is crucial, enabling law enforcement agencies to take swift action. While the human visual system is highly advanced and capable of processing images quickly and accurately, prolonged observation of similar visual data can lead to fatigue and reduced attentiveness. Additionally, large-scale surveillance systems with numerous cameras require extensive monitoring teams, which drives up operational costs. Several automatic weapon detection solutions based on computer vision have been proposed, but their performance remains limited in challenging environments. A systematic review of current literature on deep learning-based weapon detection was conducted to assess the methods employed, the characteristics of existing datasets, and the key challenges in automatic weapon detection. The most frequently utilized models were Faster R-CNN and YOLO architectures, and the integration of realistic images with synthetic data showed improved detection accuracy. Major challenges include poor lighting conditions and difficulties in detecting small weapons, with the latter being particularly significant. The focus of this review is to examine the methods used for detecting and tracking weapons using deep learning techniques.

Keyword: RGB Video_ Suspicious People dataset, object detection algorithms, object tracking algorithms.

1. Introduction

The global rise in crime rates, particularly due to the frequent use of handheld weapons during violent incidents, poses significant challenges. For a country to advance, maintaining law and order is

essential. In addition to drawing in investors, a tranquil and safe atmosphere is essential for a flourishing tourism sector. The problem of gun-related criminality is most acute in areas where it is lawful to own firearms[1].

Intelligent Surveillance System (ISS) is an autonomous system used to monitor buildings, airports, and public places for surveillance operations. Such surveillance tasks cover detection and tracking of objects (vehicles, people, etc.) in real-time as well analysis and actions to be taken thereafter automatically. This involves image and signal processing techniques as well as artificial intelligence or more specifically, machine learning, in the development of such intelligent systems. The most common surveillance system is visual cameras like CCTV. They are commonly thought about as a method of watching on properties, locations, individuals or events. Much of this work has been reviewed in several focused review papers[2-5].

In one color scheme, thermal video is run through a machine learning algorithm to make it easier for police officers in the field to detect weapons. Over the years various methods have been proposed to enhance performance of different type of surveillance systems[6]. These systems have come to the forefront due to security threats at airports. Among these thermal imaging has proven to be the best tool since it can locate heat signatures in low visibility such as smoke, fog or darkness. This is especially beneficial when using thermal cameras for security as they spot objects based on the heat their bodies radiate, a characteristic that normal day or night vision built-in with typical visible light CCTV cannot do[6].

The idea of humans in thermal video footage, especially those with a weapon, is now one of the main research areas for this type of technology within the industry that deals with surveillance and security. This literature review aims to give an overview of the recent advancements, methodologies, and issues in machine learning algorithms and deep learning algorithms that are used for this purpose.

1.1 Advances in Machine Learning Techniques

Recent studies have also illustrated the ability of deep learning models, specifically convolutional neural networks (CNNs), to differentiate between people and weapons in thermal images. The detection performance is also significantly improved with temporal cues incorporated into these models. An example of this is Temporal Convolutional Networks (T-CNNs) that use thermal image sequences and keep spatial frames consistent[7].

Real-Time Detection Systems: Some architectures such as Faster R-CNN and You Only Look Once (YOLO) have been common for real-time weapon detection. It is optimized to quickly process video feeds with high accuracy by being as fast and simple in terms of used forms as possible. Recent versions of YOLO now collect higher performance compared to other approaches in practical localizing and categorization guns[8].

Performance Metrics: Some research studies show that performance metrics such as recall, f1 result and precision are the major evaluation methods for these models to be compared to cont. number of other variables factors like how ambient temperature effects on accuracy detection[8].

1.2 Challenges in Detection

Although the researchers have made headway, there are still very big obstacles in finding people who might be concealed carry-Ing.:

Clothing choice: The style of clothing selected by people affects the thermal signatures that can be picked up through thermal imaging with respect to guns. Such as loose-fitting clothing that may camouflage the thermal signature of concealed firearms, resulting in false negatives and missed detections[7]. Being outdoor cameras, heat sensors are also very dependent on environmental factors i.e. lighting, ambient temperature etc. Many of these can introduce interference in thermal images that make the detection process complicated[8]

Limitations of the Dataset: One of the challenges we faced was – as with many other machine learning tasks —is satisfactory variety in data limiting the power of trained model. The other issue we see with most existing datasets is that they are limited to a specific set of scenarios and weapons making it hard for the models trained on such data to generalize well across different contexts[7, 8].

1.3 Weapon detection methodologies

Recent new techniques have been suggested to enhance the weapon detection capabilities:

1. **AI-Powered Video Surveillance:** In this case, there is somewhat complex software that needs to run and real-time analysis of video streams from IP cameras (which are hardware supported by AI-powered software). Sophisticated algorithms trained on massive data sets allow it to quickly and accurately detect potential weapons, leading drastically shorter response times as well correcting system error in identification[9].
2. **Deep Learning Models:** Convolutional Neural Networks (CNNs) have been widely used for object detection tasks, especially to distinguish between weapons and non-weapon objects placed in a challenging background with unwanted items[10]. By utilizing the DISARM and Gun datasets, as well as other pertinent data sources, these freeze models were fine-tuned on a rich resource of training data that makes them more practically deployable in real-life settings[10].
3. **Weapon Detection Object Detection Frameworks**: YOLO (You Only Look Once) optimized for real-time, to better detect weapons; with minimum false positives. In order to handle some issues such Mismatch of Scales (scale mismatches between training datasets and real-world events), techniques like Scale Match have been developed[10].

4. **Integrated Detection Systems:** Combining different methods in an effort to enhance the detection and reduce false positives. Such as using temporal analyses under a second classifier[10].
5. **Simulated datasets**: such as the one created in a game engine like Unity, which gives long-term training without needing any annotation data contribution or other ways to scale-up for reliable detection models[10].

2. Object Detection Algorithms

Object detection focuses on identifying and localizing objects within images or video frames. There are two main categories of object detection algorithms[11]. as shown in figure (1).

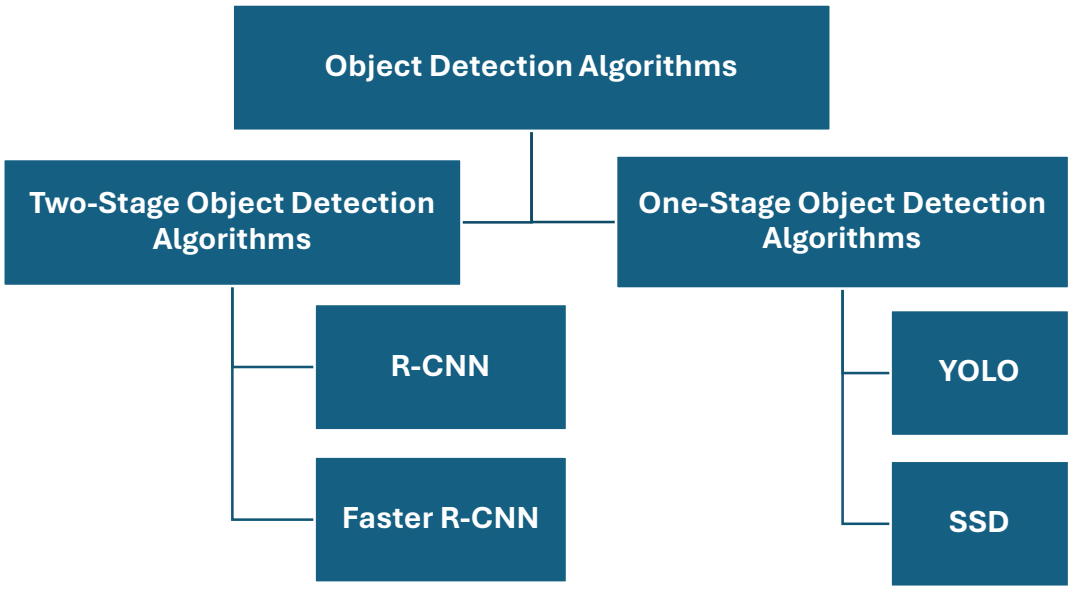

Figure 1: Object Detection Algorithms Architecture

2.1 Two-Stage Object Detection Algorithms

These Algorithms typically involve a two-step process: generating region proposals followed by classifying these proposals. Notable examples include:

A - R-CNN (Region-based Convolutional Neural Networks) Extracts regions from images and classifies them using CNNs, but it is computationally expensive [6].To recognize objects in images, it

employs a multi-stage method that blends convolutional neural networks (CNNs) with region proposal algorithms.as shown in figure (2).

- **General Structure of R-CNN**
  - **First phase: Region Proposal:** begin, create regional proposals or possible bounding boxes encircling the image's objects. Usually, algorithms like selective search are used for this, segmenting the image according to characteristics like texture and color to create roughly 2000 candidate sections [11].
  - **Second phase: Feature Extraction:** To extract feature vectors, each suggested region is scaled to a specified dimension and fed through a CNN (often a trained model like AlexNet)[12].
  - **Third phase: Classification:** To categorize each region as either an object or a background, the extracted features are input into a set of series of binary Support Vector Machines (SVMs).
  - **Fourth phase: Bounding Box Regression:** R-CNN uses bounding box regression to refine the predicted bounding boxes, modifying the initial region suggestions to better fit real objects[11].
  - **Fifth phase: Non-Maximum Suppression (NMS):** Lastly, superfluous overlapping boxes are removed using NMS, leaving only the most reliable object detections[11].

- **Advantages of R-CNN**
  - **High Accuracy**: R-CNN provides precise object localization and classification due to its detailed region-based processing.
  - **Robustness**: It effectively handles variations in object size, orientation, and complex backgrounds, making it adaptable to real-world scenarios[12].
  - **Flexibility**: The architecture can be modified for various tasks, such as instance segmentation and object tracking[12].

- **Disadvantages of R-CNN**
  - **Computational Complexity**: The multi-stage process is computationally intensive, requiring significant processing power and time[12].
  - **Slow Inference Speed**: Due to its sequential nature—processing each region independentlyR-CNN has slow inference times, making it unsuitable for real-time applications[11].
  - **Rigid Region Proposals:** Depending too much on the Selective Search method during the proposal generation phase can result in inefficiencies because the system does not learn from data adaptively[13].

- **Memory Usage**: Storing feature maps for each region proposal increases memory requirements during training[11].

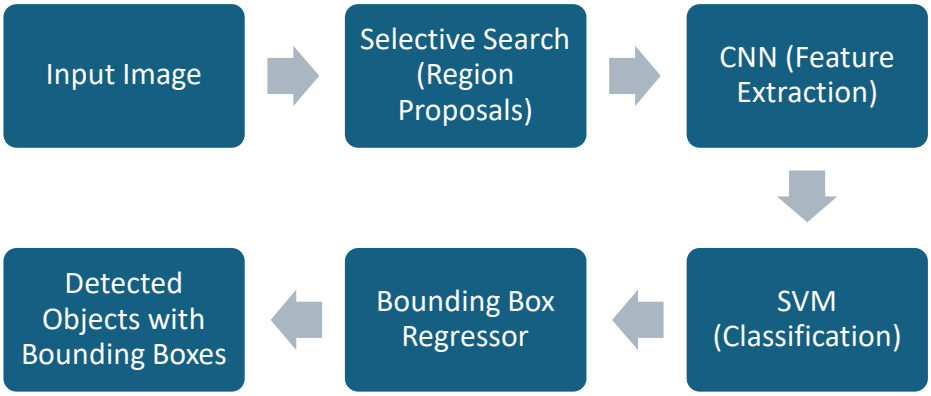

Figure 2: Diagram of Faster R-CNN Architecture

B -Faster R-CNN Builds upon R-CNN by merging region proposal networks, which maintain accuracy and increase speed[14].as shown in figure (3).

- **General Structure of Faster R-CNN**

The Faster R-CNN architecture consists of several key components working together in a unified framework:

- **First phase: Backbone Network:** this network is responsible for extracting the feature maps from an input image. It is basically a pre-trained CNN (like ResNet or VGG)[14].
- **Second phase: Proposal Network for Regions (RPN):** This module generates region proposals using the feature maps produced by the backbone directly. It employs anchor boxes having different aspect ratios and scales to predict locations of potential objects[14].
- **Third phase: The RoI (Region of Interest) Pooling Layer:** This converts the region proposals produced by RPN to a fixed-size feature map that is easy for us to analyze[14].
- **Fourth phase:** It defines the classification and bounding box regression heads which are of supreme importance in any object detection framework. Here, all the proposed regions are passing through fully connected layers for classification to identify what type of objects present in those the same time carry out regression techniques to refine upper left corner

coordinates (red point in figure) of bounding box while decreasing breadth and height. This dual strategy ensures the detection of objects and enables their categorization as well, increasing accuracy in its spatial localization within image input[14].

- **Advantages of Faster R-CNN:**
  - **End-to-End Training**: Faster R-CNN allows end-to-end training, where both region proposal generation and object classification are jointly optimized to make the process faster[14].
  - **The advancement in speed:** There is two different things in this model compared to previous versions like, Fast R-CNN· The improvement of speed. Fast R-CNN used selective search to generate region proposals, but this model improves speed by using a Region Proposal Network (RPN). As a result, this integration alleviates the computational overhead in generating region proposals which yields better processing time and accuracy of the model[12].
  - **Accuracy**: Faster R-CNN uses the deep convolutional features and provides high accuracy of object detection, which is suitable for challenging dataset[14].

- **Disadvantages of Faster R-CNN:**
  - **Computational Complexity:** Even though Faster R-CNN demonstrates a remarkable increase in speed relative to previous object detection models, it still requires high computational resources. This demand is more significant in multiple object and high-resolution image analysis at once[14].
  - **Challenges with Small Objects**: Due to limitations of design and evaluation for anchor boxes, the model may have difficulty identifying small objects or those with particular aspect ratios[14].
  - **Large Training Data Requirements:** Faster R-CNN's utility decreases severely in situations where input data suffers from scarcity, it heavily relies on annotated training data. The information collected is very difficult and time-consuming to gather, which can create obstacles for training a good model[14].

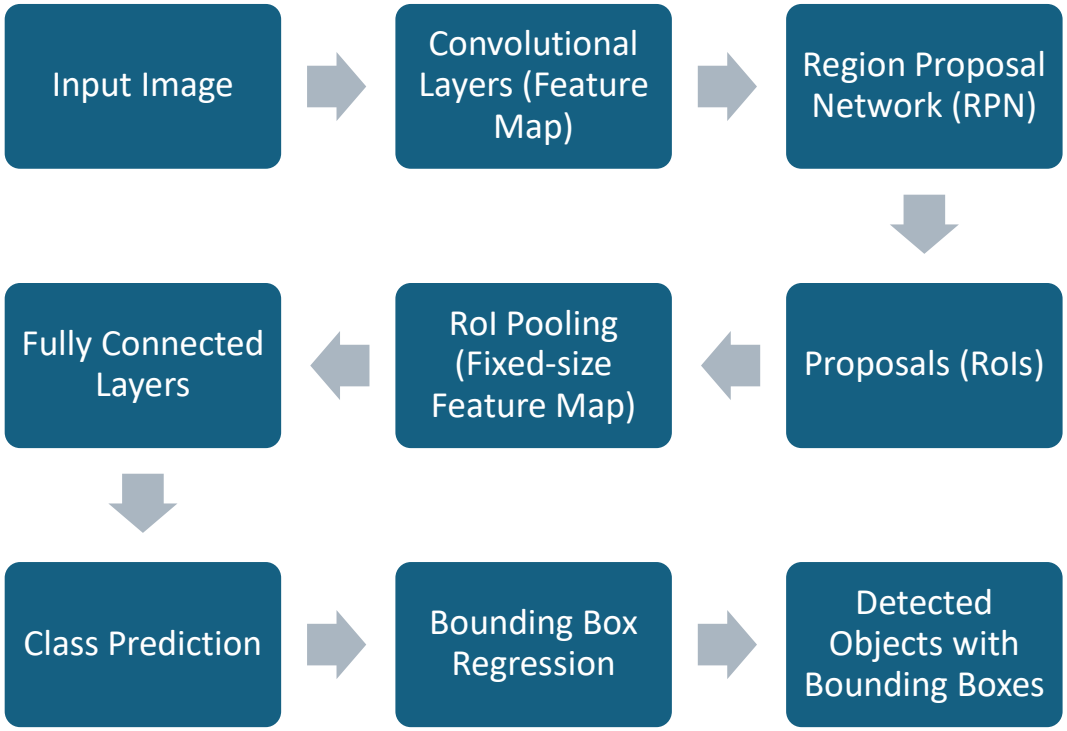

Figure 3: Diagram of Faster R-CNN Architecture

2.2 One-Stage Object Detection Algorithms

These algorithms are intended for real-time inference by predicting bounding boxes and class probabilities jointly across the full image in one go[6]. Object detection: Save time with one-stage object detectors, two-stage Object Detection algorithms are advantageous-dependent on which use case it is. The idea is that they encapsulate both predictions about the tight bounding box and object-based classification in a single pass over the neural network. Next, we summarize the main single-stage object detection algorithms together.

A. **YOLO: (**You Only Look Once) What do you think is special about this technique? Instead of previous models that have different pipelines which consist in generating region proposals firstly and following classification YOLO needs only one pass through its neural network to process the image. This architecture can do end-to-end predictions and allows YOLO to predict both bounding boxes as well as class probabilities in a single pass. That greatly increases the performance of object detection in a large image[15]. as shown in figure (4) & (5). In this review we will focus on the tenth version of YOLO.

- **General Structure of YOLO 10**

مجلة العلوم والتطـبيقات الهـندسية

Journal of Science and Engineering Application

ISSN 2521-3911

1. Backbone: Backbone network in a neural network extracts the features and converts them into an almost similar form. To solve the problems, YOLOv10 uses an improved CSPNet (Cross Stage Partial Network), which has better gradient propagation and reduces computational redundancy. This specialized backbone well preserves important features for accurate object detection[16].

2. Neck: The neck is important as a critical link between the head and spinal column that spans multiple scales. The Path Aggregation Network (PAN) layers in YOLOv10 help to combine multi-scale information, thus allowing the model to gather evidence from a wide variety of resolutions. Abstract Feature-specific transformation (FST) is an essential skill needed for objects detection of different sizes[16].

3. Header: It is the classification part of architecture using YOLOv10 This model uses a dual-head structure which improves the ability to detect and classify objects effectively. Individual heads cooperate with each other, enhancing accuracy and speed of the detection. This method of creating duality through category-agnostic approach allows us to process all object classes together in the same convolutions, and hence improve YOLOv10 simultaneous processing frequencies as most real-time methods:

- The One-to-Many Head approach allows us to produce a few predictions per object while training. This yields a wide spectrum of supervisory signals that increase learning accuracy considerably. Owing to this diversity in predictions, the model can then learn from them and improve its performance as a whole[16].
- The One-to-One Head plans to output a unique optimal prediction for each object during the inference phase. By doing so, this design actually obviates the need for Non-Maximum Suppression (NMS) thus reducing latency[16].
- Lightweight Classification Head: YOLOv10 comes with lightweight classification head, built on depth wise separable convolutions. This design decision is effective in reducing computational costs, while retaining the overall performance[16].
- Spatial-Channel Decoupled Down sampling: This optimization decouples the spatial down-sampling and channel transformation processes that results in computational overhead reduction[16].
- Rank-Guided Block Design: This design replaces redundancies originating in various model stages with compact inverted blocks. This increases efficiency and at the same time reduces the computation costs[16].
- Partial Self-Attention Module YOLOv10 includes an efficient self-attention mechanism that enhances global feature representation by not sacrificing computational overhead[16].
- NMS-Free Design: YOLOv10 achieves a breakthrough by removing Non-Maximum Suppression (NMS) when both training and inference are completed, via using an innovative dual assignment scheme. This offers optimization of the matching predictions directly and not using NMS (Modified Non-Maximum Suppression) hence improve overall efficiency of object detection. The architecture of YOLOv10: a powerful feature extractor, very high-efficient fusion neck and an accurate/low-latency head. In aggregate,

these results provide for an appealing real-time object detection backend in YOLOv10[16].

- **Advantages of YOLO:**
  - **Speed:** YOLO can process images at one time; the first versions achieved up to 45 frames per second, giving results in real-time. This performance suits it for use as a processor which needs to act immediately, such as found in autonomous driving or video surveillance systems[17].
  - **Global Context:** Since YOLO can predict bounding boxes in a single pass through the network, the entire image is used as global context to make this prediction leading to increased localization accuracy[18].
  - **Simplicity:** YOLO, with a more streamlined architecture than models like Faster R-CNN that uses the two-stage cascade of region proposal and detection[17].

- **Disadvantages of YOLO:**
  - **Lower Accuracy on Small Objects** — Owing to its grid-based design, YOLO might lack the precision necessary for accurately detecting small or tightly packed objects which may result in detection failure[18].
  - **Limited Class Predictions per Grid Cell:** In the context of heavily overlapping objects, one issue with a cell-based approach is that each grid cell must be committed to just predicting one class. However, this design limitation can degrade the performance of the model in detecting overlapping objects within a grid cell For handling such cases more intricate designs that assist us predict running predictions over multiple classes[18].
  - **Sensitivity towards Aspect Ratios:** Traditional Object Detection models which work on grid-based systems predict a single class from each cell in the grid. However, it cannot process for more than one object in same spatial region simultaneously because of this limitation with collecting local information from smaller regions restrict to classify multiple objects overlap together accurately[18].

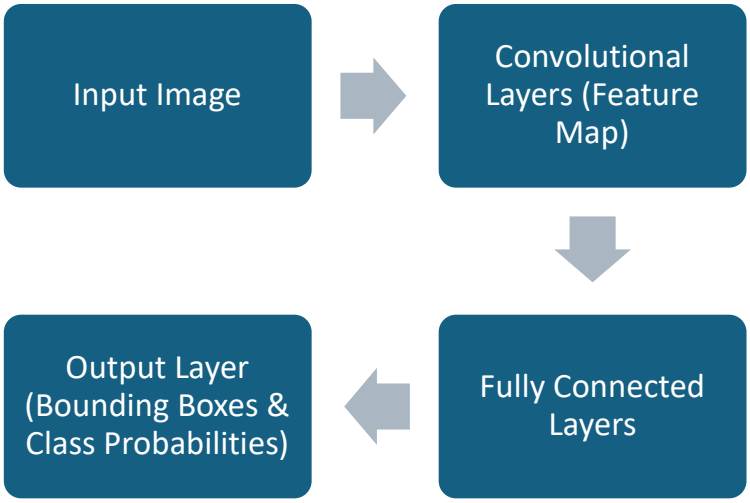

Figure 4: YOLO Architecture Diagram

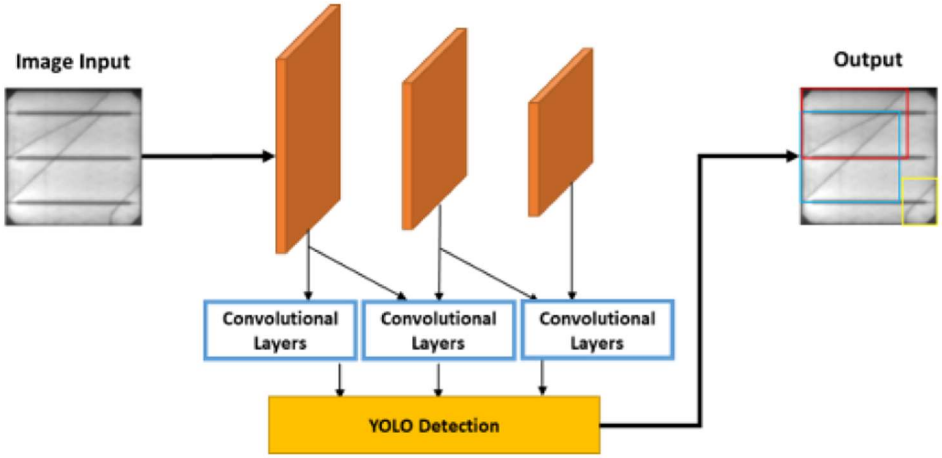

Figure 5: YOLO Architecture Diagram

B. SSD (Single Shot Multi Box Detector): One of these is the Single Shot MultiBox Detector (SSD) that represents a good compromise between speed and accuracy making it real-time applicable. It can identify and segregate multiple sets of objects in an Image concurrently which means it will be more streamlined for better performance especially in a dynamic environment using only one deep learning framework. Since this method was designed for quick turnaround (e.g., real-time feedback applications) it is structured to process quickly in its architecture thus depicted as Fig[13].shown in figure (6).

- **General Structure of SSD**

A Solid-State Drive (SSD) infrastructure is an event-driven architecture due to the acknowledgment of multiple elements during a single, full system pass.

- **Input Layer**: The input image can be resized to a fixed size, most commonly 300x300 pixels[13].
- **Backbone Network**: This is used to take input images and get the feature maps from them. Usually, you will employ backbone network such as VGG16 or Mobile Net, to make full use of their abilities in feature representation which enable the model better understanding and interpreting visual information. These architecture have been design to directly capture critical patterns and features across the image which exploit in better way for performance of computer vision tasks[13].
- **Multi-scale Feature Maps:** at multiscale — The Single Shot MultiBox Detector (SSD) uses feature maps from a few different layers of its backbone network in order to detect objects across scales. The output is the feature maps at different resolutions, which helps detect objects of in many scales since each layer has images with various scales[13].
- **Anchor Boxes with the Single Shot MultiBox Detector (SSD)**: each grid cell in a feature map is accountable for generating bounding boxes; these are what we call anchor boxes. The anchor boxes are specified with aspect ratios and scales which can significantly improve object detection for localized region nature. This makes the model more capable to consider different object shapes and sizes, which increases the overall performance of detection[13].
- **Classification and Localization Layers:** First for all the anchor boxes what class should it be?
    - **Class Scores:** The probabilities corresponding to each class explain the chances of their presence in that context. These probabilities provide quantitative estimates for each takeaway class demonstrating the potential likelihood of an occurrence, supporting assessment and analysis in decisions making[13].
    - **Bounding Box Offsets:** The output data represents the confidence that each class will exist, therefore it indicates how likely a shared box be included. Here, statistical analysis quantifies the likelihood of each category being present in the dataset[13].
- **Non-Maximum Suppression (NMS):** After prediction, NMS is used to get rid of irrelevant bounding boxes by matching the overlaps for each other along with their corresponding confidence scores. And this process will make sure that there is always only relevant predictions placed in the hands of detection systems, therefore improving them even more. Thus, by filtering the boxes with higher confidence at first and suppressing the overlapped areas that crowded into final output, NMS behaves as a post-processing method to help get more accurate sizes of bounding box for detected objects[13].

- **Advantages of SSD**

- **Real-Time Detection:** Solid State Drives (SSDs) enable the delivery of high-speed performance which allows for real-time processing and ensuring adequate accuracy. This feature makes them ideal, for example in video surveillance and autonomous driving applications where fast data access is beneficial[13].
- **Multi-Scale Detection:** Utilizing feature maps from several layers, Single Shot MultiBox Detector (SSD) is superior in detecting objects at different scales. This multi-layer design helps in better detecting smaller objects conceived by YOLO and faster predictions compared to most of the fast single-stage detectors. This feature highlights the multi-task property of SSDs on different detection setting, which specially shows its ability in dealing with complicated environments[13].
- **Unified Architecture:** The framework is also a single-shot approach, as it concatenates region proposal and classification into one network which simplifies the architecture of the system hence reducing computational overhead[13].

- **Disadvantages of SSD**
  - **Accuracy Trade-Off:** High speeds are possible with solid-state drives (SSDs) but they may lack the accuracy of two-stage detectors, like Faster R-CNN when applied to deep clustering object datasets. The variations in detection performance under different scenarios point to the need for selecting an appropriate detection framework, depending on which unique challenges of the data are brought into play[13].
  - **Sensitivity to Aspect Ratios:** The efficiency of anchor boxes can greatly vary depending upon how the aspect ratios are defined w.r.t to target objects. This often results in the degradation of detection capabilities. Consequently recursive optimization of anchor box aspect ratios to better fit distribution over how large category targets are is paramount for improving accuracy at the time detection[13].
  - **Limited Performance on Small Objects:** SSD was a leader in performance over other alternatives, but it still struggles to identify very small objects. This constraint is a function of its reliance on cell-based prediction techniques that may not effortlessly represent minuscule object characteristics [13].

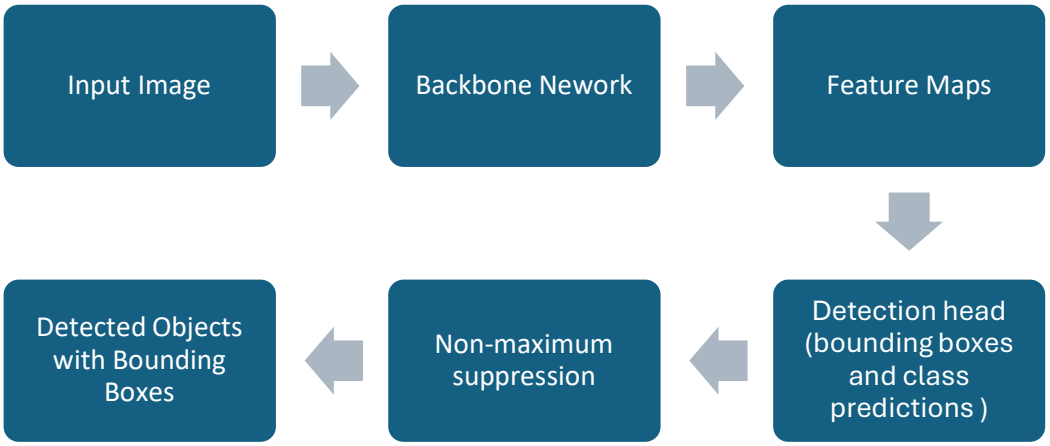

Figure 6: Diagram SSD Architecture

With the remarkable progress made in recent years using deep learning, such algorithms have become more and more viable for a variety of tasks including autonomous vehicles, security systems or robotics. These enhancements enable improved performance and higher throughput, attributes that are imperative for the applications to work in real-world environments[6].

3. Object tracking Algorithms

Recent improvements to deep learning have greatly improved the accuracy and efficiency of these algorithms, making them more broadly usable in practical systems (e.g. autonomous vehicles, security applications like face recognition, or robotics). These advances allow for stronger performance and quicker processing speed, both of which are essential to making you able to park in harmony with these new technologies[19].

- **Advantages of Object Tracking**
  - **Enhanced Precision:** updated algorithms that serve to improve object detection and tracking. This is particularly critical in industries like manufacturing and e-commerce, where it results in fewer errors compared to traditional manual tracking methods. These advanced algorithms

not only help with operational efficiency but also endeavor towards better functioning data management in different sectors as well[20].

- **Operational Efficiency**: Being high-speed systems, automated object tracking solutions aid in improving operational efficiencies specifically in environments such as warehouses and production lines. Increased efficiency leads to increased productivity and as a result of that decreased operation cost also[20].
- **Continuous Monitoring**: With the help of object tracking technology, you can constantly track objects—a critical aspect when it comes to materials introducing into inventory system as well as maintaining and assuring quality. They also have an impressive capability for rapidly adapting to changes in environmental conditions and the properties of tracked objects. It can be Figurative representation of the flexible capability (that stresses adaptivity) to tighten tracking performance aiding both operational continuity and decision-making under diverse applications[20].
- **Cost Savings**: Automated object tracking systems can have significant return on investment when considering long-term cost reduction, by reducing manual labor needed to complete tasks and the potential for human-errors[20].
- **Improved Safety**: It is obvious that safety becomes a major concern as in modern environment where humans and machines coexist, effective object-tracking systems must be implemented to ensure the human's security. These systems allow robots to recognize obstacles and navigate them more safely help reducing the risk of accidents[20].

- **Disadvantages of Object Tracking**
  - **Vulnerability to Changes:** Most object tracking systems have a poor resistance to variations of objects such as modifications of size, shape and appearance due the time including different illumination conditions. And in dynamic environments like retail, these can lead to inaccuracies. Related to the previous sets of apps, in these scenarios tracking with high-level semantics is limited by occlusions, lighting changes or simply due to variation among appearances. As a result, robust tracking performance requires sophisticated algorithms that can handle these variations while maintaining levels of accuracy and confidence[20].
  - **Background Distractions**: Since processing power is abstract and dynamic, the performance can vary significantly depending on environmental conditions due to run-time resource allocation. Understanding how that works exactly will be presented here in Algorithm Performance. Since we train our algorithms in controlled environments, they find it difficult to generalize when the backgrounds get complex as illustrations go beyond clips[20].

- **High Computational Demand**: Advanced tracking algorithms frequently necessitate significant computational resources, which can pose practical challenges for various applications[20].
- **Error Potential:** Error In Case of Challenge via dark scenes – where entities may be partially or completely shadowed behind obstacles, the tracking systems return would fail. Thus, in order to overcome these challenges and be able to handle such disruptions efficiently emptile tracking methodologies must become robust enough so as not lose track of the object through periods where it isn't visible[20].
- **Significant Initial Investment:** As with anything, there are opportunity cost savings in the long run but deploying new object tracking technologies can be expensive and may prevent smaller businesses from accessing these solutions[20].

3.1 Object Tracking in video

Video tracking involves monitoring and following a moving object across frames in avideo. The primary goal is to create a connection between the target's appearance in consecutiveframes. Essentially, video tracking analyzes video sequences to link the object's previousposition withits current one by predicting and drawing a bounding box around it. This techniqueis commonly employed in areas like traffic surveillance, autonomous vehicles, and securitysystems due to its ability to process real-time video footage[21].

3.1.1. Single object tracking

Single object tracking focuses on following a single target throughout a video or sequence of images. The target, along with its bounding box coordinates, is provided in the initial frame, and the algorithm continues to identify and track it in the following frames. These algorithms must be capable of tracking any given object, even if no classification model has been trained specifically for that object[19].

3.1.2. Multiple object tracking

Track, as in multiple objects to track. The tracking algorithms must then identify how many objects are in each of the frames and which object corresponds from frame to frame.

3.2 Object Tracking Method

Stage one: Initializing Targets

The first part of object tracking involves identifying one or more objects that you want to track — for simplicity, I will just define a single target. Typically, this begins by outlining the object with a bounding box. For a set of images such as image sequences this labeling is commonly done on the first image and in video analysis it occurs in firstframe [22].

While recognizing the object it should estimate its position in subsequent frames to keep track of targets. Such tracking requires the use of spatial-temporal analysis to compensate for changes in object appearance and motion dynamics, ensuring continuity throughout an entire sequence of frames [22]. There are two main methods for carrying out the initialization process: Manual and Automatic. Manual initialization requires the user to specify where the object is in terms of boxes or ellipses. This is again opposed to the automatic initialization where no manual input, but rather object detection algorithms are employed for this task[22].

Stage Two: Appearance Modeling

Appearance modeling is a fundamental way in the computer vision field to mimic an objects visual appearance under different conditions. Lighting conditions, angles of observation and movement can cause significant changes to the view in ways that affects perception more than necessary for moving edges between frames – leading into tracking errors when algorithms are trying to keep things consistent. The methods implemented in AppEReal leverage these limitations by incorporating adaptability to the object's motion blur-related distortions. That strengthens the reliability of tracking systems, to cope with that very volatility within visual information[22].Optimization includes improving two key component the first Construction of Reliability Visual Representation: Describing objects including identifying and extracting relevant visual features, development accurate positive examples etc. and the other is Building Robust Statistical models to improve object recognition Providing More Advanced statistical learning approach[22].

Stage three: Motion Estimation

Once the object has been identified and its appearance modelled, we use our estimates of motion to accurately predict where in the next frame that same (or similar) looking object will be located. This involves solving a state estimation problem, typically using methods like linear regression or Kalman filters (or more contemporary approaches such as the particle filter) to produce accurate predictions. All of these approaches offer unique benefits for accuracy and flexibility which enables a strong tracking regardless the different contexts[22].

Stage four: Target Positioning

Motion estimation is a critical technique used to ascertain the likely region where an object is situated. Once this approximate location is established, a visual model can be deployed to refine the search and precisely pinpoint the target's exact position. This methodology typically involves either a greedy search approach or maximum posterior estimation, both of which leverage the insights gained from the initial motion estimation to enhance accuracy and efficiency in localization. By integrating these techniques, researchers can achieve a more reliable determination of object locations in dynamic environments[22].

4. Human Detection in Thermal Videos

Motion estimation is a very important technique used to figure out how far the object may have moved, after which it then follows this likely location. Artificial intelligence and machine learning technology provides a formula to identify an approximate location, which can allow for the subsequent deployment of visual models in order to further narrow down their exact position. This is usually performed using greedy-search algorithms or maximum posterior estimation which exploit the knowledge gained from initial motion estimates to improve accuracy and performance in localization. If these methods are combined, then researchers can receive a better way of things positions at dynamic environments[22].

5. literature Review

Amanda Berg et al. (2016), in a more thorough investigation performed to extend detection and tracking performance inside thermal infrared photos, basically orientated on deciphering demanding situations less than low visibility conditions for strengthening surveillance devices. Fundamentally, their research critically assesses Gaussian Mixture Models for background subtraction, Support Vector Machines (SVM) models for classification and Cellular Models to tracking objects. In addition, the authors applied semantic analysis-based methods for behavior recognition and found that accuracy ranged across a wide spectrum depending on the method used. The vision-based cellular model proved to be highly sensitive, with detection rates reaching from 93.5%–98.2%. At the same time, semantic-based methods were able to recognize fraudulent events at 95% accuracy [21]. This study confirms the efficient operation of these algorithms to help solve specific problems from thermal imagery[23].

Wassim A. El Ahmar et al. (2022), investigated the performance of cold imaging on object detection and tracking, especially under difficult light circumstances. To compare their performance in thermal images with the general RGB image, we applied task-aligned object detection (TOOD) and variable feature network (VFNET), which are advanced algorithms for object detection [24]. on a dataset of RGB-images from Teledyne FLIR Thermal Dataset developed for Advanced Driver-Assistance Systems, where comparing performances helps to draw experience-based conclusion about strengths/weaknesses within TOOD/VFNET. An instance of the latter is a novel dataset appropriately titled City Scene RGB-Thermal MOT Dataset comprised manually labeled thermal images along with associated RGB counterparts[24].The dataset consists of FLIR infrared and accompanying visible-light data collected from the same camera. The results suggest that thermal trackers well-known outperform RGB image-based tracks by a large margin, especially in scenes with low light conditions; The study also suggested a dynamic cut-off threshold in tracking-by-detection pipelines can improve tracker associatively by taking into consideration the bounding box size of detector predictions[24]. Although specific accuracy measures were not reported, the thermal models had what was described as high recall abilities; it shows strong detection performance across different conditions. This work highlights thermal imaging as an enabler of object detection and tracking systems, particularly in adverse conditions[24].

Garima Mathur et al. (2016), introduce study focuses on Intelligent Video Surveillance (IVS) systems that automatically can identify suspicious behaviors such as loitering, unattended objects and unauthorized intrusions are required to improve security in public places. Its innovative take on improving public safety is by using sophisticated algorithms to monitor and analyze surveillance videos in real-time. By applying ML (machine learning) mechanisms, IVS systems promise to deliver more efficient and fast security encapsulations within different surroundings. Gaussian Mixture Model for Subtraction of Background and Some Algorithms. Classification: SVM, ANN A Self-Adaptive Background Matching Mechanism with Cauchy Distribution. Behavior classification: Semantic Analysis A variety of datasets from the literature were reviewed by these researchers; however, we do not know which ones will be tested with proposed methods. The changes in performance across different algorithms reflect differences accuracy between approaches — detection rates ranging from 93.5% to 98.2%, with the vision-based cellular model performing best. Detecting suspicious activities up to 95% was reported by semantic based approaches. Nevertheless, no method had perfect performance suggesting future areas of research and continued efforts to refine the experiments[25].

Amanda Berg et al. (2015), this work presented the objectives of a study on establishing a standardized benchmark for short-term single object tracking in thermal infrared imagery to address lack of dedicated effort towards it. The research discussed various tracking methods, including: Kalman Filters, Particle Filters, Mean Shift and Camshafts algorithms These algorithms are evaluated within the context of the proposed benchmark. The benchmark introduces the LTIR dataset, which consists of 20 thermal image sequences collected from multiple sources. The dataset is annotated following the Visual Object Tracking (VOT) protocol, allowing for effective evaluation of tracking methods. While specific accuracy metrics for the algorithms tested were not detailed in the summary, the paper emphasizes that

the performance of tracking methods differs significantly between visual and thermal imagery, highlighting the importance of tailored benchmarks for thermal tracking[26].

Ángel Torregrosa-Domínguez et al. (2024), introduce study focuses on enhancing real-time weapons detection systems to improve security measures in industrial environments, particularly addressing the challenge of detecting small weapons. The research utilizes: You Only Look Once (YOLO) versions, specifically YOLOv5, YOLOv7, and YOLOv8, for object detection. A Scale Match method to improve detection accuracy for small-aspect ratio weapons. The authors created a new dataset called Disarm-Dataset by combining existing datasets with their own. This dataset includes complex images with difficult-to-identify weapons and simpler images with clearly detectable weapons. The experimental results indicate: An improvement of +13.23 in average precision when using the Scale Match method. A 71% reduction in false positives compared to the baseline model[10].

Tomás Santos et al. (2024), conducted a systematic review on deep learning-based weapon detection in surveillance footage, focusing on the methods employed, dataset characteristics, and challenges in automatic weapon detection. The review highlights the use of several models Faster R-CNN, YOLO (You Only Look Once) architecture, The study discussed datasets that incorporate: Realistic images and synthetic data, which have shown to improve detection performance. The review notes that while various models demonstrated improvements in accuracy, specific numerical metrics were not consistently provided. However, it emphasizes that the performance of these models significantly decreases under challenging conditions, such as poor lighting and small weapon detection. The main challenges identified include Poor lighting conditions affecting detection accuracy[27].

Saksham Gosain et al. (2021), introduced The paper on Detecting Hidden Weapons through Image Processing and Machine Learning presents an approach that aims to replace traditional X-ray detection methods with automated and less error-prone techniques. The study explored several algorithms, including: YOLOv6 for object detection. Convolutional Neural Networks (CNN) for weapon detection, specifically using VGG Net. Logistic Regression for classifying images into weapon and non-weapon categories. Fuzzy KNN (K-Nearest Neighbors) as a potential technique for weapon detection in X-ray images. The dataset included images that facilitate the fusion of thermal/infrared images with traditional RGB or HSV images. experimental results referred: Test accuracy: 0.9375 and training accuracy: 0.85[28].

Pavinder Yadav et al. (2023), present a study on advancements in weapon detection technology, underscoring the necessity for automated systems that can identify criminal activities in CCTV footage without human intervention. The research explores various algorithms, focusing primarily on classical machine learning techniques and deep learning methods, though specific deep learning algorithms are not detailed in the findings. The authors point out the absence of a dedicated real-time dataset for weapon detection, noting that most existing datasets are derived from virtual environments like films and video games. The paper concludes that while deep learning techniques surpass traditional methods in both speed and accuracy, there remains a significant gap in creating a unified and robust deep learning approach for real-time weapon detection[9].

Muhammad Tahir Bhatti et al. (2021), introduced This work focused on providing a secure environment by utilizing CCTV footage as a source for detecting harmful weapons, applying state-of-the-art open-source deep learning algorithms. them implemented binary classification, assuming the pistol class as the reference class, and introduced the concept of including relevant confusion objects to reduce both false positives and false negatives. Since no standard dataset was available for real-time scenarios, them created our own dataset by capturing weapon images with our cameras, manually collecting images from the internet, extracting data from YouTube CCTV videos, utilizing GitHub repositories, and gathering data from the University of Granada and the Internet Movie Firearms Database (IMFDB) at imfdb.org. Two approaches were employed: sliding window/classification and region proposal/object detection. The algorithms used include VGG16, Inception-V3, Inception-ResnetV2, SSDMobileNetV1, Faster-RCNN Inception-ResnetV2 (FRIRv2), YOLOv3, and YOLOv4. Precision and recall are prioritized over accuracy when performing object detection, so all algorithms were tested based on these metrics. YOLOv4 stood out as the best among all algorithms, achieving an F1-score of 91% and a mean average precision of 91.73%, surpassing previously achieved results[29].

The reviewed studies demonstrate a variety of approaches for detecting suspicious individuals carrying weapons using machine learning techniques in thermal video. Deep learning models, such as YOLO, have been notably effective in real-time applications across multiple studies. As the field progresses, future research will likely aim at improving accuracy, minimizing false positives, and incorporating multiple sensor modalities to further enhance detection capabilities.

It is possible to summarize the previous works and focus on the most important data in each paper as shown in the table (1).

Table 1: survey the previous works

NO	Title of Research	Year	Algorithms Used	Publisher
1	Detection and Tracking in Thermal Infrared Imagery	2016	Various detection and tracking methods; includes benchmarks for thermal tracking	DiVA Portal
2	Multiple Object Detection and Tracking in the Thermal Spectrum	2022	YOLO, Byte Track, TOOD, VFNET	CVPR 2022 Workshop
3	Research on Intelligent Video Surveillance Techniques for Suspicious Activity Detection	2016	Various intelligent surveillance techniques	International Conference on Recent Advances and Innovations in Engineering (ICRAIE)
4	A Thermal Object Tracking Benchmark	2015	Proposed benchmarks for short-term single object tracking in thermal imagery	IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)
5	Effective Strategies for Enhancing Real-Time Weapons Detection in Industry	2024	YOLOv5, YOLOv7, YOLOv8	MDPI - Applied Sci
6	Systematic Review on Weapon Detection in Surveillance Footage through Deep Learning	2024	Various deep learning methods	Elsevier
7	Concealed Weapon Detection Using Image Processing and Machine Learning	2021	Image Processing, Machine Learning	International Journal of Research and Scientific Technology

NO	Title of Research	Year	Algorithms Used	Publisher
8	A Comprehensive Study Towards High-Level Approaches for Weapon Detection	2023	Classical ML, Deep Learning	Elsevier
9	Real-Time Gun Detection in CCTV Using Deep Learning	2021	YOLOv3, CNN	IEEE Access

6. Challenges

- **Data Scarcity**: There is a limited availability of annotated thermal datasets specifically designed for weapon detection Many existing datasets predominantly concentrate on general object detection or the identification of human presence, often neglecting to provide specific labels for weapon detection. This limitation hinders the development of advanced models capable of accurately recognizing and classifying firearms and other weapons in various contexts[30]. The absence of specific information on the annotation of weapons not only hinders existing algorithms from doing their best; it also limits the jobs that security, surveillance and law enforcement applications would like to accomplish. Tackling these deficiencies requires creating contextual weapon specific training data, to improve model performance on the detection of such instances and untangle those features that make manifest a complex narrative with weapons.
- **Environmental Variability:** Thermal imaging technology is heavily influenced by ambient environmental conditions such as temperature changes and the presence of clutter signals. So, at times of standing water and cooler temperatures these can obscure or misrepresent a heat signature affecting clarity & precision in what the thermal image shows. For example, transformations in surrounding temperature led to thermal gradients that can hamper the discovery of certain heating sources. In the same vie, background noise (e.g. reflections of neighboring objects) can reject the thermal signatures essential for accurate imaging and evaluation[30].
- **Integration with RGB Data:** While coupling thermal and RGB data can help improve the detecting power, it needs to involve sophisticated fusion sensor methodologies that could prescribe objects identification wihttps://www.quora.com/What-is-threshold-detectionth different modalities[30].
- **Future Directions:** Subsequent research should focus on building extensive datasets of a range of scenarios related to weapon-based activities. Such an approach is necessary to train models that are robust against a variety of operational issues. Also, work is needed in improving the algo-rhythms that can deal with occlusions and getting through cluttered environments in which bad behavior could happen. There is also an urgent need to investigate the latest developments in real-time processing capabilities that would provide near instantaneous responses within security use

cases. To sum up, the use of machine learning in conjunction with thermal imaging may be an effective approach to detect and follow individuals bearing weapons. Ongoing research and innovation in this area have the potential to provide substantial public health benefits across multiple settings[30].

7. Conclusion

The jobs of object detection and tracking have seen some significant progress due to the inventions created by deep learning. Which means choosing one of the best algorithms is relying on several very important facets like computational pace, accuracy and assets availability for processing. Research has therefore suggested efforts to improve the robustness of these algorithms about many adversarial challenges in recent years. Specifically, the work needs to focus on improving occlusion performance (i.e. when objects are partially covered), building systems that can adapt under a better variety of lighting conditions which affects detection accuracy directly and optimizing algorithms for real-time processing because some applications require it like Security, autonomous vehicles or interactive system etc.… The next steps in research will probably be to focus on improving the generalization of algorithms across diverse environments, which may include new architectures and training methodologies. This could mean using synthetic data to boost training datasets, making algorithms more robust by resorting to adversarial testing or combining different sensor types in multimodal approaches. Above all else, however, the focus is on developing robust and scalable systems that are up to snuff in both dynamic or adverse environments[27].

Reference

[1] M. C. Monuteaux, L. K. Lee, D. Hemenway, R. Mannix, and E. W. J. A. j. o. p. m. Fleegler, "Firearm ownership and violent crime in the US: an ecologic study," vol. 49, no. 2, pp. 207-214, 2015.

[2] R. J. Radke, S. Andra, O. Al-Kofahi, and B. J. I. t. o. i. p. Roysam, "Image change detection algorithms: a systematic survey," vol. 14, no. 3, pp. 294-307, 2005.

[3] M. Cristani, M. Farenzena, D. Bloisi, and V. J. E. J. o. A. i. s. P. Murino, "Background subtraction for automated multisensor surveillance: a comprehensive review," vol. 2010, pp. 1-24, 2010.

[4] T. Bouwmans, F. El Baf, and B. Vachon, "Statistical background modeling for foreground detection: A survey," in *Handbook of pattern recognition and computer vision*: World Scientific, 2010, pp. 181-199.

[5] T. J. R. P. o. C. S. Bouwmans, "Recent advanced statistical background modeling for foreground detection-a systematic survey," vol. 4, no. 3, pp. 147-176, 2011.

[6] K. Du and A. J. E. P. Bobkov, "An overview of object detection and tracking algorithms," vol. 33, no. 1, p. 22, 2023.

[7] A. H. Ashraf *et al.*, "Weapons detection for security and video surveillance using cnn and YOLO-v5s," vol. 70, no. 4, pp. 2761-2775, 2022.

[8] K. Vijayakumar, K. Pradeep, A. Balasundaram, A. J. M. B. Dhande, and Engineering, "R-CNN and YOLOV4 based Deep Learning Model for intelligent detection of weaponries in real time video," vol. 20, no. 12, pp. 21611-21625, 2023.

[9] P. Yadav, N. Gupta, and P. K. J. E. S. w. A. Sharma, "A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods," vol. 212, p. 118698, 2023.

[10] Á. Torregrosa-Domínguez, J. A. Álvarez-García, J. L. Salazar-González, and L. M. J. A. S. Soria-Morillo, "Effective Strategies for Enhancing Real-Time Weapons Detection in Industry," vol. 14, no. 18, p. 8198, 2024.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[12] L. Du, R. Zhang, and X. Wang, "Overview of two-stage object detection algorithms," in *Journal of Physics: Conference Series*, 2020, vol. 1544, no. 1, p. 012033: IOP Publishing.

[13] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016, pp. 21-37: Springer.

[14] S. Ren, K. He, R. Girshick, J. J. I. t. o. p. a. Sun, and m. intelligence, "Faster R-CNN: Towards real-time object detection with region proposal networks," vol. 39, no. 6, pp. 1137-1149, 2016.

[15] T. Diwan, G. Anirudh, J. V. J. m. T. Tembhurne, and Applications, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," vol. 82, no. 6, pp. 9243-9275, 2023.

[16] A. Wang *et al.*, "Yolov10: Real-time end-to-end object detection," 2024.

[17] M. G. Ragab *et al.*, "A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023)," 2024.

[18] R. J. L. h. w. v. l. c. b. y.-o.-d. L. Kundu, "YOLO: Algorithm for Object Detection Explained," vol. 19, p. 2023, 2023.

[19] A. Yilmaz, O. Javed, and M. J. A. c. s. Shah, "Object tracking: A survey," vol. 38, no. 4, pp. 13-es, 2006.

[20] J. Xie, E. Stensrud, and T. J. S. Skramstad, "Detection-based object tracking applied to remote ship inspection," vol. 21, no. 3, p. 761, 2021.

[21] Y. Chen *et al.*, "Satellite video single object tracking: A systematic review and an oriented object tracking benchmark," vol. 210, pp. 212-240, 2024.

[22] Z. Soleimanitaleb, M. A. Keyvanrad, and A. Jafari, "Object tracking methods: a review," in *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2019, pp. 282-288: IEEE.

[23] A. Berg, "Detection and tracking in thermal infrared imagery," Linköping University Electronic Press, 2016.

[24] W. A. El Ahmar, D. Kolhatkar, F. E. Nowruzi, H. AlGhamdi, J. Hou, and R. Laganiere, "Multiple object detection and tracking in the thermal spectrum," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 277-285.

[25] G. Mathur and M. Bundele, "Research on intelligent video surveillance techniques for suspicious activity detection critical review," in *2016 international conference on recent advances and innovations in engineering (ICRAIE)*, 2016, pp. 1-8: IEEE.

[26] A. Berg, J. Ahlberg, and M. Felsberg, "A thermal object tracking benchmark," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015, pp. 1-6: IEEE.

[27] T. Santos, H. Oliveira, and A. J. C. S. R. Cunha, "Systematic review on weapon detection in surveillance footage through deep learning," vol. 51, p. 100612, 2024.

[28] S. Gosain, A. Sonare, S. J. I. J. f. R. i. A. S. Wakodkar, and E. Technology, "Concealed weapon detection using image processing and machine learning," vol. 9, no. 12, pp. 1374-1384, 2021.

[29] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. J. I. A. Fiaz, "Weapon detection in real-time cctv videos using deep learning," vol. 9, pp. 34366-34382, 2021.

[30] P.-F. Tsai, C.-H. Liao, and S.-M. J. S. Yuan, "Using deep learning with thermal imaging for human detection in heavy smoke scenarios," vol. 22, no. 14, p. 5351, 2022.