PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
الجمهـوريـــة الجزائـريـــة الديـمقراطيـــة الشعبيـــة
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
وزارة التعليـم العالـي والبحـث العلمـي
HIGHER SCHOOL OF COMPUTER SCIENCE AND DIGITAL TECHNOLOGIES-BEJAIA
المـدرسة العـليا في علـوم وتكنـولوجيـات الاعـلام الالـي بجـاية

Dissertation Submitted to the Department Of Computer Science in Partial Fulfillment of the Requirements for Master's Degree in Computer Science

Specialty: Artificial Intelligence and Data Sciences

Submitted By:

Ms. MELIZOU Ouassila & Ms. TOUATI Hayet

Review of Techniques for Optimizing Intelligent Video Recording Using Activity Detection in Surveillance Systems

Supervised by:

Dr. ELMIR Youssef
Estin

Members of jury:

▪ **Dr.** Leila CHELOUAH	President	Estin
▪ **Dr.** Lamia CHEKLAT	Examiner	Estin
▪ **Mr.** Anis Chawki ABBES	Examiner	Estin
▪ **Mrs.** Macilia Boukhama	Examiner	Estin

Academic year: 2023/2024

Acknowledgments

First and foremost, All Praises are for Allah almighty ,The most Gracious and most merciful, who gave us the strength, courage, health, and patience to successfully conduct our research.

In this regard, we would like to express our deepest gratitude to our esteemed supervisor, Dr. Elmir Youssef, for the time he dedicated and the valuable information he provided with interest and understanding.

We also extend our heartfelt thanks to our internship supervisor, Mr. Ridha Chekroune, for the time he invested, his assistance in the success and smooth running of our internship under the best conditions, and the precise information he shared with us. We are also deeply grateful to the examiners of our thesis and for agreeing to evaluate our work.

Moreover, we acknowledge all the teachers and administrative staff of the Graduate School of Computer Science and Digital Technologies (ESTIN of Bejaia) which contributed to our education throughout our university course.

Finally, we extend our profound and sincere thanks to our families for their encouragement and moral and financial support. We do not let this occasion pass without thanking everyone who contributed, directly or indirectly, to the completion of this thesis.

Dedication

I dedicate this modest work to my dear family: my father, my mother, and my siblings. To my father, who guided me and fought for me to study and remain strong during the challenging periods of school, I want to thank him infinitely. To my dear, patient mother, thank you for your support and love. I also want to express my gratitude to my extended family—my grandparents, uncles, aunts, and cousins.

Finally, to my friends, whose love and encouragement meant a lot to me, thank you for having my back when I needed you and for listening to my exam complaints, especially my two best friends, Hanane and Kenza.

Touati Hayet.

With my sincere thanks and deep gratitude, I dedicate this humble work to my parents, who have sacrificed and given everything for my happiness and success. To my brother Mami, who has always been there for me. His unwavering belief in me and his constant encouragement have been invaluable. To my best friends Zahra, Dyhia khawla and Imene. To my cousins Nesrine, and Hanan and to all my friends.

I am profoundly grateful for the unwavering support and love from my entire family, my uncles, and my cousins, whose encouragement and belief in me have been a constant source of strength. Their presence in my life has made all the difference, and I am truly blessed to have such a wonderful family.

Melizou Ouassila.

List of Figures

Contents

General Introduction

The widespread installation of surveillance cameras, now a common sight in our urban environment, has resulted in a massive surge of visual data. According to a study of IHS Markit , estimated that as of 2021, there are over 1 billion surveillance cameras installed worldwide, generating astronomical amounts of video data daily. This profusion of data poses major challenges in terms of storage, with hundreds of petabytes required to store this crucial information. In the face of this reality, the search for efficient methods to manage this mass of data and facilitate search becomes an imperative necessity.

Surveillance cameras, although crucial for security and surveillance, have created an information-saturated environment. Indeed, according to estimates, a single HD camera can generate hundreds megabytes of data per minute, which, multiplied by thousands of cameras in an urban area, creates a mountain of data that is difficult to manage. These data, which are often redundant or irrelevant, require careful consideration of how to store them optimally while ensuring the rapid retrieval of critical information.

The problem that emerges from this critical situation is centered on how to optimize the storage of surveillance camera data. The use of Intelligent algorithms is emerging as a promising path to filter sequences, recording only those that are of particular interest. This approach thus poses itself as a potential response to the major challenge of information overload in the field of video surveillance.

This thesis systematically explores the existing methods of activity detection through video surveillance systems. The is structured as follows:

- Chapter 1: a comprehensive review of foundational concepts in machine learning, deep learning, computer vision, and pattern recognition.
- Chapter 2: an in-depth look at the components and architectures of video surveillance systems, including analog, digital, and network-based systems.
- Chapter 3: focuses on the state of the art, offering a detailed analysis of current methodologies, identifying their strengths and weaknesses, and highlighting gaps and limitations. Based on these insights, we propose solutions to address common limitations identified in existing research, such

as enhancing real-time processing capabilities, improving generalization to diverse conditions, and optimizing computational requirements.

This research aims to provide practical recommendations for advancing data management in the demanding field of video surveillance.

Chapter 1

Foundations and Definitions

Introduction

The first chapter of this thesis is dedicated to laying the foundations for understanding the key concepts and definitions that are essential for delving deeper into the intricacies of Intelligent Video Recording Optimization using activity d etection for surveillance systems. This chapter aims to provide a comprehensive overview of the foundational elements that underpin the methodologies and technologies discussed in subsequent chapters.

1.1 Machine learning

1.1.1 Definition

Machine learning (ML) is a subset of Artificial Intelligence (AI) and computer science that focuses on enabling AI systems to learn from data and improve over time without being explicitly programmed. It involves the use of algorithms that analyze data, identify patterns, and make predictions or decisions based on that data. In essence [IBM24b], ML works through a three-step process :

1. Decision Process: ML algorithms analyze input data, which can be labeled or unlabeled, to make predictions or classifications about patterns in the data.

2. Error Function: an error function evaluates the accuracy of the model's predictions by comparing them to known examples or labels in the data.

3. Model Optimization Process: if the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate.

The algorithm will repeat this iterative "evaluate and optimize" process, updating weights autonomously until a threshold of accuracy has been met [IBM24b].

1.1.2 Machine learning methods

1. Supervised machine learning : relies on labeled datasets to train algorithms for accurate classification or prediction tasks.As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This process, often accompanied by cross-validation, aims to prevent over-fitting or under-fitting of the model. Supervised learning plays a crucial role in addressing various real-world challenges, such as segregating spam emails from legitimate ones. Common techniques employed in supervised learning include neural networks, Naïve Bayes, linear regression, logistic regression, random forest, and support vector machine (SVM) [IBM24b].

2. Unsupervised learning: employs machine learning algorithms to scrutinize and categorize unlabeled datasets into subsets known as clusters. These algorithms autonomously uncover latent patterns or data groupings without human guidance. This methodology's adeptness in identifying similarities and disparities in data renders it invaluable for exploratory data analysis, devising cross-selling strategies, segmenting customers, and performing tasks like image and pattern recognition. Additionally, unsupervised learning facilitates the reduction of model features through dimensionality reduction techniques, with principal component analysis (PCA) and singular value decomposition (SVD) being prominent methods. Neural networks, k-means clustering, and probabilistic clustering approaches are among the various algorithms utilized in unsupervised [IBM24b].

3. Semi-supervised learning : combines elements of supervised and unsupervised learning by using both labeled and unlabeled data for training. The algorithm leverages the labeled data to guide the learning process while also exploiting the additional information present in the unlabeled data [IBM24b].

4. Reinforcement learning : involves training an agent to interact with an environment and learn optimal actions through trial and error. The agent receives feedback in the form of rewards or penalties based on its actions, and the goal is to maximize cumulative rewards over time [IBM24b].

1.2 Deep learning

1.2.1 Definition

The simple machine learning algorithms work well on a wide variety of important problems. However,they have not succeeded in solving the central problems in AI such as recognizing speech or recognizing objects. The development of deep learning was motivated in part by the failure of traditional algorithms to generalize well on such Al tasks. the challenge of generalizing to new examples becomes exponentially more difficult when working with high-dimensional data, and how the mechanisms used to achieve generalization in traditional machine learning are insufficient to learn complicated functions in high-dimensional spaces. Such spaces also often impose high computational costs. Deep learning was designed to overcome these and other obstacles [GBC16].

The above describes the simplest type of deep neural network in the simplest terms. However, deep learning algorithms are incredibly complex, and there are different types of neural networks to address specific problems or datasets, for example :

- Convolutional neural networks (CNNs) are a specific type of neural network, which is composed of node layers, containing an input layer, one or more hidden layers and an output layer. Each node connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network [HS24].

- Recurrent neural networks (RNNs) use their "memory" as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of RNNs depends on the prior elements within the sequence. While future events would also be helpful in determining the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions [HS24].

the figure 1.1 of the Venn diagram visually illustrates how these different concepts and classes overlap or are distinct from each other within the broader landscape of machine learning, providing a useful tool for understanding the relationships and distinctions between various approaches and techniques in the field.

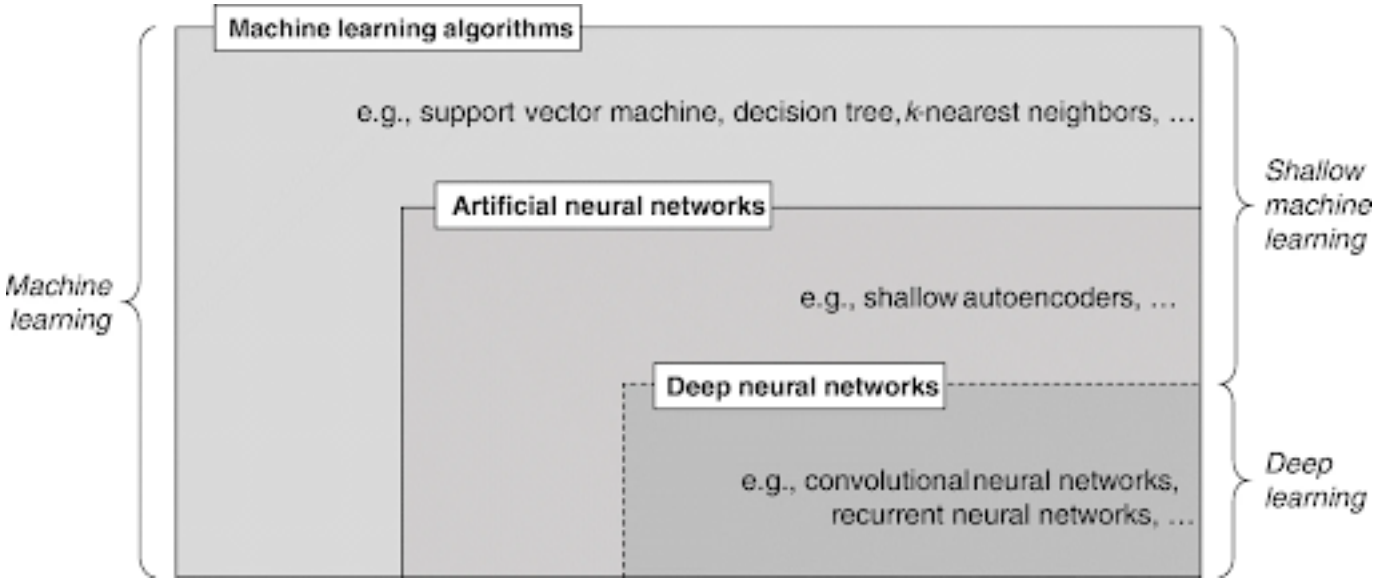

Figure 1.1: Venn diagram of machine learning concepts and classes [GBC16].

1.3 Computer vision

1.3.1 Definition

Computer vision, as a field of study, is primarily concerned with enabling computers to perceive and understand visual data . The overarching objective is to decipher the contents of digital images, often achieved by developing techniques that mimic human visual capabilities. This entails extracting meaningful descriptions from images, which could range from identifying objects to generating textual descriptions or constructing three-dimensional models. Essentially, computer vision involves automating the process of extracting information from images, which can encompass tasks like object detection, object recognition, determining camera positions, and organizing image content. It's important to differentiate computer vision from image processing. While image processing involves modifying or enhancing existing images, focusing on aspects like brightness or color adjustments, it doesn't concern itself with interpreting the content of the image. However, computer vision systems may incorporate image processing techniques as a preliminary step, such as pre-processing raw images before analysis [Bro19].

Figure 1.2 illustrates the hierarchical relationship among Artificial Intelligence (AI), Machine Learning (ML), and specific application areas like Computer Vision. AI encompasses the entire field, providing the foundational goals and objectives. Machine learning is a critical component within AI, offering the methodologies and algorithms that enable systems to learn from data. Building upon machine learning, computer vision applies these techniques to visual data, allowing computers to gain insights from images and

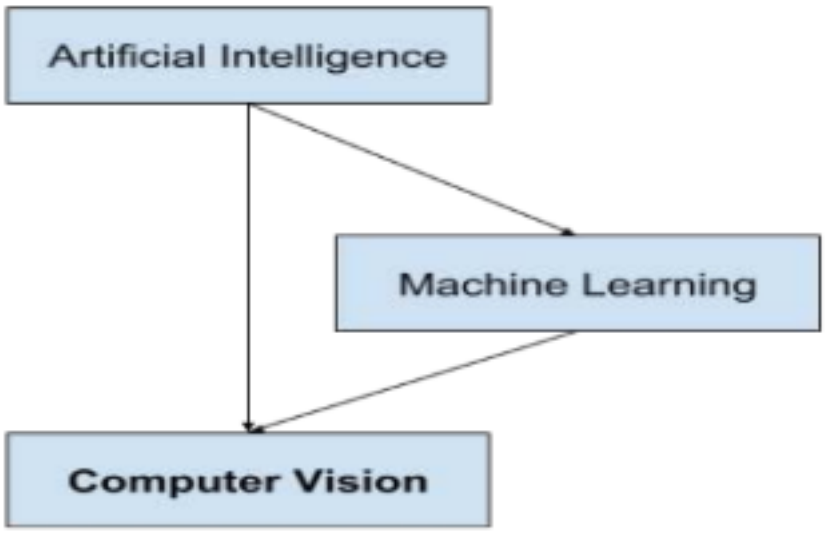

Figure 1.2: Overview of the Relationship of Artificial Intelligence and Computer Vision [Bro19].

videos.

1.3.2 Tasks and applications

As resources for developing computer vision applications become more accessible, it's crucial to address a fundamental question: What specific functions will these applications perform? Clarifying and defining precise computer vision tasks can focus and validate projects and applications, streamlining the initiation process. Below are several examples of established computer vision tasks [IBM24a]:

- Image classification involves analyzing an image and assigning it to a particular category (such as identifying a dog, an apple, or a person's face). Specifically, it accurately predicts that a given image belongs to a specific class. For instance, a social media platform might utilize image classification to automatically detect and segregate objectionable images uploaded by users [IBM24a].

- Object detection utilizes image classification to identify a specific class of objects and then locates and records their presence within an image or video. Applications include detecting defects on assembly lines or identifying machinery in need of maintenance [IBM24a].

- Object tracking involves monitoring or tracing an object once it has been detected. This task is often performed using sequential images or real-time video feeds. For example, autonomous vehicles must not only classify and detect objects like pedestrians, vehicles, and road infrastructure but also track their movements to avoid collisions and comply with traffic regulations [IBM24a].

- Content-based image retrieval leverages computer vision to browse, search, and retrieve images from extensive databases based on their content rather than metadata tags. This task can incorporate automatic image annotation to replace manual tagging, benefiting digital asset management systems and enhancing search and retrieval accuracy [IBM24a].

1.4 Pattern recognition and image analysis

Pattern recognition and image analysis are closely related fields within the broader domain of computer vision and artificial intelligence. While pattern recognition focuses on identifying regularities or patterns in data, image analysis specifically deals with processing and interpreting visual information contained in images [Arm] [IBM24a].

Pattern recognition techniques are applied to analyze various types of data, including images, signals, text, and more. In the context of image analysis, pattern recognition algorithms are used to detect and interpret patterns within images, such as shapes, textures, objects, or structures. These algorithms enable computers to understand and extract meaningful information from visual data [Arm]. Advanced topics in pattern recognition include:

- Statistical, structural, and syntactic pattern recognition : these are different approaches to classify data based on statistical analysis, structural relationships, or syntactic rules [Arm].

- Feature extraction and reduction: this involves identifying the most relevant data attributes (features) and reducing the dimension of the dataset to improve the performance of pattern recognition algorithms [Arm].

Image analysis, on the other hand, is a subset of computer vision that deals with processing and interpreting visual information contained in images. It encompasses various tasks such as image enhancement, feature extraction, image segmentation, and object detection, aiming to understand and analyze the visual content of images [Spr] [MDP]. Some of the current trends in image analysis are:

- Color and texture analysis: These techniques analyze the color and texture patterns within an image to identify objects or regions of interest [Arm].

- Image segmentation and compression: Segmentation divides an image into parts for easier analysis, while compression reduces the image size for storage and transmission without significantly affecting the quality [Arm].

Both fields are rapidly evolving with the advent of deep learning and neural networks, which have significantly improved the capabilities of both pattern recognition and image analysis systems. For instance, convolutional neural networks (CNNs) have become a staple in image classification tasks due to their ability to learn hierarchical representations of visual data.

Conclusion

After we have explored the foundational concepts of machine learning, deep learning, computer vision, and pattern recognition. we are able to receive the subsequent discussions on intelligent video recording optimization and activity detection for surveillance systems. In the next chapter, we will delve into the intricacies of video surveillance systems. We will define what constitutes a video surveillance system, examine its key components, and explore various system architectures, including analog, digital, and network surveillance systems. This comprehensive understanding of video surveillance systems is crucial for appreciating the advancements and addressing the challenges in the realm of smart surveillance technologies.

Chapter 2

Video Surveillance System

Introduction

Video surveillance has experienced significant expansion both technologically and economically in recent years. It has become one of the essential components of government security policies, serving as a cornerstone in ensuring public safety amidst the challenges posed by escalating criminal activities. This evolution addresses every citizen's fundamental need for security, offering reassurance and protection in the face of increasing delinquency and crime rates. In this chapter, we will delve into the intricacies of video surveillance systems, examining their diverse components, structural frameworks, various types, and expansive domains of application.

2.1 Definition of video surveillance system

Video surveillance constitutes a sophisticated system incorporating strategically positioned cameras within a defined area to ensure continuous monitoring. These cameras are linked to a computer system capable of processing and analyzing incoming data. Originating in Germany in 1942 with Siemens AG's development for rocket observations, video surveillance systems have undergone significant evolution. Modern advancements have led to automated data analysis and integration, thereby reducing the necessity for extensive human involvement [Mok12].

Utilizing a network of cameras, monitors, and video interfaces, video surveillance involves observing scenes and scrutinizing behaviors indicative of impropriety or potential threats. Integration with autonomous artificial intelligence enhances the system's capabilities, elevating its efficiency and effectiveness.

Deployed in diverse settings, both indoor and outdoor, spanning structures and properties, these systems

operate continuously. They can be programmed to record upon motion detection or during specified periods, offering comprehensive monitoring and surveillance functionalities.

2.2 Video surveillance system Components

2.2.1 Surveillance cameras

Surveillance cameras are the eyes of the system, capturing video in real-time. They come in various types, including analog and digital (IP cameras), each suitable for different applications. Features like night vision, motion detection, and thermal imaging are common, enhancing the camera's ability to record in diverse conditions. Cameras can also vary in housing configuration to suit different installation environments.

Here are some of the IP camera models available on the market, each designed to meet specific surveillance needs [Kol18]:

1. Fixed IP Camera: unlike motorized cameras, a fixed IP camera monitors a specific location. Its main advantage is its effectiveness in deterring ill-intentioned individuals who quickly realize they are being filmed when they see the camera's lens directed toward them. This makes it an excellent choice for protecting residences or businesses.

2. Motorized Dome IP Camera: this type of camera can be installed anywhere, such as on walls or ceilings. A key feature is that it can be remotely controlled via a controller. It can zoom in on objects or people and perform 360-degree rotations, which is useful for sweeping the surroundings and obtaining a comprehensive view. Additionally, it allows for the scheduling of surveillance rounds at convenient times.

3. Box IP Camera: shaped like a box, this camera includes various types of lenses, making it versatile for different video surveillance needs. PTZ (Pan-Tilt-Zoom) and Dome PTZ IP Camera The PTZ camera can make horizontal movements and zooms, making it ideal for tracking a suspicious person and observing their actions in real-time. The dome PTZ camera offers the ability to monitor areas by moving through a 360-degree angle.

4. Vandal-Proof IP Camera as the name suggests, this camera is designed to withstand external assaults, such as vandalism and tampering attempts. It is equipped with special surfaces that are resistant to breaks and shocks, making it the ideal surveillance camera for protecting valuable locations and enhancing security.

5. Spy IP Camera: due to its small size, the spy camera can easily be hidden in motion detectors and other small objects. It can also come in various forms, such as a pen, keychain, or watch. Its main advantage is that it can better catch intruders attempting to infiltrate our home or business. Since it is easily concealed, it is not readily noticeable at first glance.

6. Infrared IP Camera: this night vision surveillance device can record in complete darkness with precision up to 30 meters. The infrared camera is practical for protecting against thieves and burglaries while you sleep peacefully.

2.2.2 Recorder and Storage

The video recorder processes and stores the videos captured by the cameras. There are two main types of recorders:

1. Digital Video Recorders (DVRs):these are used with analog cameras and are part of traditional surveillance systems.

2. Network Video Recorders (NVRs): these work with IP cameras and are more suited for modern, digital setups. NVRs often provide all-in-one solutions that simplify the surveillance system, especially in entry-level applications.

2.2.3 Transmission and Connectivity

The method of transmitting the video signal from the cameras to the monitoring site is crucial. Options include hard-wiring with cables such as CAT5e or CAT6 Ethernet cables, or using wireless technologies where cameras connect to a network via routers. The choice between wired and wireless systems depends on the specific installation site and the required flexibility [BHKNA16].

2.2.4 Storage and Compression

Storage is a critical component, with options ranging from local hard drives to cloud storage. Video compression technologies like MJPEG, H.264, or H.265 help in managing storage space and bandwidth by reducing the size of the video files without significant loss of quality (the most commonly used is H.264).

2.2.5 Supporting Technologies

Additional components like cables, routers, and Wi-Fi extenders might be necessary, depending on the system's configuration, for instance, wireless systems need a robust network setup to maintain a reliable connection across the surveillance area.

2.2.6 Remote Viewing and Access

Modern IP video surveillance systems typically offer remote viewing capabilities, allowing users to monitor their property from anywhere via a mobile app. This feature requires internet access to connect to the NVR remotely.

2.3 Video surveillance system architectures

The architecture of video surveillance systems is based on five fundamental decisions: encoding, storage, analytics, management, and monitoring. These decisions determine where the video is encoded, stored, analyzed, managed, and monitored :

1. Encoding: The first decision in a video surveillance system is where the video will be encoded. This can occur directly in the cameras, in separate encoders, or at the recorders. The choice affects the system's flexibility, cost, and complexity.

2. Storage: Video data can be stored in several locations :

   (a) Local Storage: directly on the camera or on a local recorder like DVR or NVR.

   (b) Network Storage: on network-attached storage (NAS) or storage area networks (SAN).

   (c) Cloud Storage: offsite storage that offers scalability and remote accessibility.

3. Analytics: video analytics can be processed at different points in the system:

   (a) Edge Analytics: directly on the camera.

   (b) Centralized Analytics: on dedicated servers or cloud-based services. This placement impacts the responsiveness and scalability of the analytic solutions.

4. Management: the management of video data and system operations can be handled through various platforms:

(a) Video Management Software (VMS): for centralized management of video feeds.

(b) Cloud-Based Management: for remote and scalable management options.

(c) Monitoring: the final output of the video surveillance system is the monitoring interface, which can be stationed locally or accessed remotely, often via cloud services.

However, the complexity of interactions between system components, particularly across different equipment and models, presents challenges. Various methods have been proposed to address this issue, focusing on connecting cameras and gathering information through different architectural approaches. Different architectural approaches categorize connections between main stations and devices into analog, digital, and network systems [EAAM21].

2.3.1 Analog Surveillance System

Historically, surveillance systems relied on analog technology for over two decades. Analog signal processing formed the foundational model for image transmission, exchange, and recording. This involved utilizing short-distance coaxial cables and long-distance transceiver optical fibers. While analog systems offered advantages such as image restoration, they also presented limitations including restricted transmission distance, complex engineering cabling, and inflexible application. The diagram below illustrates the components of an analog-based system, encompassing acquisition, storage, visualization, switching, and processing [EAAM21].

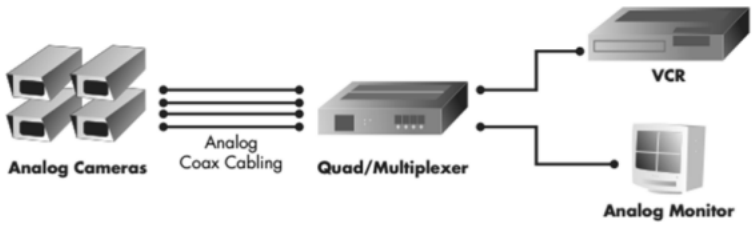

Figure 2.1: Example of an analog video surveillance system. [EAAM21] .

2.3.2 Digital video surveillance

The transition from analog to digital surveillance systems began with the introduction of digital video recorders (DVRs), stemming from analog technology. During this phase, digital image files are transmitted via computer network systems. Cameras connect to a video server through an IP network, enabling transmission over existing computer LANs or even the Internet. This setup allows for camera control and scene zooming via a computer terminal network, effectively transforming traditional systems into distributed ones.

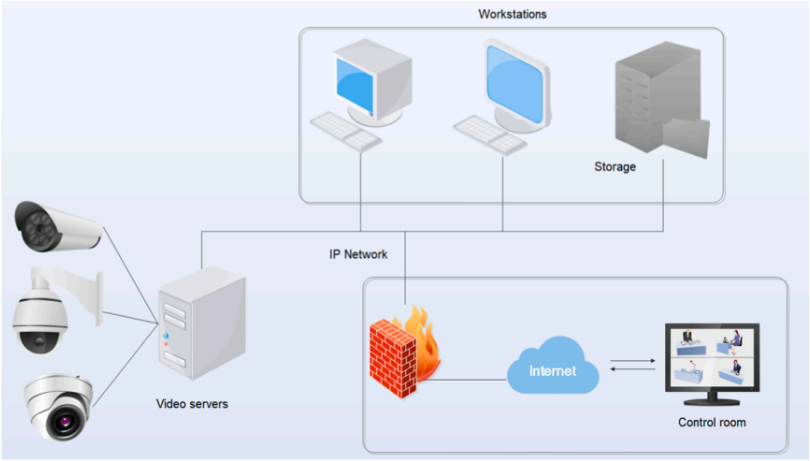

Figure 2.2: Example of a digital video surveillance system [EAAM21].

2.3.3 Network surveillance system

The network surveillance system is founded on digital signal processing and incorporates network cameras or IP cameras, which are digital video cameras. This system allows for the direct connection of as many IP cameras as necessary to the IP network. Networking techniques are employed to achieve signal transmission, exchange, control, and video storage. Additionally, the system can perform centralized control and management of all devices, including cameras and sensors.

The figure 2.3 illustrates the architecture of a network video surveillance system, detailing all devices and their connection methods [EAAM21].

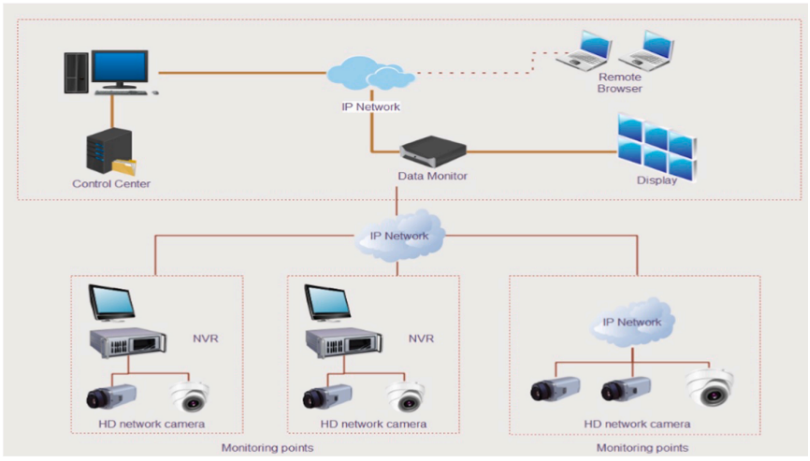

Figure 2.3: Example of a network video surveillance system [EAAM21].

Conclusion

The evolution of surveillance systems from analog to digital and network-based technologies has revolutionized the way data is captured, processed, and stored. Analog systems, despite their limitations in transmission distance and cabling complexity, laid the foundation for modern surveillance solutions. The transition to digital surveillance introduced significant advancements, such as digital video recorders and IP network connectivity, enabling remote monitoring and enhanced control capabilities. Network surveillance systems, leveraging digital signal processing and IP cameras, offer scalability and centralized management options. As technology continues to evolve, understanding the strengths and limitations of each surveillance system type is crucial for organizations seeking to implement effective and efficient security solutions in various environments.

Chapter 3

State of The Art

Introduction

In today's world, with the emergence of computer vision, many researchers are exploring how to integrate this technology into surveillance systems to increase storage capacity and reduce processing time, in order to find the best solution that yields optimal results while considering available resources. In this chapter, we will conduct a Related Work study to understand the different techniques used, and extract the strengths and weaknesses of each technique in order to perform a comparative analysis.

3.1 Related Work

Authors of [ASGB23] present a comprehensive real-time object detection system that integrates motion detection, face detection, and human activity recognition. The system, trained on a large dataset of annotated images, is shown to be effective in various real-world applications such as pedestrian and vehicle detection in crowded urban environments and industrial settings. However, the paper falls short in providing a thorough comparison with existing systems and lacks sufficient details about the training dataset. Additionally, a more detailed discussion on the system's limitations and potential areas for future work would have offered a more balanced perspective and directions for further improvements.

In this research paper [PBJ+23], the authors address the challenge of efficient motion estimation in video sequences, proposing a composite block matching algorithm that integrates object detection to enhance accuracy and efficiency for surveillance video coding. While the method demonstrates higher accuracy and faster processing times compared to traditional approaches, its integration introduces significant computational

complexity, potentially limiting its real-time applicability in large-scale surveillance systems. The reliance on PSNR and SSIM metrics for evaluation may not fully capture the method's practical applicability, warranting a detailed comparative study with other techniques to address these concerns and justify its applicability to various datasets.

The thesis [AY16] addresses challenges such as changes in illumination and shadow detection, aiming to create an algorithm for detecting and tracking mobile objects in a complex scene. The proposed solution involves background modeling and subtraction, handling sudden changes in illumination, and an object tracking model. While the model successfully detects mobile objects and draws their trajectories, it has limitations, especially regarding fixed and mobile cameras. The method's reliance on background modeling may not accurately account for changes in the scene, and it may detect irrelevant movements, leading to false positives and reduced effectiveness in surveillance systems. Additionally, the results presented lack exact statistics, making it difficult to assess the reliability percentage of the model. Further quantitative measures are essential to understand the real performance of the algorithm in various scenarios.

The paper [SD19], provides a comprehensive overview of how deep learning techniques are employed in the realm of intelligent video surveillance, particularly for crowd analysis. The authors systematically examine various deep learning models and their applications in detecting and analyzing crowd behaviors, identifying anomalies, and ensuring security in densely populated areas. They highlight the advancements in convolutional neural networks (CNNs) and recurrent neural networks (RNNs) that have significantly improved the accuracy and efficiency of surveillance systems. The review critically discusses the challenges such as computational complexity, real-time processing, and the need for large annotated datasets. While the paper effectively synthesizes current methodologies and technological advancements, it also points out gaps in the existing research, particularly the need for more robust models that can operate effectively under diverse environmental conditions and varying crowd densities. The authors call for future research to address these limitations, suggesting that integrating multi-sensor data and developing more adaptive algorithms could enhance the robustness and scalability of intelligent surveillance systems.

The article [XHW20] presents a hybrid model combining Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) for improved human activity recognition (HAR). The authors propose this architecture to leverage the spatial feature extraction capability of CNNs and the temporal sequence modeling strength of LSTMs. The study demonstrates superior accuracy and robustness over traditional methods, validating the effectiveness of the LSTM-CNN model. However, it also highlights challenges such

as high computational cost and the need for extensive labeled data, which may hinder real-time applications. The authors suggest further research to optimize computational efficiency and data labeling processes. While promising, the model's practical implementation requires addressing these operational challenges for broader applicability.

The study [AGKVA20] addresses the problem of detecting suspicious individuals and hostile behavior. To achieve this, the authors propose a solution that utilizes two neural networks for human detection and behavior classification. Video frames are first fed into the YOLO model which is pre-trained on the COCO dataset to identify humans in the frame. These frames are then processed by ResNet-34(trained on dataset consisting of video clips of human interactions), which recognizes the activities displayed by the detected individuals. The model demonstrates significant results, achieving a precision of 82%, the highest compared to related works mentioned in the paper. Although the proposed model employs some of the most effective algorithms and delivers efficient results, it relies on two deep learning models, which require substantial computational resources.

In [LO23] the author aims to solve the problem of creating a video surveillance system that is capable of motion detection and object recognition. As a solution, he proposes first detecting the motion and then detecting the object. The author didn't specify the algorithm used; he only specified that he used OpenCV as a library to implement the model. For the results of detection, they didn't provide exact results on how efficient the model is in detection. They just provided statistics on resource consumption, indicating that it reduces the use of CPU and RAM.

The work in [ATSA19] addresses the problem of inefficiency in traditional surveillance systems, particularly in accurately and quickly identifying individuals in video footage. The authors propose a solution involving advanced AI techniques, specifically deep learning models, to enhance human identification capabilities. Their methodology includes training convolutional neural networks (CNNs) on large datasets of facial images and evaluating the performance of these models in real-time surveillance scenarios. The results indicate that AI-based identification significantly improves accuracy and processing speed, making surveillance systems more reliable and responsive. However, the paper also highlights several criticisms: the high computational demands of AI models, potential privacy issues related to extensive data collection, and ethical concerns surrounding surveillance. Future work is suggested to focus on optimizing these AI systems for better efficiency and addressing the associated ethical implications.

The article [UMDS+19] addresses the problem of accurately recognizing human activities in video sequences. The authors propose a method that combines temporal optical flow features from CNNs with multilayer LSTM networks to capture both spatial and temporal information. Their results show improved accuracy in activity recognition over traditional methods. However, the study notes challenges with the high computational demands of their model, which could hinder real-time application. Additionally, while the proposed model performs well, its evaluation on a broader range of datasets would strengthen the findings.

The dissertation [B+22] aims to create a real-time object detection model. The authors chose persons and cats as the instances to detect. To develop their model, they use the YOLO (You Only Look Once) model, which they train on their own dataset composed of images collected by the two students. The results show that the model achieved a Mean Average Precision (mAP) of 76% for detecting cats and 72% for detecting persons. However, the dataset, being limited to images collected by the students, might lack diversity, potentially impacting the model's generalization capability in real-world scenarios. Expanding the dataset to include a wider variety of images could help improve the model's robustness. Additionally, the model's performance could benefit from further hyperparameter tuning and experimentation with different architectures or training techniques. By combining these strategies, the model's accuracy and generalization capabilities can be enhanced, making it more effective in real-time object detection tasks.

The article [SKJG15] addresses the inefficiency of manual surveillance systems in detecting motion. The authors propose an automatic surveillance system that utilizes motion detection algorithms to identify and alert about moving objects in real-time. Their methodology involves implementing background subtraction and frame differencing techniques to detect motion, followed by alert generation. The results indicate that their system effectively reduces the need for constant human monitoring and improves response time to potential security threats. However, the article does not adequately address the system's performance in varying lighting conditions and complex environments, which are critical for real-world application. Furthermore, the paper could benefit from a comparative analysis with existing motion detection systems to better highlight its advantages and limitations.

The paper [DGS+17] presents a system for detecting human body objects and suspicious activities. To develop the system, the authors devise his program in two steps. Firstly, they begin by detecting motion after comparing three motion detection algorithms (background and frame subtraction with optical flow). They decide to use optical flow to detect motion, even though it requires more calculations, as it provides more efficient results. For object detection, they apply a template matching algorithm, which takes an image of

the object as a template and searches for this object in the input image. However, this algorithm is based on pixel comparison, for objects like humans with different forms, styles, sizes, and positions (walking, sitting). To enhance this, it would be beneficial to explore more sophisticated techniques for object detection and identification, such as deep learning and convolutional neural networks, which may be more suitable for detecting complex and varied objects.

The paper [KSG+20] aims to develop a robust model for object detection and tracking. Their solution combines YOLOv3 and RetinaNet to leverage their respective strengths: YOLOv3 is known for its speed and performance on common objects, while RetinaNet excels in detecting small and densely packed objects. This combination ensures comprehensive detection capabilities across various scenarios.

In the framework, detections from each frame are processed using non-max suppression (NMS) to eliminate redundant detections, resulting in a refined set of bounding boxes representing newly located objects. These detections are then fed into a pre-trained CNN model on a person re-identification dataset, which contains 1,261 IDs with 200,000 tracklets. For tracking, a Kalman Filter is employed to optimally estimate state variables during motion, even when the precise location of the system is unknown. The Kalman Filter, combined with Deep SORT, uses the Hungarian algorithm to perform matching based on a cost matrix that considers both the Mahalanobis distance for motion consistency and the cosine distance for appearance similarity. This deep association metric allows Deep SORT to maintain tracking through short periods of occlusion.

The model's performance was evaluated on the VisDrone 2018 dataset using standard MOT evaluation metrics, demonstrating high performance compared to other state-of-the-art models. The combination of YOLOv3 and RetinaNet provides a powerful detection framework, but to avoid the two-pass algorithm and enhance speed, using YOLOv3 alone could be more efficient. The accuracy of YOLOv3 in detecting objects can be improved to match RetinaNet's performance by training on a larger dataset, making it an ideal option for models that can leverage extensive training data.

The paper [D+22] proposes a real-time, online action detection system designed to robustly generalize across unknown surveillance videos. The system addresses key challenges in action classification, such as class imbalance and multi-label actions, by employing techniques like the PLM method and LSEP loss. It achieves state-of-the-art performance on the ActEV-SDL UF-full dataset and secures second place in the TRECVID 2021 ActEV challenge. The methodology includes tracklet generation, activity classification, and prediction refinement using a post-processing algorithm. Additionally, the approach leverages knowledge distillation to enhance computational efficiency. Despite its success, the paper highlights that existing datasets do not fully capture real-world surveillance challenges, such as the untrimmed nature of videos, tiny actor

bounding boxes, and the complexity of multi-label actions.

The article [PB14] addresses the challenge of improving surveillance efficiency by automating the detection and recording of suspicious activities, especially in organizations after office hours. The authors propose a system that uses a webcam to capture video and a motion detection algorithm to identify movements that exceed a predefined threshold. Each detected motion triggers the recording and storage of a new video clip. The methodology aims to enhance security by automating the monitoring process and reducing the need for physical presence. Results indicate improved detection rates and reduced false positives. However, the paper's lack of evaluation across diverse environments and its limited discussion on handling occlusion and varying lighting conditions highlight areas needing further research to ensure the system's robustness and reliability in real-world applications.

In the research paper [BMB21] the authors tackle the problem of small object detection in video sequences. They propose STDnet-ST, a Small Target Detection spatio-Temporal convolutional neural network (ConvNet) that uses two branches and a correlation module to link detections from two input frames, creating spatio-temporal small object tubelets. These tubelets are refined using the Viterbi algorithm and a tubelet suppression procedure to discard unprofitable tubelets while preserving high-quality ones. Enhancements to the original STDnet structure, resulting in STDnet++ and STDnet-ST++, include improved 0-padding operations and a cascaded header to reduce false positives.

The authors validate their architecture on three datasets: USC-GRAD-STDdb, UAVDT, and VisDrone2019-VID, showing that STDnet-ST++ achieves state-of-the-art performance with 2.3% improvement on USC-GRAD-STDdb, 1.0% on UAVDT, and 1.2% on VisDrone2019-VID. The key factors contributing to this success are high-resolution feature maps, correlation over RCN regions, and the combination of correlation-based tubelet linking with tubelet suppression. However, the authors acknowledge the limitation of current datasets in terms of the number of small objects and propose future work involving the use of Generative Adversarial Networks (GANs) to generate synthetic small objects and place them in various contexts.

The article [BKH+17] focuses on developing an automated system for recognizing human activities in surveillance videos using neural networks and digital image processing. The methodology includes techniques like background subtraction, binarization, and morphological operations to extract activity features, followed by a multi-layer feed-forward perceptron network for classification. The system achies a high recognition rate of 94%, demonstrating its effectiveness in identifying daily human activities. However, the study primarily tested the system in controlled indoor environments with static cameras, and the paper does not discuss

limitations or potential challenges in more complex or varied settings. Future work suggested includes testing the system on more challenging datasets and complex activities to enhance robustness and generalizability.

The article [DCG19] addresses the challenge of rapid object detection in compressed videos, crucial for various practical applications but often constrained by computational resources needed to process each frame individually. The authors propose leveraging motion vectors and residual errors available in compressed video streams to significantly reduce processing time. They introduce a cross-resolution feature fusion module (CR-eFF) to enhance detection accuracy by utilizing existing compressed video information. Evaluations demonstrate that this method improves detection precision while cutting down on computational time compared to traditional frame-by-frame processing methods. The algorithm used in this approach is a combination of motion vector-based detection and residual error analysis, integrated with the CR-eFF module to fuse features across different resolutions. However, the approach may be limited by the quality of motion vectors and residual errors in compressed videos and may not perform as well in scenarios with rapidly moving objects or highly compressed videos, which can result in critical information loss.

The article [RHGS16] addresses the problem of achieving fast and accurate object detection in images. The proposed solution is the Faster R-CNN algorithm, which integrates a Region Proposal Network (RPN) with a Region-based Convolutional Neural Network (R-CNN). The RPN efficiently generates region proposals, which are then classified by the R-CNN. This approach was evaluated on benchmark datasets, including PASCAL Visual Object Classes and MS COCO (Microsoft Common Objects in Context). On the PASCAL VOC dataset, Faster R-CNN achieved a mean Average Precision (mAP) of 73.2%, significantly outperforming previous methods. On the MS COCO dataset, it achieved an mAP of 21.9%. The results demonstrated that Faster R-CNN not only enhances detection accuracy but also improves speed, making it suitable for real-time applications. However, the algorithm's computational intensity poses a limitation, particularly for deployment on devices with limited processing power.

The paper [PP21] addresses the challenge of analyzing human crowds in video surveillance due to non-rigid shapes and occlusions. The authors propose a methodology that treats the crowd as a single entity and uses motion patterns to differentiate behaviors. The process involves pre-processing (converting video to frames and noise removal), background subtraction (extracting foreground blobs), motion tracking (estimating optical flow for velocity, position, and direction), clustering (using an adjacency matrix-based clustering algorithm), and feature extraction/classification (using Harris algorithm and SVM). The system achieved high recognition rates of 95.6% and 95.1% on PETS and University of Minnesota (UMN) datasets, respectively. However, the

paper's focus on controlled datasets may not fully represent real-world variability, and the approach may face challenges with highly dynamic or densely packed crowds.

3.2 Detailed Analysis of Methodologies

3.2.1 Background subtraction method

This algorithm involves creating a background model of the scene and subtracting it from the current frame to identify moving objects. Background subtraction is a foundational method for detecting moving objects in videos captured by static cameras. The essence of this approach involves comparing each frame of the video with a reference frame, often termed the "background copy" or "background replica." This reference frame ideally represents the scene without any moving objects and needs to be regularly updated to account for changes in lighting conditions and scene geometry. Traditionally, background subtraction involves detecting regions of motion by analyzing the differences between the current frame and the background reference frame. This method is highly effective for motion detection and can provide valuable data, including information about detected objects.

.

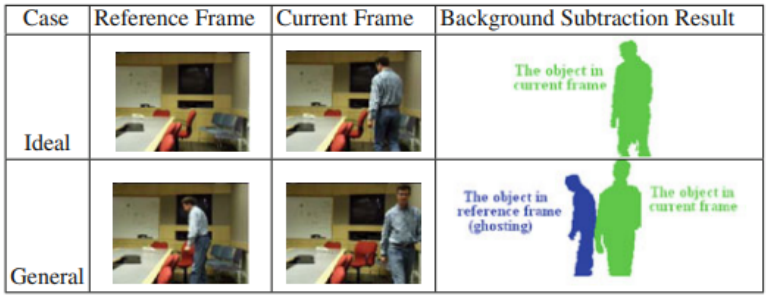

Figure 3.1: Background subtraction results when a background (frame) subtraction technique is used [MMdP12].

3.2.1.1 Strengths

1. Simplicity and Efficiency: background subtraction is straightforward to implement and computationally efficient , making it suitable for real-time applications [CCX18].

2. Effective in Static Camera Scenarios: this method is particularly effective in scenes captured by static cameras, as it can reliably detect moving objects even if they temporarily stop moving. This makes it highly applicable in various monitoring systems.

3. Low Computational Cost: traditional background subtraction methods have low computational costs compared to more sophisticated techniques like deep learning-based methods. This makes them suitable for deployment on lightweight architectures and in scenarios where computational resources are limited [YYL19].

3.2.1.2 Weaknesses

1. This method is limited in dynamic environments or non-fixed camera setups where lighting conditions and scene geometry may change frequently.

2. Noise from poor-quality image sources, gradual variations in lighting conditions.

3. Small movements of non-static objects such as tree branches and bushes blowing in the wind .

4. Undeviating variations of objects in the scene such as cars that park or depart after a long period.

5. Sudden changes in light conditions such as sudden rain or the presence of a light switch, and movements of objects in the background that leave parts of it different from the background model.

6. Sensitivity to parameters: the performance of background subtraction algorithms can be highly dependent on the choice of parameters, making it challenging to achieve optimal results in all scenarios.

However, more sophisticated models have expanded the concept of background subtraction beyond its literal interpretation. Such as using filters to eliminate the noise or background updating, which has been proposed to solve scene change problems. Essentially, this time-differencing method proposes that a pixel is considered part of a moving object if its intensity has significantly changed between the current frame and the previous one [RP13]. Mathematically, this is expressed as $|I_t(x) - I_{t-1}(x)| < \tau$ where t is a predefined threshold [MMdP12] [YL15].

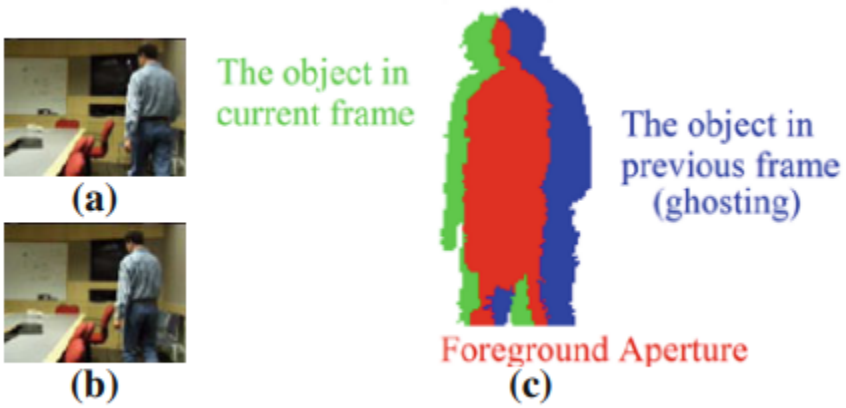

Figure 3.2: Drawbacks of adjacent frame difference approach [MMdP12].

3.2.2 Optical Flow

The optical flow method detects moving objects or humans by analyzing their velocities. It serves as a crucial feature for both object detection and tracking, offering insights into the spatial arrangement of viewed objects and the rate of change in this arrangement. The optical flow describes the direction and time pixels in a time sequence of two consequent dimensional velocity vectors, carrying direction, and the velocity of motion is assigned to each a given place of the picture. Optical flow algorithms are commonly divided into

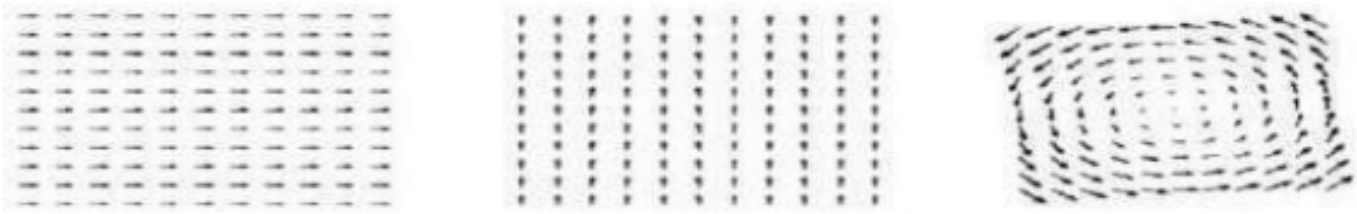

Figure 3.3: Example of optical flow [UMDS+19] .

two classes based on regularization: feature-based methods and variational methods. Feature-based methods calculate the optical flow solution for each pixel and its neighborhood independently of other pixels in the image. In contrast, variational methods consider the optical flow solutions of neighboring pixels and typically incorporate smoothness assumptions into the flow field.

3.2.2.1 Strengths

1. Accuracy: optical flow algorithms can provide high accuracy in estimating the motion of objects in a video sequence.

2. robustness: optical flow methods are often robust to noise and can handle complex motion patterns in the video.

3. Versatility: optical flow can be used in various applications, such as object tracking, motion estimation, and image stabilization.

4. Real-time processing: some optical flow algorithms can achieve real-time performance, making them suitable for applications that require fast processing speeds.

3.2.2.2 Weaknesses

1. Aperture problem: Optical flow algorithms may struggle to estimate motion accurately in regions where the local image gradients are parallel to the motion direction, a phenomenon known as the aperture problem.

2. Discontinuities: Optical flow methods may encounter difficulties in handling motion discontinuities or occlusions in the video sequence.

3. Computational complexity: optical flow algorithms are computationally expensive, especially when dealing with high-resolution videos or videos with complex motion patterns.

4. Sensitivity to parameters: The performance of optical flow algorithms can be highly dependent on the choice of parameters, making it challenging to achieve optimal results in all scenarios.

3.2.3 Block Matching Algorithms

Pixel-based Block Matching Algorithms (BMA) are traditional algorithms used in video compression, motion estimation, and object tracking by finding correspondences between blocks of pixels in two sub-sequence frames. Block matching methods are effective yet computationally intensive. [BRZ24] The process involves calculating a cost function at each possible location within a search window to find the best match for a macro-block in the reference frame. This helps in discovering temporal redundancy in the video sequence, thereby enhancing inter-frame video compression by referencing the contents of a macro-block to a known macro-block with minimal differences.

3.2.3.1 Strengths

1. High accuracy: Block matching algorithms, particularly the Full Search (Exhaustive Search) method, provide the highest peak signal-to-noise ratio (PSNR) for motion-compensated images. This method

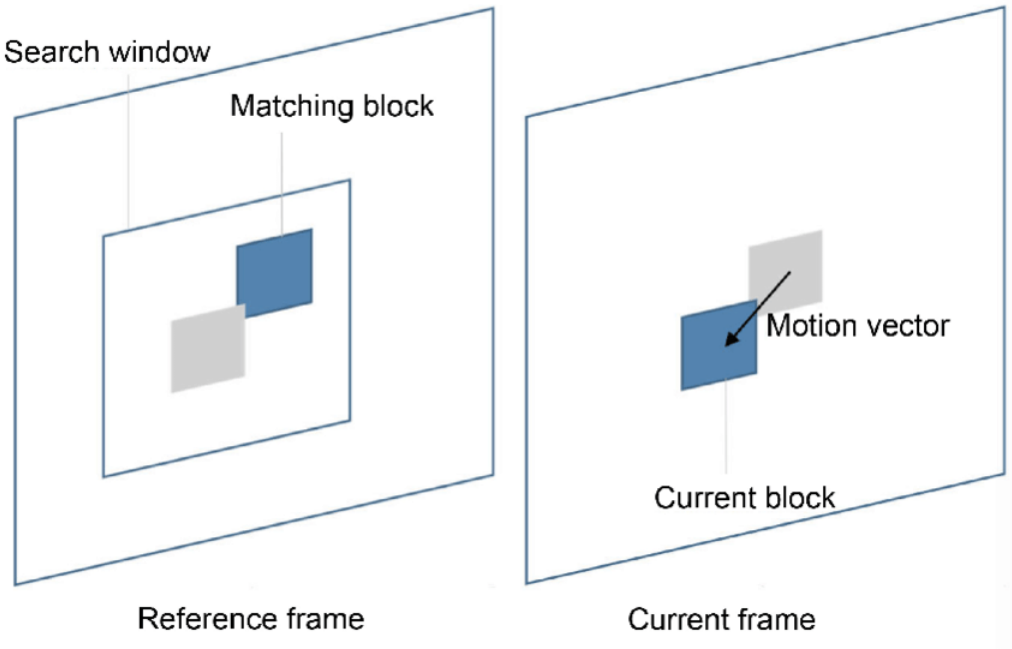

Figure 3.4: Block-matching motion estimation [YCS17].

evaluates all possible positions within the search window, ensuring the best possible match.

2. Versatility: these algorithms are versatile and can be adapted to various video compression standards, from MPEG1/H.261 to MPEG4/H.263. They are widely accepted in the video compression community and have been implemented in numerous standards [Bar04].

3. Robustness: advanced block matching algorithms, such as those using hierarchical or multi-layer approaches, offer robustness in motion estimation. These methods can handle different types of motion and provide reliable performance across various video sequences.

4. Improved compression efficiency: by effectively estimating motion, block-matching algorithms significantly improve video compression efficiency. They reduce the amount of data required to represent motion between frames, leading to better compression ratios.

3.2.3.2 Weaknesses

1. Computational complexity: the full search method, while accurate, is computationally intensive. It requires evaluating all possible positions within the search window, making it the most expensive computationally block matching algorithm.This high computational demand can be a limitation for real-time applications.

2. Difficulty with small and fast-moving objects: these algorithms may struggle with accurately estimating the motion of small or fast-moving objects. The fixed block size can limit the algorithm's ability to

capture fine details and rapid movements [KSS17].

3. Complexity in implementation: developing and fine-tuning block matching algorithms can be complex. It requires expertise in both algorithm design and the specific application domain. Additionally, optimizing these algorithms for different video sequences and compression standards can be challenging.

Table 3.1: Comparative Table of motion detection algorithms

algorithm	papers	precision	speed	computational complexity	Real-Time Efficiency
Background substraction	[AY16], [BKH+17], [RP13]	High	high	low	Ideal
Optical flow	[KMK16], [UMDS+19]	very high	low	high	Less Efficient
Block matching algorithme	[PBJ+23]	very high	low	high	Less Efficient

3.2.4 Machine learning algorithms

3.2.4.1 SVM

Support Vector Machines are pivotal in surveillance due to their robust classification, especially in complex, high-dimensional data. SVMs work by establishing a hyperplane to separate different classes, aided by support vectors that influence its position and a margin that maximizes the distance from this hyperplane to the nearest data points of each class. In human detection, SVMs excel at feature extraction using methods like Histograms of Oriented Gradients (HOG), which capture relevant image features. Trained on labeled data to discern humans from other objects, SVMs then classify new data, such as images or video frames, effectively identifying human presence amidst varying backgrounds or complexities. **Strengths**

- High accuracy : SVMs are known for their high accuracy in classification tasks , essential for reliable human detection.
- Effective in High Dimensions : SVMs perform well in high-dimensional spaces handling the complex features extracted from images and videos .

Weaknesses

- Computationally intensive: training SVMs can be resource-intensive, which is a limitation for real-time applications.
- Sensitivity to parameters: in particular, performance with the kernel strongly depends on other hyperparameters and, therefore, is highly sensitive to the settings of these hyperparameters.
- Scalability problems: these could involve issues with training time and memory usage with massive datasets.

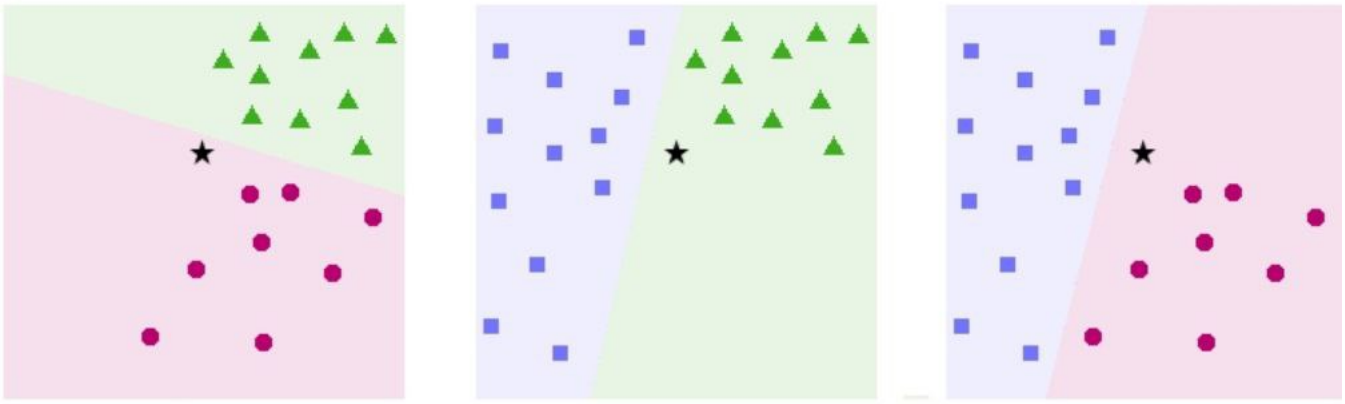

Figure 3.5: Exemple of use of the svm [Chr20].

3.2.5 Deep Learning

3.2.5.1 CNN

Convolutional Neural Networks (CNNs) operate by employing convolutional layers that extract hierarchical features from input data through convolutional filters, capturing important patterns like edges, textures, and shapes. Subsequently, pooling layers downsample these feature maps, preserving essential information while reducing spatial dimensions for computational efficiency. Fully connected layers then perform classification based on the extracted features, enabling CNNs to accurately detect humans in images or video frames amidst complex backgrounds. This process involves training the network using supervised learning with labeled data, where the CNN learns to recognize human presence through iterative adjustments of its internal parameters, ultimately leading to proficient classification capabilities in human detection tasks within surveillance systems.

Strengths

1. Hierarchical feature learning: CNN learns hierarchical features to capture very fine-grained patterns needed for human detection.

2. Translation Invariance: CNNs are by design invariant concerning translation and capture an object in an image independently of its position or orientation.

3. Top Performance: CNN architectures, such as ResNet, VGG, and EfficientNet, have attained high accuracy in object-detection tasks.

Weaknesses

1. Computational complexity: deep CNN training is computationally heavy; it demands an enormous amount of computation.

2. High dependency on large labeled data: CNNs are architectures that require a lot of labeled data.

3. Interpretability: the extreme convolution of CNNs at their core makes it very hard to interpret how they arrive at their decisions. Thereby, models face specific problems in interpretability.

3.2.5.2 YOLO

You Only Look Once (YOLO) marked a significant advancement in object detection due to its innovative approach of performing detection and classification simultaneously in a single pass through a convolutional neural network (CNN). Unlike traditional methods that propose regions of interest in a separate initial step, YOLO divides the input image into a grid, with each cell responsible for predicting bounding boxes and class probabilities directly from features extracted by the CNN. This eliminates the need for multiple image scans, combining real-time speed with high accuracy. YOLO's architecture includes several convolutional and pooling layers that capture useful patterns at different spatial scales, significantly reducing computational costs. It also uses predefined "anchors," or bounding boxes of various sizes and shapes, to improve detection across different object types and scales, enhancing its precision.

Strengths

1. Instantaneous detection: YOLO detects objects instantly, significantly reducing the number of computations required.

2. Optimized resource usage: it calculates shared features only once, making efficient use of computational resources.

3. High performance and accuracy: YOLO's pipeline approach allows it to generalize well across objects of various shapes, enhancing robustness in diverse scenarios.

4. High-Resolution image processing: its efficient architecture allows for the processing of larger images without sacrificing speed, which is beneficial for applications like aerial or satellite detection.

Weaknesses

1. Localization errors: YOLO tends to have more localization errors compared to region-based methods, which can affect the precision of bounding box placement.

2. Difficulty with small objects: it can struggle with detecting small objects in an image due to the coarse division of the image grid.

3. Less accurate on complex backgrounds: YOLO might be less accurate when objects are set against complex backgrounds, as its grid-based approach can oversimplify the context.

Figure 3.6: Image shows an input image divided into a grid , the second image displays the detected objects with bounding boxes and the final image highliths the identified object with refined bounding box [Dat23].

3.2.5.3 Faster R-CNN

Faster R-CNN, short for "Faster Region-Convolutional Neural Network," is a state-of-the-art object detection architecture within the R-CNN family. It was introduced by Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun in 2015. The primary goal of the Faster R-CNN network is to develop a unified architecture that not only detects objects within an image but also precisely locates these objects. It combines the benefits of deep learning, convolutional neural networks (CNNs), and region proposal networks (RPNs) into a cohesive network, significantly improving the speed and accuracy of the model [gee]. Faster R-CNN consists of two main components:

- Region Proposal Network (RPN): this module is responsible for generating region proposals. It applies the concept of attention in neural networks, guiding the Fast R-CNN detection module to where to look for objects in the image .
- Fast R-CNN: this module detects objects in the proposed regions generated by the RPN. The convolutional computations are shared across the RPN and the Fast R-CNN, reducing computational time and improving efficiency.

Strengths:

1. Unified architecture: Faster R-CNN integrates object detection and localization into a single, cohesive network, which enhances both speed and accuracy.
2. Region Proposal Network (RPN): the RPN efficiently generates region proposals, applying the concept of attention to guide the detection process.

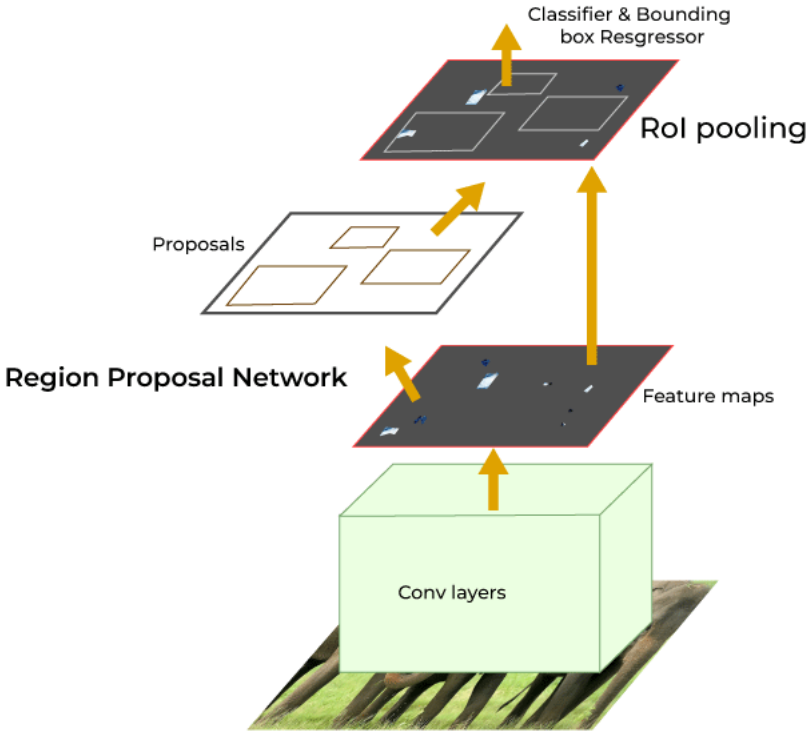

Figure 3.7: Faster R-CNN [gee].

3. Shared convolutional computations: Sharing convolutional computations between the RPN and Fast R-CNN modules reduces computational time and increases efficiency.

4. High accuracy: the combination of deep learning, CNNs, and RPNs results in a highly accurate object detection model.

Weaknesses:

1. Complexity: the architecture is complex, requiring substantial computational resources and expertise to implement and train effectively.

2. Training time: training Faster R-CNN can be time-consuming due to its dual-module structure and the need for fine-tuning.

3. Resource-Intensive: despite its efficiency improvements, the model still requires significant computational power, making it less suitable for real-time applications on resource-constrained devices.

	CNN	Faster R-CNN	YOLOV
Precision	Generally high precision for classification tasks	High precision for both large and small objects	Good precision but may struggle with small or densely packed objects
Speed	Basic CNNs are relatively slower for object detection tasks	Slower than YOLO due to two-stage process	Very fast due to single-pass detection

Table 3.2: Comparison of CNN, Faster R-CNN, and YOLOV

3.3 Discuss the common limitations identified in the reviewed papers

3.3.1 Fixed Camera Requirements :

Many surveillance systems proposed in the related work, such as those discussed in [AY16] and [SKJG15] work only on fixed camera setups. These systems use background modeling and subtraction algorithms, which are effective only when the camera remains stationary. This dependence limits their applicability in dynamic environments where camera positions may change or the scene itself is highly variable, such as mobile surveillance units or areas with frequent layout changes.

3.3.2 Lack of Real-time Processing

A significant number of reviewed systems, like [PBJ+23] And [SD19], struggle with real-time data processing. The high computational complexity of algorithms such as composite block matching and deep learning models (e.g., CNNs and RNNs) often prevents them from operating in real-time. This limitation severely impacts their effectiveness in live monitoring scenarios, where immediate response is crucial, such as in emergency response or active threat detection.

3.3.3 Indoor/Outdoor Specific Constraints

Some systems are designed to work exclusively in either indoor or outdoor environments, limiting their versatility. For example, the study [BKH+17] was tested in controlled indoor environments and does not discuss potential challenges in more complex or varied settings, such as outdoor environments.

3.3.4 Computational Requirements

High computational demands can hinder the deployment of surveillance systems, especially on resource-constrained devices.The Faster R-CNN algorithm, as discussed in [RHGS16], and RetinaNet, as discussed in

[KSG+20], are both known for their high computational demands, making deployment on resource-constrained devices challenging, especially for real-time detection where processing must be parallel with video capturing. The computational intensity of these algorithms poses a significant barrier to achieving real-time performance, particularly on devices with limited processing power. This is crucial for ensuring that the processing and video capture occur simultaneously, demanding efficient and optimized computational resources for seamless real-time surveillance applications.

3.3.5 Generalization Issues

The challenge of generalizing models to diverse real-world conditions due to limited and specific datasets is evident in multiple sources, for instance, [B+22] acknowledges that existing datasets fail to capture the full complexity of real-world scenarios, such as untrimmed videos and multi-label actions, impacting the generalization of the action detection system. Similarly, [A+19] uses deep learning for object detection and tracking in video surveillance, but the model's performance may degrade in diverse real-world conditions not represented in the training data. Additionally, [LO23] develops a video surveillance system for motion detection and object recognition, but the system's generalization to varied environments, different classes of objects, and varying image quality, such as changes in lighting and resolution, is not thoroughly evaluated. These limitations highlight the need for more comprehensive datasets and robust algorithms capable of adapting to diverse and challenging real-world conditions.

3.4 Proposed Solutions and Research Focus

Given the identified gaps and limitations in current human detection and surveillance methodologies, several solutions and research directions can be proposed to address these challenges and enhance the effectiveness of these systems.

3.4.1 Hybrid Approaches

Combining Multiple Methodologies: To leverage the strengths and mitigate the weaknesses of different techniques, hybrid approaches can be developed. For instance, combining background subtraction with deep learning models can enhance robustness in dynamic environments, while integrating optical flow with SVM can improve motion tracking accuracy.

3.4.2 Generalization Improvements

Expanding and Diversifying Training Datasets: Increasing the diversity and size of training datasets can help improve model generalization to diverse real-world conditions. Synthetic data generation, data augmentation, and transfer learning from pre-trained models can also enhance generalization capabilities.

3.4.3 Environment-specific Adaptations

Developing Versatile Models, creating models that can perform well in both indoor and outdoor environments can increase system versatility. This can involve training models on mixed datasets and using context-aware mechanisms to adjust to different environments automatically.

Conclusion

In this chapter, we have conducted a thorough review of the state of the art in video surveillance systems, focusing on the integration of computer vision techniques to enhance storage capacity and reduce processing time. By examining various methodologies, including background subtraction, optical flow, machine learning, and deep learning approaches, we have identified the strengths and weaknesses of each technique. This comparative analysis provides valuable insights into the current advancements and limitations in the field, offering a solid foundation for future research aimed at optimizing surveillance systems. Our findings suggest that while significant progress has been made, there are still critical areas that require further exploration to achieve more efficient and effective surveillance solutions.

General Conclusion and Perspectives

In this thesis, we explored the domain of intelligent video recording optimization by leveraging activity detection techniques. Our primary objective was to identify and evaluate the most effective algorithms for detecting important scenes in video surveillance systems before initiating recording.

We conducted a comprehensive literature review of related work in this field, focusing on various methods employed for activity detection, object detection, and tracking. This review encompassed techniques such as motion detection (including background subtraction and optical flow) as well as object detection and tracking algorithms like YOLOv3, CNN, Faster R-CNN, and block matching. Through detailed analysis, we discussed the strengths and weaknesses of these methods under different real-world conditions, using various camera setups.

Through our research, we concluded that a hybrid approach, combining the strengths of various methods, could offer a more balanced solution. Specifically, integrating motion detection techniques for initial activity screening with advanced object detection algorithms for precise identification and tracking could optimize both performance and resource usage.

It is important to highlight that many models, including YOLO, are in a state of rapid evolution. New versions are continuously being developed to enhance performance and address existing limitations. This trend is evident across various models, as researchers strive to improve speed, accuracy, and efficiency. The ongoing advancements in deep learning models are promising, suggesting that more robust and capable solutions will emerge in the near future.

In conclusion, future research should focus on refining model structures to enhance detection and localization precision while maintaining computational efficiency. Expanding the dataset with more diverse and annotated data can significantly improve model training and performance. Additionally, leveraging more powerful training devices can greatly boost the accuracy and speed of these models. The methodologies explored in this research have potential applications that extend beyond the security sector, benefiting any field requiring real-time object detection and intelligent video recording. These improvements will pave the way for more effective and efficient surveillance systems, ensuring higher precision with optimal resource usage.

Appendix A

Glossary

API (Application Programming Interface): A set of routines, protocols, and tools for building software applications.

Deep Learning: A subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.

IoU (Intersection over Union): A metric used to measure the accuracy of an object detector on a particular dataset.

mAP (mean Average Precision): A metric used to evaluate the accuracy of object detection models.

SVM (Support Vector Machine): A supervised machine learning algorithm which can be used for both classification or regression challenges.

PCA (PrincipalComponentAnalysis):

A statistical procedure that transforms possibly correlated variables into a smaller number of uncorrelated variables called principal components

SVD (SingularValueDecomposition): A mathematical technique used in machine learning to reduce the dimensionality of data

Appendix B

Acronyms

AI	Artificial Intelligence
AI HLEG	Artificial Intelligence High-Level Expert Group
API	Application Programming Interface
AVI	Audio Video Interleave
CCTV	Closed Circuit Television
CLI	Command-line interface
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DNN	Deep Neural Network
DQN	Deep Q-Networks
DW	Directed Work
GFLOPs	Giga Floating Point Operations Per Second
GPU	Graphics Processing Unit
HD	High Definition
IoU	Intersection over Union
LSTM	Long Short-Term Memory
mAP	mean Average Precision
mAP@0.5	mean Average Precision at Intersection over Union threshold of 0.5
ML	Machine Learning
MP4	MPEG-4 Part 14

Ms COCO	Microsoft Common Objects in Context
NMS	Non-Maximum Suppression
PCA	Principal Component Analysis
PIL	Python Imaging Library
PR	Pattern Recognition
PX	Pixel
PW	Practical Work
RCNN	Region-based Convolutional Neural Network
RPN	Region Proposal Network
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SS	Surveillance System
SSD	Single Shot MultiBox Detector
STDnet-ST++	Spatio-temporal ConvNet for small object detection
TP	True Positive
TPU	Tensor Processing Unit
TN	True Negative
UI	User Interface
YOLO	You Only Look Once
YAML	Yet Another Markup Language

Bibliography

[A+19] HV Ravish Aradhya et al. In *Object detection and tracking using deep learning and artificial intelligence for video surveillance applications*, volume 10. Science and Information (SAI) Organization Limited, 2019.

[AGKVA20] S Adarsh, S Poorvaja Giridhar Kannan, BS Vidhyasagar, and J Arunnehru. Suspicious activity detection and tracking in surveillance videos. volume 7, pages 75–79, 2020.

[Arm] Arm's glossary entry on pattern recognition.

[ASGB23] Mohammed Arham, Amisha Srivastava, Akshatha G, and Rajendra A B. In *Motion Detection And Human Activity Recognition For Security*, volume 11, 2023.

[ATSA19] Eman Alajrami, Hani Tabash, Yassir Singer, and M.-T. El Astal. On using ai-based human identification in improving surveillance system efficiency. *2019 International Conference on Promising Electronic Technologies (ICPET)*, pages 91–95, 2019.

[AY16] Sadoun Abdelbaki and Ouellabi Yacine. Détection et suivi des objets mobiles: Application dans un environnement de foule. Master's thesis, Université ECHAHID HAMMA LAKHDAR D'EL OUED, Algérie, 2016.

[B+22] Noureddine BOUMEDIENE et al. *Détection d'objet en temps réel en utilisant une approche basée sur l'apprentissage profond*. PhD thesis, Université Ibn Khaldoun-Tiaret-, 2022.

[Bar04] Aroh Barjatya. In *Block matching algorithms for motion estimation*, volume 8, pages 225–239, 2004.

[BHKNA16] Amal Ben Hamida, Mohamed Koubaa, Henri Nicolas, and Chokri Ben Amar. Video surveillance system based on a scalable application-oriented architecture. In *Multimedia Tools and Applications*, volume 75, pages 17187–17213. Springer, 2016.

[BKH+17] Mohanad Babiker, Othman Omran Khalifa, Kyaw Kyaw Htike, Aisha Hassan, and Muhamed Zaharadeen. Automated daily human activity recognition for video surveillance using neural network. *2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, pages 1–5, 2017.

[BMB21] Brais Bosquet, Manuel Mucientes, and Víctor M. Brea. In *STDnet-ST: Spatio-temporal ConvNet for small object detection*, volume 116, page 107929, 2021.

[Bro19] Jason Broenlee. *Deep learning for computer vision*. 2019.

[BRZ24] Mohammadreza Bayat, Liu Rongke, and Haleh Zarrini. Utilizing motion direction of video capturing satellites for complexity reduction in video compressor block-matching techniques. 2024.

[CCX18] Sheng-Yi Chiu, Chung-Cheng Chiu, and Sendren Sheng-Dong Xu. In *A background subtraction algorithm in complex environments based on category entropy analysis*, volume 8, page 885. MDPI, 2018.

[Chr20] Walid Chrimni. Comprendre les support vector machines (svm), september 2020.

[D+22] Ishan R. Dave et al. Gabriellav2: Towards better generalization in surveillance videos for action detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, 2022.

[Dat23] DataScientest. You only look once (yolo): Tout savoir, 2023.

[DCG19] Benjamin Deguerre, Clément Chatelain, and Gilles Gasso. Fast object detection in compressed jpeg images. In *2019 ieee intelligent transportation systems conference (itsc)*, pages 333–338. IEEE, 2019.

[DGS+17] PA Dhulekar, ST Gandhe, Anjali Shewale, Sayali Sonawane, and Varsha Yelmame. Motion estimation for human activity surveillance. In *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, pages 82–85. IEEE, 2017.

[EAAM21] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. In *A review of video surveillance systems*, volume 77, pages 103–116. Elsevier, 2021.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. The MIT Press, 2016.

[gee] Faster r-cnn | ml - geeksforgeeks. https://www.geeksforgeeks.org/faster-r-cnn-ml/.

[HS24] Jim Holdsworth and Mark Scapicchio. Ibm ,deep learning. https://www.ibm.com/topics/deep-learning?, june 2024.

[IBM24a] IBM. What is computer vision? https://www.ibm.com/topics/computer-vision, june 2024.

[IBM24b] IBM. What is machine learning? https://www.ibm.com/topics/machine-learning?, june 2024.

[KMK16] Ibrahim Kajo, Aamir Saeed Malik, and Nidal Kamel. An evaluation of optical flow algorithms for crowd analytics in surveillance system. In *2016 6th International conference on intelligent and advanced systems (ICIAS)*, pages 1–6. IEEE, 2016.

[Kol18] Maheshkumar H Kolekar. *Intelligent video surveillance systems: an algorithmic approach*. Chapman and Hall/CRC, 2018.

[KSG+20] Shivani Kapania, Dharmender Saini, Sachin Goyal, Narina Thakur, Rachna Jain, and Preeti Nagrath. Multi object tracking with uavs using deep sort and yolov3 retinanet detection framework. In *Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems*, pages 1–6, 2020.

[KSS17] T Kalaiselvi, P Sriramakrishnan, and K Somasundaram. In *Survey of using GPU CUDA programming model in medical image analysis*, volume 9, pages 133–144. Elsevier, 2017.

[LO23] Ruslan Lys and Yurii Opotyak. Development of a video surveillance system for motion detection and object recognition. *Advances in Cyber-Physical Systems*, 2023.

[MDP] Mdpi's topic collection on "applications in image analysis and pattern".

[MMdP12] Ester Martínez-Martín and Ángel P del Pobil. *Robust motion detection in real-life scenarios*. Springer Science & Business Media, 2012.

[Mok12] Djamila Mokhtari. Université de montréal département d'informatique et de recherche opérationnelle faculté des arts et des sciences,". *Détection des chutes par calcul homographique*, 2012.

[PB14] Rucha D. Pathari and Sachin M. Bojewar. In *Automated Surveillance System Using Clustered Matching*, 2014.

[PBJ+23] Arup Kumar Pal, Bhaskar Biswas, Mihir Digamber Jichkar, Adarsh Nandan Jena, and Manish Kumar. Object detection driven composite block motionestimation algorithm for high-fidelitysurveillance video coding. 2023.

[PP21] Shankargouda Patil and Kappargaon S. Prabhushetty. An efficient motion based group level activity recognition for intelligent video surveillance. *2021 Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–8, 2021.

[RHGS16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. In *Faster R-CNN: Towards real-time object detection with region proposal networks*, volume 39, pages 1137–1149. IEEE, 2016.

[RP13] Rupali S Rakibe and Bharati D Patil. In *Background subtraction algorithm based human motion detection*, volume 3, pages 2250–3153. Citeseer, 2013.

[SD19] G. Sreenu and S. Durai. In *Intelligent video surveillance: a review through deep learning techniques for crowd analysis*, volume 6, pages 1–27. Springer, 2019.

[SKJG15] Harshal Suradkar, Aniket Kolte, Shreenath Jamdade, and Sailee Gokhale. Automatic surveillance using motion detection. volume 3, page 525. IJERGS, 2015.

[Spr] Pattern recognition and image analysis. *Journal by Springer*.

[UMDS+19] Amin Ullah, Khan Muhammad, Javier Del Ser, Sung Wook Baik, and Victor Hugo C. de Albuquerque. In *Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM*, volume 7, pages 51177–51188. IEEE, 2019.

[XHW20] Kun Xia, Jianguang Huang, and Hanyu Wang. In *LSTM-CNN Architecture for Human Activity Recognition*, volume 8, pages 56855–56866, 2020.

[YCS17] Zhuge Yan, Siu-Yeung Cho, and Sherif Shaker. In *Predictive Block-Matching Algorithm for Wireless Video Sensor Network Using Neural Network*, volume 05, pages 66–77, 01 2017.

[YL15] Amara Yasmine and Akliouat Lynda. *Détection et suivi automatiques d'objets en mouvement*. PhD thesis, Université Mouloud Mammeri, 2015.

[YYL19] Tianming Yu, Jianhua Yang, and Wei Lu. In *Refinement of background-subtraction methods based on convolutional neural network features for dynamic background*, volume 12, page 128. MDPI, 2019.

Summary

This master thesis presents a comprehensive review of various activity detection methods that can be used in surveillance systems to capture relevant footage. The review encompasses twenty articles, covering a wide range of intelligent video system approaches. The key methodologies examined include background subtraction, optical flow, machine learning techniques such as Support Vector Machines, and deep learning techniques including You Only Look Once, Convolutional Neural Networks, and Faster Region-based Convolutional Neural Networks. Each method is thoroughly explained, with an emphasis on their respective strengths and weaknesses. The analysis provides insights into the current state of the art and identifies potential areas for future research and development in surveillance systems.

Key Words : Activity detection , Surveillance systems , Background subtraction , Optical flow , Machine learning, Deep learning, YOLO, CNN, Faster R-CNN

Résumé

Ce memoire de master présente une revue complète de diverses méthodes de détection d'activité pouvant être utilisées dans les systèmes de surveillance pour capturer des séquences vidéo pertinentes. L'objectif est d'analyser l'optimisation des enregistrements des systèmes de surveillance grâce à ces différentes méthodes. La revue couvre vingt articles, abordant une large gamme d'approches de systèmes vidéo intelligents. Les principales méthodologies examinées incluent la soustraction de fond, le flux optique, les techniques d'apprentissage automatique telles que les machines à vecteurs de support, et les techniques d'apprentissage profond incluant You Only Look Once, les réseaux de neurones convolutifs et les réseaux de neurones convolutifs régionaux rapides. Chaque méthode est expliquée en détail, en mettant l'accent sur leurs forces et faiblesses respectives. L'analyse fournit des informations sur l'état de l'art actuel et identifie les domaines potentiels de recherche et de développement futurs dans les systèmes de surveillance.

Mots clés : Détection d'activités, Systèmes de surveillance, Soustraction de fond , Flux optique , Apprentissage automatique , Apprentissage profond, YOLO, CNN , Faster R-CNN

ملخص

تقدم هذه المذكرة مراجعة شاملة لمختلف طرق اكتشاف النشاط التي يمكن استخدامها في أنظمة المراقبة لالتقاط لقطات الفيديو ذات الصلة. الهدف هو تحليل تحسين تسجيلات نظام المراقبة باستخدام هذه الطرق المختلفة. تحتوي المجلة على عشرين مقال تغطي مجموعة واسعة من أساليب أنظمة الفيديو الذكية. تشمل المنهجيات الرئيسية التي تم فحصها طرح الخلفية، والتدفق البصري، وتقنيات التعلم الآلي مثل آلات المتجهات الداعمة، وتقنيات التعلم العميق بما في ذلك ءنت تنظر مرة واحدة فقط، الشبكات العصبية التلافيفية والشبكات العصبية التلافيفية الإقليمية السريعة. يتم شرح كل طريقة بالتفصيل، مع التركيز على نقاط القوة والضعف الخاصة بكل منها. يوفر التحليل معلومات عن الوضع الحالي للفن ويحدد المجالات المحتملة للبحث والتطوير في المستقبل في أنظمة المراقبة.

الكلمات المفتاحية

الكشف عن الأنشطة، أنظمة المراقبة، الطرح الخلفي، التدفق البصري، تعلم الآلة، التعلم العميق، $YOLO, CNN, FasterR-CNN$.
